2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION
Paper ID #50382

Assessing Student Learning on AI Ethics Through a Case Study of Algorithmic Bias

Ms. Annika Haughey, Duke University

PhD student in Mechanical Engineering at Duke University

Siobhan Oca, Duke University

Siobhan Rigby Oca is an assistant professor of the practice in the Thomas Lord Department of Mechanical Engineering and Materials Science at Duke University, NC, USA. She received her B.Sc. from Massachusetts Institute of Technology and Master in Translational Medicine from the Universities of California Berkeley and San Francisco. She completed her Ph.D. in Mechanical Engineering in 2022 from Duke University. Her research interests include applied medical robotics, human robot interaction, and robotics education.

Assessing Student Learning on AI Ethics Through a Case Study of Algorithmic Bias

Annika Haughey, Brian Mann, Siobhan Oca
Mechanical Engineering & Material Science
Duke University

Abstract

Artificial intelligence is increasingly deployed in decision-making contexts, making it crucial for future engineers to understand the risks of algorithmic bias. This study evaluates a classroom case study based on Amazon's discontinued resume screening model, utilizing an open-source synthetic dataset to engage students in identifying and mitigating bias. The module was implemented in graduate-level Ethics ($n_1 = 16$) and Machine Learning ($n_2 = 21$) courses to assess four learning objectives: recognizing bias, interpreting outcomes, evaluating mitigation strategies, and ethical reflection. We observed positive normalized gains across all objectives: recognizing bias showed a 68.2% gain (3.3 ± 1.0 vs. 4.5 ± 0.55); interpreting results, 50.4% (2.6 ± 0.87 vs. 3.8 ± 0.67); mitigation, 45.0% (2.64 ± 0.96 vs. 3.7 ± 0.87); and reflection, 50.5% (3.0 ± 0.99 vs. 4.0 ± 0.78). While Ethics students demonstrated higher baseline familiarity, Machine Learning students achieved higher overall normalized gains (55% vs. 45%), suggesting the module's effectiveness across varying levels of prior exposure. These results indicate that the case study successfully builds both technical competency and ethical awareness.

Introduction

Artificial Intelligence (AI) is rapidly becoming integral to daily life; however, without ethical implementation, it carries significant risks[1]. As future engineers, students must understand these risks and learn to mitigate them before they assume leadership roles in deploying AI models. This study evaluates the effectiveness of a case-study-based module designed to teach these critical concepts.

Case studies effectively bridge the gap between theory and practice, making them vital tools for teaching engineering ethics. While previous works have illustrated AI fairness challenges in defense[2] and healthcare[3], this study utilizes a case focused on AI hiring practices. Uniquely, we incorporate a synthetic dataset to provide students with hands-on experience in analyzing and mitigating algorithmic bias.

This module focuses on Amazon's attempt to automate its hiring process using AI[4]. Amazon trained a model on historical employee resumes to screen applicants. However, due to the

historical gender imbalance in the tech industry, which was also present in the simulated dataset, the model learned to systematically bias against women.

To bring this case into the classroom, we utilized a synthetic dataset and an interactive Python Jupyter Notebook developed in prior work[5]. The dataset mimics resumes found in a traditional tech environment. In the activity, students train a machine learning model to hire certain applicants, observe the initial bias, and then actively implement various mitigation techniques. This hands-on process highlights the "bias-accuracy tradeoff" inherent in Machine Learning (ML) models and equips students with practical tools to manage it. Specifically, this tradeoff arises when technical interventions aimed at improving fairness metrics constrain the model's optimization process. By prioritizing equitable outcomes across protected groups, the model may experience a marginal decline in its overall classification accuracy or predictive performance. Navigating this tension forces students to critically evaluate the ethical balance between mathematical optimization and social equity in automated decision-making.

We established four learning objectives for this module: (1) understand how biases in training data propagate through machine learning models; (2) interpret model results to communicate implications for fairness and accuracy; (3) evaluate the strengths and limitations of bias mitigation strategies; and (4) reflect on the ethical challenges of deploying ML systems, with an emphasis on transparency, interpretability, and responsible AI practices.

Methods

We implemented this case study in two graduate-level courses: "Ethics in Automation," a cross-listed course with approximately 20 undergraduate and master's students, and "Introduction to Machine Learning & Data Science," which enrolled approximately 50 undergraduate, master's, and PhD students. The module was presented as a single active-learning session in both classes. The session opened with a lecture on the case study background and algorithmic bias, followed by a hands-on activity where students worked in pairs to implement bias-mitigation strategies using a Jupyter Notebook.

To evaluate the module, we administered identical pre- and post-class surveys containing 13 Likert-scale items (Table 1). Students rated their agreement with statements ranging from (1) Strongly Disagree to (5) Strongly Agree. Additionally, the post-class survey included two short-answer questions to solicit qualitative feedback. We collected valid responses from 16 students in the Ethics course ($n_1 = 16$) and 21 students in the Machine Learning course ($n_2 = 21$).

Results

We assessed the effectiveness of this case study by analyzing the differences between pre- and post-class survey responses. To provide a granular analysis of student outcomes, the 13 survey items were separated into three distinct categories: Conceptual Understanding, Professional Attitudes, and Self-Efficacy. Items Q1–Q5 and Q7 were grouped under Conceptual Understanding to measure the acquisition of technical knowledge of machine learning bias. Items Q6, Q8, and Q9 constituted the Professional Attitudes domain, designed to evaluate students'

Table 1: Survey Statements

No.	Survey Statements
1	I understand what bias in machine learning means.
2	I can explain how historical data can introduce bias in machine learning models.
3	I am familiar with real-world examples where machine learning has shown bias.
4	I understand how training data influences a model's decision-making process.
5	I can identify potential sources of bias in machine learning datasets.
6	Bias in machine learning is a serious issue that needs to be addressed.
7	Machine learning models are only as unbiased as the data they are trained on.
8	AI developers have a responsibility to actively mitigate bias in their models.
9	I think existing strategies for mitigating bias in machine learning are effective.
10	I feel confident discussing bias in machine learning with others.
11	I believe I can critically evaluate whether an ML model is biased.
12	I feel prepared to contribute to discussions about fairness in AI development.
13	I feel comfortable proposing solutions to mitigate bias in ML models.

recognition of their ethical responsibilities and the perceived gravity of bias in engineering practice. Finally, items Q10–Q13 were categorized as Self-Efficacy, assessing students' perceived confidence and readiness to apply these skills in professional environments, such as critically evaluating models or proposing mitigation strategies. These results are presented in Figure 1, where each column presents the individual course results (machine learning course and ethics course) and the row represents the question category. In these plots, blue violins represent the pre-class distributions, while yellow violins represent the post-class results.

Across all statements, we observed positive average learning gains and decreased variance in the post-class survey results. Figure 2 depicts the changes in mean score (Average Post Score - Average Pre Score) for each survey question from pre-survey to post-survey for the Ethics course (a) and the Machine Learning Course (b). Ethics students showed greater baseline familiarity with bias and ethics pre-class average for all questions was 3.53 ± 0.52 vs. 3.06 ± 0.61 for the machine learning students. The machine learning students, however, demonstrated higher average normalized gains across all survey questions (55% vs. 45%) calculated as $g = \frac{Post-Pre}{5-Pre}$. This strongly suggests that the case study can be effectively adapted to a variety of course settings, even where prior exposure is limited.

To assess the attainment of our learning objectives, we analyzed the survey statements most closely aligned with each goal. For LO1 (Understand biases in ML models), we selected Statement 1: "I understand what bias in machine learning means." As shown in Figure 3a, where

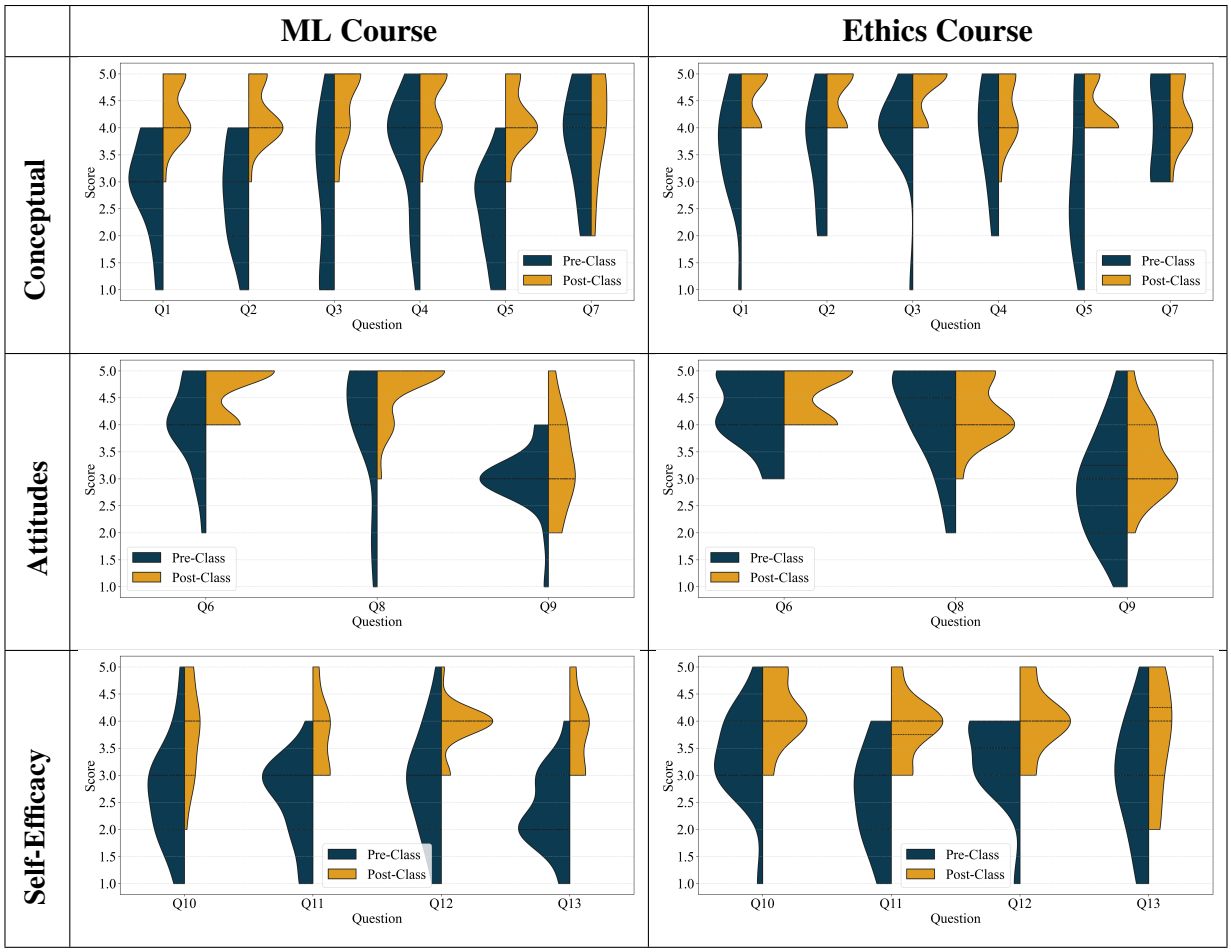
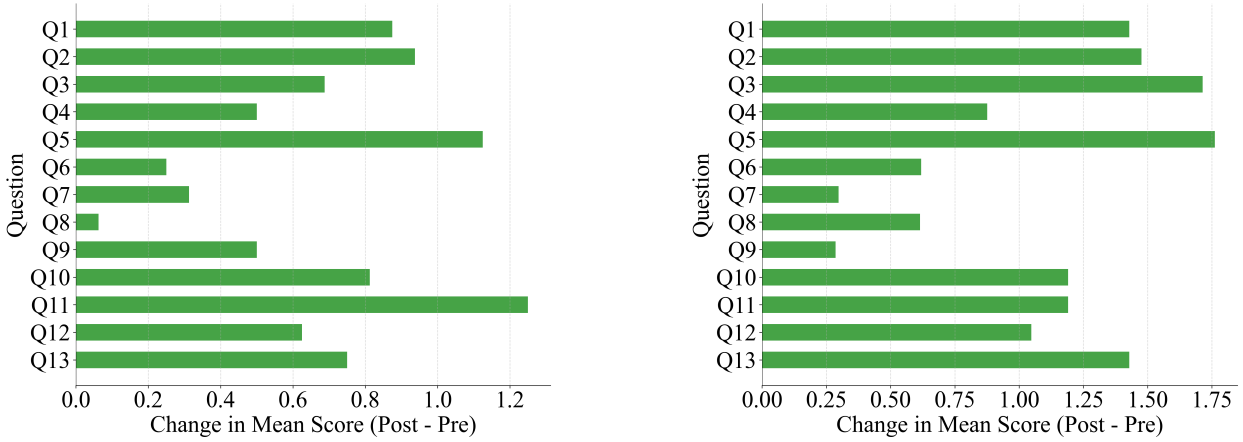


Figure 1: Survey evaluation results by groupings, separated by machine learning course (column 1) and the ethics course (column 2).

yellow lines represent the ethics course and blue dashed lines represent the machine learning course, we observed a normalized gain of 68.2% (3.3 ± 1.0 vs. 4.5 ± 0.55) following the case study. This normalized gain was calculated as $g = \frac{Post-Pre}{5-Pre}$. This high agreement indicates the module was highly effective in achieving this objective.

For LO2 (Interpret ML model results regarding fairness and accuracy), we utilized Statement 11: "I can evaluate if a ML model is biased" (Figure 3b). We observed a normalized average gain of 50.4%, (2.6 ± 0.87 to 3.8 ± 0.67) with most students demonstrating substantial improvement. While the average pre-class response (2.57) was lower than for other statements, the post-class average increased significantly to 3.80.

For LO3 (Evaluate the strengths and limitations of bias mitigation strategies), we examined Statement 13: "I feel comfortable proposing solutions to mitigate bias in ML models" (Figure 3c). The overall normalized average gain was 45% (2.64 ± 0.96 to 3.7 ± 0.87); however, the gain was notably lower in the ethics course (36.36%) compared to the machine learning course (53.57%). This suggests that the machine learning students' greater familiarity with Python and ML frameworks facilitated their achievement of this objective.



(a) Ethics course

(b) ML course

Figure 2: Average survey changes from pre to post-survey for a) the ethics course and b) the ML course.

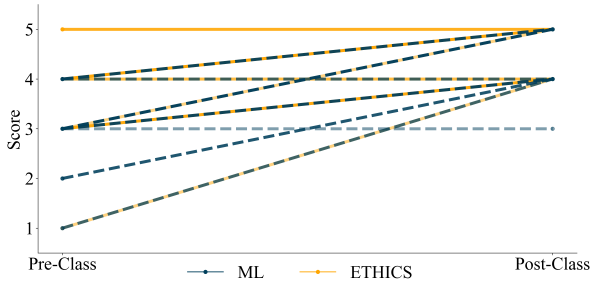
Finally, for LO4 (Reflect on the ethical challenges of deploying ML systems), we tracked Statement 12: "I feel prepared to contribute to discussions about fairness in AI development" (Figure 3d). We observed an average normalized learning gain of 50.5% (3.0 ± 0.99 to 4.0 ± 0.78). Similar to LO3, gains were higher in the machine learning course (48.89%) than in the ethics course (38.46%). This disparity is likely attributable to the ML cohort's technical background combined with their lower initial pre-class confidence.

The learning gains for each of our learning objectives were substantially positive ($> 45\%$). These outcomes suggest that the case study effectively increased student awareness and confidence in addressing the ethical dimensions of AI.

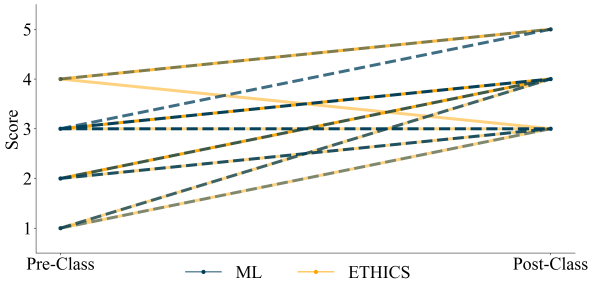
To supplement the quantitative results, the post-class survey included two open-ended questions: (1) "Does this lecture change how you will develop machine learning algorithms in the future? If so, how?" and (2) "What was the most interesting thing you learned?"

For the first question, the majority of students indicated a shift toward more rigorous data scrutiny and the adoption of specific mitigation strategies. Students expressed that they would be "more cognizant" of their datasets, specifically looking for "potential biases" and "training value" before modeling begins. Beyond general awareness, many participants explicitly mentioned they would incorporate the technical tools introduced in the module, stating they would "consider tools like Fairlearn" and "develop algorithms that force equality if needed." Notably, even students who do not write code personally reported that they now felt "knowledgeable to contribute to conversations regarding ML," suggesting the case study effectively empowered different types of students to engage in ethical technical discussions.

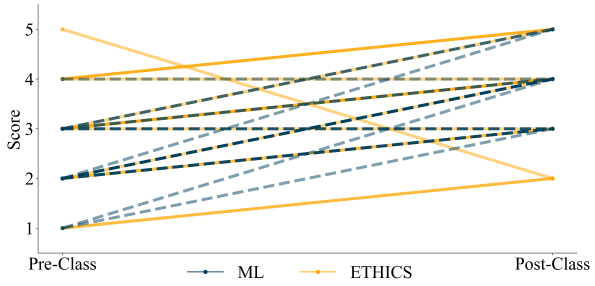
When asked about the most interesting aspects of the module, responses centered on the complexity of the "bias-accuracy tradeoff" and the availability of open-source solutions. Students frequently highlighted the difficulty of balancing performance, noting that "you can't eliminate bias without sacrificing accuracy" and that "undersampling and oversampling have major



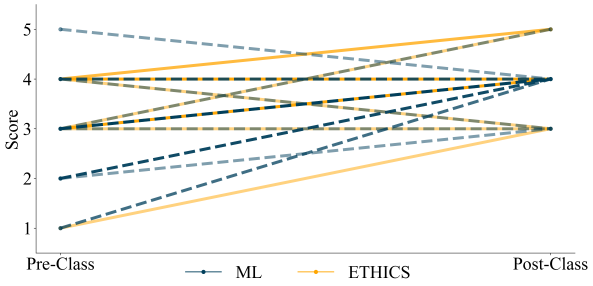
(a) Q1: ”I understand what bias in ML means”



(b) Q11: ”I can evaluate if a ML model is biased”



(c) Q13: ”I feel comfortable proposing solutions to mitigate bias in ML models”



(d) Q12: ”I feel prepared to contribute to discussions about fairness in AI development”

Figure 3: Responses for pre- and post-class surveys, results for statements related to a) LO1, b) LO2, c) LO3, d) LO4. Lines show individual student responses for the ML class (blue) and the Ethics (yellow), highlighting the changes before the class and after. Each individual line has transparency, so darker lines indicate more student responses following this trend.

weaknesses.” However, this challenge was balanced by optimism regarding technical solutions; many students cited the introduction of libraries like SHAP and Fairlearn as the most valuable component, finding it “reassuring to see developers wanting to open the conversation” through open-source tools.

Discussion

The results of this study demonstrate that a technical, case-study-based approach is highly effective for teaching AI ethics to students across different disciplinary backgrounds. The consistently positive learning gains across all four learning objectives suggest that anchoring ethical discussions in real-world scenarios, in our case the Amazon hiring case study, helps students move beyond abstract definitions of fairness to understanding the concrete mechanisms of algorithmic bias.

A key finding of this work is the relationship between technical competency and the ability to successfully employ mitigation strategies. While students in the Ethics course began with a higher baseline familiarity with the concepts of bias, they showed lower normalized gains in LO3 (evaluating mitigation strategies) compared to the Machine Learning cohort (36.36% vs. 53.57%). This disparity suggests that while the recognition of bias (LO1) is accessible to all

students regardless of technical background, the mitigation of bias (LO3) is more dependent on familiarity with the underlying tools (e.g., Python, Pandas, Scikit-learn). For educators implementing this module in non-technical ethics courses, this indicates a need for additional scaffolding, such as paired programming with technical students or low-code alternatives, to ensure students can fully engage with the mitigation exercises.

Despite the technical gap, the qualitative feedback indicates that the module successfully increased students' understanding of the ethical challenges in AI. Non-technical students explicitly reported feeling "knowledgeable to contribute to conversations regarding ML". This aligns with LO4 (reflecting on ethical challenges), where even the less technical Ethics cohort saw substantial gains (38.46%).

The qualitative results also highlight that students grasped the complexity of engineering decisions, specifically the "bias-accuracy tradeoff". Students noted that "you can't eliminate bias without sacrificing accuracy," realizing that ethical AI is not merely a checklist but a design constraint that requires negotiation. This realization is critical for engineering students, it frames ethics as a core technical competency rather than an external constraint. This feedback suggests the module successfully moved students beyond abstract ethical concerns to understanding the concrete technical realities of implementing fair AI.

While the results are promising, the study is limited by a relatively small sample size ($N = 37$) across two specific graduate-level courses. Future work will focus on adapting this module for undergraduate engineering curriculum, where students may have less prior exposure to both ethics and machine learning. Additionally, longitudinal studies are needed to determine if the awareness and confidence gained in this module translate into ethical decision-making in students' future capstone projects and professional careers.

References

[1] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 2016.

[2] Erik Lin-Greenberg. Wrestling with killer robots: The benefits and challenges of artificial intelligence for national security, 2021. URL `https://mit-serc.pubpub.org/pub/wrestling-with-killer-robots/release/2`.

[3] Haoran Zhang, Thomas Hartvigsen, and Marzyeh Ghassemi. Algorithmic fairness in chest x-ray diagnosis: A case study, 2023. URL `https://mit-serc.pubpub.org/pub/algorithmic-chest/release/2`.

[4] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, Oct 2018.

[5] Annika Haughey, Brian P. Mann, and Siobhan Oca. Case study: Using synthetic datasets to examine bias in machine learning algorithms for resume screening. In *2025 ASEE Annual Conference & Exposition*, Montreal, Quebec, June 2025. ASEE Conferences.