Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Fairness Audits of Institutional Risk Models in Deployed ML Pipelines

Kelly M^{c}Convey[†,*], Dipto Das[†], Maya Ghai[†], Angelina Zhai[†], Rosa Lee[†], Shion Guha[†]
[†] University of Toronto, Toronto, Ontario, Canada

Abstract

Fairness audits of institutional risk models are critical for understanding how deployed machine learning pipelines allocate resources. Drawing on multi-year collaboration with Centennial College, where our prior ethnographic work introduced the ASP-HEI Cycle, we present a replica-based audit of a deployed Early Warning System (EWS), replicating its model using institutional training data and design specifications. We evaluate disparities by gender, age, and residency status across the full pipeline (training data, model predictions, and post-processing) using standard fairness metrics. Our audit reveals systematic misallocation: younger, male, and international students are disproportionately flagged for support, even when many ultimately succeed, while older and female students with comparable dropout risk are under-identified. Post-processing amplifies these disparities by collapsing heterogeneous probabilities into percentile-based risk tiers. This work provides a replicable methodology for auditing institutional ML systems and shows how disparities emerge and compound across stages, highlighting the importance of evaluating construct validity alongside statistical fairness. It contributes one empirical thread to a broader program investigating algorithms, student data, and power in higher education.

Keywords: Fairness auditing, educational AI, algorithmic bias, machine learning, higher education, early warning systems

1. Introduction

Public higher education institutions (HEIs) under sustained fiscal pressure have increasingly turned to algorithmic decision-making as a tool for triaging students [1]. Our prior ethnographic work at Centennial College, a public college in Ontario, Canada, theorized this turn as the *ASP-HEI Cycle*: a self-reinforcing pattern in which the prioritization of financial sustainability drives the adoption of data-driven practices that reproduce and expand institutional power [2]. That work's argument about the *formalization of inequity* (§5.4.2) held that deployed risk models encode existing demographic inequities into institutional policy, because biases in training data, opaque categorization, and the quantitative framing of risk interact with pre-existing institutional inequities to disproportionately harm students from equity-seeking groups.

That argument rested on interviews and stakeholder perception; the pipeline-level mechanisms were not, and could not be, tested from ethnographic data alone. At which stages of a deployed Early Warning System (EWS) does inequity enter, and where does it compound? Does the shift from continuous predictions to percentile tiers dampen or magnify demographic disparity? Answering these questions requires something rare in the public-sector ML literature: access to a deployed institutional model together with the training data and design decisions used to build it [3].

This paper tests the §5.4.2 claim and extends the framework. Through continued collaboration with the college at the center of the original ASP-HEI ethnography, we replicate its deployed EWS using the institution's training data and documented design specifications, then audit the full pipeline for disparities by gender, age, and residency status. Our audit

[*] kelly.mcconvey@mail.utoronto.ca

substantiates the claim: each stage of the pipeline is a distinct site of inequity *production* rather than mere reflection, with baseline group differences learned by the model and compounded downstream. Beyond confirming the claim, we identify the specific mechanism of amplification: *percentile-based post-processing* is where baseline demographic differences are converted into amplified institutional policy. We ask:

- **RQ1:** How are different population groups represented and treated by the EWS's (a) training data, (b) predictions, and (c) interpretation?
- **RQ2:** How does the EWS operationalize risk in terms of student drop-out, and how does that operationalization mediate the ASP-HEI cycle's inequity claim?

Our contributions are: (1) **empirical substantiation of ASP-HEI §5.4.2**, showing with real institutional data that the deployed model formalizes historical disparity as current allocation, and specifying the pipeline stages at which this occurs; (2) an **extension of the framework** that identifies percentile-based post-processing as a specific, locatable mechanism of amplification; and (3) a **replicable audit methodology** that generalizes beyond this site to other institutional risk models where training data and design specifications are accessible.

2. Related Work

Research on algorithmic systems in child welfare [4] and homelessness [5] has documented how predictive systems perpetuate disparities through problem formulation and target-variable choice. Raji et al. [3] argue that internal algorithmic audits are essential for accountability, and Simbeck [6] proposes concrete criteria—fairness, transparency, and robustness—for auditing learning analytics systems specifically; our work operationalizes such criteria on a deployed institutional EWS. Obermeyer et al. [7] showed how poor proxy choices produce differential mismeasurement across groups; Passi and Barocas [8] show how problem-formulation decisions embed values; and Jacobs and Wallach [9] highlight how institutional convenience often takes precedence over construct validity. Within higher education, Perdomo et al. [10] evaluate Wisconsin's EWS, demonstrating how prediction-based approaches may fail their stated goals due to problem-formulation issues. Our prior systematic review of algorithms in higher education documented a trend toward less interpretable models built on protected attributes and identified a critical lack of empirical study of how these systems function in practice [11]; our own prior ethnographic work at the present site introduced the ASP-HEI Cycle and identified the formalization of inequity as a central mechanism, but rested on interview data [2]. Our audit addresses that gap at the pipeline level.

3. Methods

Through continued collaboration with Centennial College — the same site and same Early Alert System examined in the original ASP-HEI ethnography [2] — we obtained training data and documented design specifications, enabling a *replica-based audit*—a fairness evaluation conducted on a faithful re-implementation of the deployed model, trained on the same institutional data and following the same documented design specifications—with access to actual training data, documented institutional decisions, and the operational context in which fairness is a resource-allocation question affecting real students.

3.1. Data and Modeling

After research ethics approval, we obtained student records spanning 2011–2019 (102,353 records: 61,375 domestic, 40,978 international). To replicate the intake EWS, we filtered

to first-semester students with features available at intake and trained separate XGBoost models for each population, following institutional design; feature sets differ (English test scores for international, high-school grades for domestic) due to distinct admissions processes. We retained 45 intake/program/admissions features after removing those with substantial missing data. Sensitive attributes (age, gender, residency, funding, first-generation status) were included as model features in line with the replicated system; race/ethnicity and disability data were unavailable. Students were labeled *Successful* if they completed their program within the allowable period and *Unsuccessful* otherwise (withdrawal, failure, transfer). Data were split chronologically (70/15/15), balanced using SMOTE [12], and tuned via grid search.

The EWS pipeline has three stages: (1) intake data collection, (2) separate XGBoost models producing success probabilities, and (3) percentile-based categorization into Low Risk (top 50%), Medium Risk (next 27%), and High Risk (bottom 23%). Test predictions yielded 15,461 students (9,209 domestic, 6,252 international) for fairness analysis.

3.2. Fairness Metrics and Risk Operationalization

We evaluate disparities using Statistical Parity Difference (SPD; difference in positive prediction rates between groups), Equal Opportunity Difference (EOD; difference in true positive rates), False Positive Rate gaps (ΔFPR; difference in false positive rates), and Calibration Error (CE; deviation between predicted probabilities and observed outcome frequencies), computing pairwise disparities and reporting maximum absolute differences [3]. A central methodological decision is the target definition: the institution's EWS predicts whether a student will be *unsuccessful*, collapsing dropout, transfer, and program change into a single label, treating dropout as a proxy for *student need for support*. Following Jacobs and Wallach [9], we treat this as a construct-validity issue, and for this audit we adopt the institution's definition to faithfully replicate the deployed system. We return to its implications in Section 5.

4. Findings

We audit the EWS across three demographic attributes: gender, age, and residency status.

Baseline and model-level disparities. Training data reveal significant group-level differences in program completion: international students achieve 85% success compared with 67% for domestic ($\chi^2 = 2847.3$, $p < 0.001$); female students outperform male in both populations (domestic 73% vs. 59%; international 89% vs. 82%, $p < 0.001$); and students aged 26+ outperform those $\leq$20. International-student models achieve significantly higher accuracy (91%) than domestic (82%, $p < 0.001$), translating into differential access to interventions; both models rely heavily on program-level features (credential type, program length), raising questions about whether student-level prediction is necessary when program-level patterns dominate. Table 1 summarizes error rates across gender and age groups. Female students face systematically higher false-positive rates (domestic: 32% vs. 23%; international: 26% vs. 18%) but lower false-negative rates; all metrics exceed commonly used fairness thresholds (SPD > 0.1). Age effects are larger: students aged 36–40 show false-positive rates above 0.60 in both models, while students aged 19–20 exhibit false-negative rates of 0.21 in the domestic model.

Post-processing amplification. The EWS converts continuous probabilities into three tiers using percentile thresholds ($\approx$0.80 and $\approx$0.39), meaning students with vastly different success probabilities receive identical interventions. The Medium Risk category exhibits poor calibration (Brier Score: 0.18 vs. 0.04 for Low Risk); approximately 21% of High Risk students ultimately succeed. These thresholds amplify demographic disparities beyond

Table 1. Model performance by gender and age. FPR = false positive rate; FNR = false negative rate. Accuracy (gender): Dom. F 0.84, M 0.81; Intl. F 0.92, M 0.91.

		Domestic			International		
	Group	**FPR**	**FNR**	**F1**	**FPR**	**FNR**	**F1**
Gender	Female	0.32	0.10	0.89	0.26	0.05	0.96
	Male	0.23	0.17	0.83	0.18	0.08	0.94
Age	0–18	0.21	0.18	0.84	0.19	**0.09**	0.93
	19–20	0.20	**0.21**	0.81	0.18	**0.10**	0.92
	21–25	0.29	0.10	0.88	0.18	0.06	0.96
	26–30	0.33	0.09	0.90	0.25	0.03	0.97
	31–35	0.41	0.07	0.92	0.29	0.05	0.96
	36–40	**0.61**	0.08	0.90	**0.70**	0.02	0.97
	41–50	0.53	0.05	0.91	**0.69**	0.02	0.95

those in raw predictions. Unsuccessful male students are 10 percentage points more likely to be categorized High Risk than females (74% vs. 63%, $p = 1.56 \times 10^{-13}$), corresponding to roughly 300 additional male students per year receiving intensive interventions. Students $\leq$25 have a much higher probability of High Risk categorization than those 36+ (94% vs. 75%), and unsuccessful international students are 1.12× more likely to receive High Risk classification than domestic students, exceeding the 1.08× difference in raw model accuracy. The EWS additionally operationalizes "risk" as probability of non-completion, conflating dropout, transfer, and program change; percentile categorization further weakens construct validity by treating prediction uncertainty (Medium Risk) as equivalent to moderate intervention need, institutionalizing model error as a resource-allocation rule. This amplification is the technical site at which the ASP-HEI cycle's formalization claim becomes observable: the gap between groups is not a property of the model's predictions but of the decision to translate those predictions into a fixed-quota intervention policy.

5. Discussion

Each stage of the ML pipeline is a distinct technical site at which an element of the ASP-HEI cycle operates, and at each site inequity is measurably produced rather than merely reflected: the pipeline is the cycle instantiated in code.

Stages 1–2 — Historical inequity becomes formal decision rule. The cycle predicts that algorithmic decision-making learns from a history shaped by pre-existing inequity and converts that diffuse condition into a coded claim [2]. Baseline gaps of 14–18 percentage points in completion between female and male domestic students and between international and domestic populations are not errors in the data; they are the college's institutional history, and an EWS that optimizes for accuracy is asked to preserve it. The conversion to formal claim happens concretely in the model: false-positive rates for domestic students aged 36–40 exceed 0.60, and the model encodes age as a proxy for risk without any causal account of why. Advisors inherit these classifications without access to the underlying reasoning and cannot contest them—exactly the loss of discretionary override the original ethnography attributed to the automation of the faculty–student relationship.

Stage 3 — Post-processing as institutional policy. This is our principal extension of the framework. The cycle argues that algorithmic decisions function as policy once implemented; we identify *percentile-based post-processing* as the specific, locatable mechanism through which this transition happens. Percentile thresholds collapse probabilities from 0.41 to 0.80 into a single "Medium Risk" bin—the bin with the worst calibration—and widen the male–female High Risk gap from 36% to 40%. This gap is not produced by the model's

predictive errors; it is produced by the decision to convert a continuous score into three fixed-quota intervention tiers. A defensible ranking becomes an indefensible allocation rule.

5.1. Three Mechanisms of the Cycle

Locating the mechanism sharpens *why* technical fairness interventions cannot on their own break the cycle. Our findings identify three mechanisms through which the cycle's inequity-exacerbation effect operates, each sharing a common structure: a property of the institution or its tooling is relabeled as a property of the student.

Task formulation mismatch. The EWS predicts dropout; advisors interpret its outputs as indicators of who would benefit from support. These are distinct constructs—students drop out for reasons unrelated to academic need (transfer, finances, personal circumstances), and students who persist may still struggle. The mismatch is not a technical oversight; "who needs help" has quietly been replaced by "who threatens retention metrics."

Institutional priorities versus student-centered success. The choice of program non-completion as target reflects the cycle's driving force. The EWS was implemented during funding cuts and tuition freezes [2], and its operational definition of "success" follows directly. Students and advisors may understand success as transferable skills, appropriate career paths, or wellbeing—none of which are observable in the institution's student information system, and none of which make it into the model.

Prediction uncertainty relabeled as moderate need. The Medium Risk bin is where the model is least confident; practitioners treat it as where students are moderately in need. Model uncertainty is thus laundered into intervention intensity. This is the cycle's characteristic move applied to epistemic rather than demographic content: a hidden institutional property (how much the model does not know) is recoded as a student property (how much support the student requires).

6. Implications, Limitations, and Conclusion

Historical marginalization enters as training-data disparity, is formalized by the model as decision rules, and is institutionalized by post-processing as allocation policy—and at each stage the institution can intervene, if it chooses to.

This reframes what trustworthy ML in public higher education needs to look like. Audits should not stop at aggregate statistical metrics; they must trace how disparity is transformed across stages and interrogate whether the predicted target is the construct the institution claims to care about. Concretely, the post-processing amplification we document points to several pipeline-level interventions worth investigating: replacing fixed-quota percentile bins with calibrated probability thresholds, applying group-conditional calibration [13], or reformulating the task as multi-objective optimization that balances retention prediction against allocation parity. None of these substitute for revisiting the target definition, but each would attenuate the specific amplification mechanism we identify. Our three mechanisms (task formulation mismatch, institutional-priority drift, and uncertainty relabeled as moderate need) generalize beyond the EWS we audit and can serve as a checklist for auditing other institutional risk models.

Our analysis is based on a replica trained on the institution's data and documented design protocol, not the production model itself; fidelity is high but not perfect. We lack data on actual intervention delivery or student responses, so the audit evaluates predictions and categorizations rather than downstream effects, and the outcome label conflates dropout, transfer, and program change—a construct-validity concern we foreground rather than a flaw we can correct. Future work should extend this pipeline-level interrogation to the other elements of the ASP-HEI cycle for which we currently have only ethnographic evidence, and pursue longitudinal evaluation of downstream impact on the students the model governs.

Acknowledgments

We would like to thank our collaborators and study participants at Centennial College for allowing us to conduct this work. Additionally, we thank the anonymous reviewers whose suggestions and comments helped improve this manuscript. Generative AI tools were used as an editing aid during manuscript preparation; all intellectual content, analysis, and conclusions are solely our own.

References

[1] K. McConvey, M. Ghai, R. Lee, and S. Guha. "Risk, Retention, and the Algorithmic Institution: Artificial Intelligence as a Policy Response to Higher Education in Crisis". In: *Canadian Public Policy / Analyse de politiques* (2026). DOI: 10.3138/cpp.2025-030.

[2] K. McConvey and S. Guha. ""This Is Not a Data Problem": Algorithms and Power in Public Higher Education in Canada". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024, pp. 1–14.

[3] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2020, pp. 33–44. DOI: 10.1145/3351095.3372873.

[4] D. Saxena, K. Badillo-Urquiola, P. J. Wisniewski, and S. Guha. "A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare". In: *Proceedings of the ACM on Human-Computer Interaction*. Vol. 5. CSCW2. 2021, pp. 1–41. DOI: 10.1145/3476089.

[5] E. S. Y. Moon and S. Guha. "A Human-Centered Review of Algorithms in Homelessness Research". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024, pp. 1–15. DOI: 10.1145/3613904.3642392.

[6] K. Simbeck. "They Shall Be Fair, Transparent, and Robust: Auditing Learning Analytics Systems". In: *AI and Ethics* 4.2 (2024), pp. 555–571.

[7] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations". In: *Science* 366.6464 (2019), pp. 447–453. DOI: 10.1126/science.aax2342.

[8] S. Passi and S. Barocas. "Problem Formulation and Fairness". In: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2019, pp. 39–48. DOI: 10.1145/3287560.3287567.

[9] A. Z. Jacobs and H. Wallach. "Measurement and Fairness". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2021, pp. 375–385. DOI: 10.1145/3442188.3445901.

[10] J. C. Perdomo, T. Britton, M. Hardt, and R. Abebe. "Difficult Lessons on Social Prediction from Wisconsin Public Schools". In: *arXiv preprint* (2023). arXiv: 2304.06205.

[11] K. McConvey, S. Guha, and A. Kuzminykh. "A Human-Centered Review of Algorithms in Decision-Making in Higher Education". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2023, pp. 1–15. DOI: 10.1145/3544548.3580658.

[12] M. A. Kabir, M. U. Ahmed, S. Begum, S. Barua, and M. R. Islam. "Balancing Fairness: Unveiling the Potential of SMOTE-Driven Oversampling in AI Model Enhancement". In: *Proceedings of the 2024 9th International Conference on Machine Learning Technologies*. ICMLT '24. New York, NY, USA: Association for Computing Machinery, Sept. 2024, pp. 21–29. ISBN: 979-8-4007-1637-9. DOI: 10.1145/3674029.3674034.

[13] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. "Fairness through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. Association for Computing Machinery, 2012, pp. 214–226. DOI: 10.1145/2090236.2090255.

Appendix A. Supplementary Tables and Figures

This appendix contains detailed fairness metrics and supplementary visualizations supporting the findings reported in the main text. Figure 1 shows the bimodal prediction distribution and risk-category thresholds; Tables 2 and 3 give pairwise fairness metrics across age groups for domestic and international students respectively; Table 4 reports accuracy by risk level and age group.

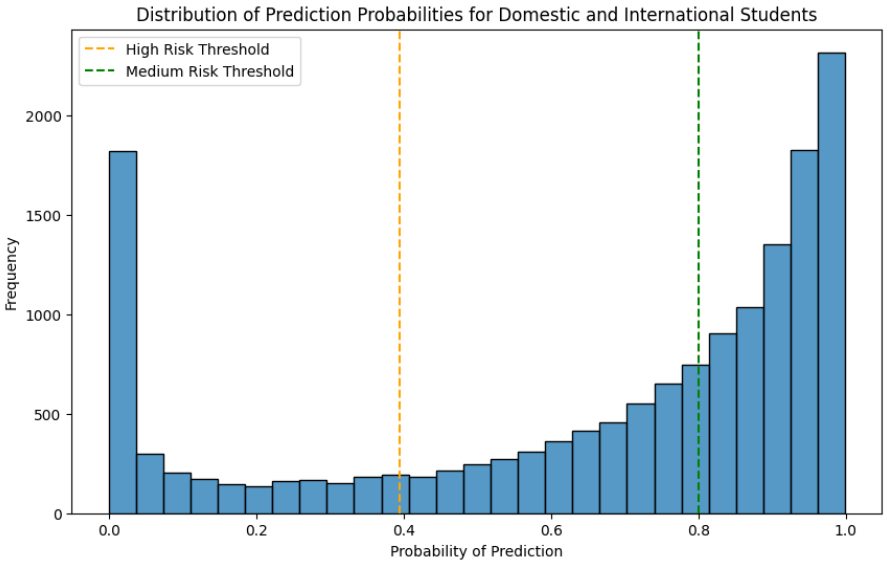

Figure 1. Distribution of success prediction probabilities from the EWS. The orange dashed line (≈0.4) marks the High Risk threshold and the green dashed line (≈0.8) marks the Medium Risk threshold.

Table 2. Domestic fairness metrics across age groups (selected pairs).

Age Groups	**SPD**	**EOD**	**AOD**	**DI**
21 vs 36	-0.1710	-0.1660	-0.0855	0.8034
21 vs 31	-0.1286	-0.1447	-0.0643	0.8446
21 vs 41	-0.1665	-0.1551	-0.0833	0.8076
21 vs 19	0.1669	0.1589	0.0834	1.3137
21 vs 66	-0.3012	-0.3971	-0.1506	0.6988
36 vs 19	0.3378	0.3249	0.1689	1.6351
31 vs 19	0.2954	0.3035	0.1477	1.5554
41 vs 19	0.3334	0.3140	0.1667	1.6268
19 vs 66	-0.4681	-0.5560	-0.2340	0.5319

Table 3. International fairness metrics across age groups (selected pairs).

Age Groups	**SPD**	**EOD**	**AOD**	**DI**
26 vs 36	-0.0575	-0.0402	-0.0287	0.9401
26 vs 19	0.1844	0.2069	0.0922	1.2568
21 vs 36	-0.1050	-0.0855	-0.0525	0.8906
21 vs 19	0.1369	0.1616	0.0684	1.1906
36 vs 19	0.2419	0.2471	0.1209	1.3369
31 vs 19	0.1771	0.1993	0.0886	1.2467
19 vs 41	-0.2310	-0.2108	-0.1155	0.7566
19 vs 51	-0.2821	-0.2442	-0.1410	0.7179

Table 4. Accuracy by risk level and age group for domestic and international models, with differences.

Risk	**Age**	**Dom.**	**Intl.**	**Diff.**
High	0	0.81	0.73	0.08
	19	0.81	0.79	0.02
	21	0.84	0.72	0.12
	26	0.84	0.82	0.02
	31	0.77	0.62	0.15
	36	0.65	0.60	0.05
	41	0.77	0.57	0.19
Medium	0	0.70	0.83	-0.14
	19	0.71	0.80	-0.09
	21	0.72	0.83	-0.11
	26	0.74	0.79	-0.05
	31	0.83	0.80	0.02
	36	0.70	0.80	-0.10
	41	0.77	0.73	0.04
Low	0	0.91	0.97	-0.06
	19	0.90	0.97	-0.07
	21	0.92	0.99	-0.07
	26	0.94	0.98	-0.04
	31	0.91	0.99	-0.07
	36	0.93	0.98	-0.05
	41	0.92	0.95	-0.03