Human-Computer Interaction & Emerging Technologies, Vol. 233, 2026, 373–382
https://doi.org/10.54941/ahfe1007537

Eliciting Fairness via Micro-Ethics Embedded Interfaces for Machine Learning Workflows

Wangfan Li and Carlos Toxtli-Hernández

Human-AI Empowerment Lab, Clemson University, Clemson, SC 29634, USA

ABSTRACT

Automated and no-code ML tools make model building accessible but can obscure harms that arise when users include sensitive attributes. We embed micro-ethics nudges at key workflow moments and evaluate downstream fairness outcomes and user experience. In a between-subjects experiment (N = 34) participants used a simplified AutoML web tool on a subset of the HMDA mortgage dataset with a 10-minute modeling task. The intervention combined in text notice when selecting sensitive attributes such as race, gender, ethnicity and post-training model explanation visualizations, while the base condition showed only model performance metrics. Participants in the Intervention group included significantly fewer sensitive features and produced models with substantially smaller equal-opportunity gaps, while System Usability Scale scores did not differ significantly across conditions. Moral acceptability did not significantly differ between conditions, though it trended lower under the intervention. We conclude that minimal, well timed fairness feedback can meaningfully reduce bias in rapid prototyping workflows. We also discuss design patterns for embedding fairness and the implications of increased moral sensitivity for tool adoption.

Keywords: Explainable AI, Human-centered AI, Algorithmic fairness

INTRODUCTION

Machine learning systems are increasingly deployed to make consequential decisions about people's lives, from determining credit worthiness and employment opportunities to influencing criminal sentencing and healthcare allocation (Barocas and Selbst, 2016). While these systems promise greater efficiency and consistency than humans, they also risk perpetuating and amplifying biases present in historical data (Angwin et al., 2016). Recent investigations have revealed discriminatory patterns in commercial lending algorithms that deny loans to qualified minority applicants at higher rates than similarly situated white applicants (Bartlett et al., 2022). These documented harms have shown a need for more responsible practices that consider fairness alongside traditional performance metrics. The proliferation of automated machine learning (AutoML) platforms and no-code tools has made access to ML capabilities significantly easier, enabling practitioners without formal data science training to build and deploy models (Amershi et al., 2014). Platforms like Google AutoML, Amazon SageMaker, and Microsoft Azure ML for example, allow users to create sophisticated models

Received February 7, 2026; Revised March 28, 2026; Accepted April 13, 2026; Available online July 20, 2026

through visual interfaces only, abstracting away technical complexity. While this democratization expands who can use ML technology, it also means that many model builders lack awareness of potential bias issues or knowledge of fairness-enhancing techniques (Holstein et al., 2019). These practitioners often work under time pressure to deliver functional models, with success measured primarily through accuracy metrics. The combination of limited expertise, time constraints, and performance-focused incentives creates conditions where fairness considerations are easily overlooked.

Existing approaches to promoting fairness in ML development face significant adoption challenges. Comprehensive fairness toolkits like IBM's AI Fairness 360 citepbellamy2019ai and Microsoft's Fairlearn citepbird 2020fairlearn provide sophisticated bias detection and mitigation, but require statistical expertise and familiarity with fairness concepts that many practitioners lack. Visual analytics systems for fairness (Cabrera et al., 2019) enable detailed exploration of bias, but assume users have time for extended analysis and can interpret complex visualizations. Automated fairness constraints citepagarwal2018reductions remove human judgment from the process, potentially create misalignment with domain-specific requirements. These tools, while powerful, exist outside standard workflows and require practitioners to actively seek them out, learn new interfaces, and integrate back into their development process.

This paper investigates whether lightweight, embedded interventions can promote fairness-aware model building without requiring expertise or disrupting existing workflows. Specifically, we examine whether combining just-in-time warnings when selecting potentially problematic features with simple post-hoc visualizations of model bias can nudge users toward fairer modeling choices. This approach draws on behavioral insights from HCI research showing that small environmental changes can trigger more deliberative decision-making citepbuccinca2021trust,kim2024m.

To evaluate this approach, We conducted a controlled study (N=34) in an AutoML-like, time-limited loan-modeling task using HMDA data. Participants either saw only performance metrics (accuracy/F1) or additionally received a sensitive-attribute warning and post-training disparity visualization. The intervention reduced inclusion of sensitive features and lowered demographic disparities, without significantly degrading usability. Perceived moral acceptability did not significantly differ, but trended lower under the intervention.

This work contributes: (1) an empirical evaluation showing that micro-ethics nudges embedded at key workflow decision points reduce sensitive-feature use and downstream disparity without harming usability, and (2) evidence of an awareness–confidence paradox, in which increased fairness feedback improves objective outcomes while introducing greater ethical caution, motivating design implications for fairness interventions that inform without inducing uncertainty.

BACKGROUND

Visual analytics systems have emerged as a promising approach to make fairness concepts more accessible. One approach (Cabrera et al., 2019)

uses visual encodings to reveal intersectional bias patterns, allowing users to explore how multiple protected attributes interact, while a causal graph approach to help users understand sources of bias and test interventions is also shown to be effective (Yan et al., 2020). FairSight (Ahn and Lin, 2019) provides rankings and clustering visualizations to identify fairness issues across subgroups. These systems demonstrate the value of visual representations for understanding complex fairness relationships. However, they assume users can interpret sophisticated visualizations and have time for extended exploration. Kaur et al. (Kaur et al., 2020) found that even data scientists struggle to correctly interpret ML interpret-ability tools, suggesting simpler approaches may be needed for non-experts. Our work employs basic bar charts and binary warnings, prioritizing immediate comprehension over analytical depth. Visual tools that present fairness metrics and potential harms in a graphic and intuitive format can also efficiently steer users away from biased deicions during model development, where interactive causal graphs and fairness panels can enable users to quickly identify and mitigate sources of social bias within the models (Yan et al., 2020).

Recent research in human computer interaction also highlights the effectiveness of low-burden interventions for guiding non-expert users toward better behavioral choices without negatively affecting overall task performance. These interventions are often designed to be lightweight and relevant to specific context, aiming to nudge user decisions at key points in the modeling workflow. Simple or timed prompts can significantly alter user behavior, where brief uncertainty cues from AI models can reduce users' blind agreement with AI outputs and improved their accuracy (Kim et al., 2024). It is also found that short cognitive forcing prompts at decision time could decrease over-reliance on AI recommendations, resulting in more critical engagement without adding extra cognitive burden (Buçinca et al., 2021).

METHODOLOGY

Study Design

We conducted a between-subjects online experiment to investigate whether embedding fairness-oriented interventions in a model-building interface influences non-expert users' modeling decisions and perceptions of the resulting models. Participants were randomly assigned to one of two conditions (baseline vs. intervention). The study addressed the following research questions:

RQ1: Do lightweight fairness interventions reduce the inclusion of sensitive features in machine learning models built by non-experts?

RQ2: Do these interventions lead to models with smaller disparities across demographic groups, as measured by equal opportunity gap?

RQ3: How do fairness interventions affect users' perceptions of model acceptability and system usability?

Based on prior work on behavioral nudges and value-sensitive design, we formulated the following hypotheses:

H1: Participants exposed to fairness interventions will include fewer sensitive features (race, gender, ethnicity) in their final models compared to baseline.

H3: Models built with fairness interventions will exhibit smaller equal opportunity gaps than those built without such interventions.

H3: The fairness interventions will not significantly reduce system usability, as lightweight nudges can guide behavior without adding substantial cognitive burden.

H4: Participants exposed to fairness interventions will report higher moral acceptability for their models, reflecting increased confidence in ethical decision-making.

The experiment simulated a time-constrained model development scenario using a subset of 10,000 records from the Home Mortgage Disclosure Act (HMDA) dataset, which contains loan decisions along with applicant demographics and financial information. This dataset was selected because (1) lending decisions have clear fairness implications with legal protections for demographic groups; (2) the domain is accessible to non-experts; and (3) the presence of both sensitive attributes (race, gender, ethnicity) and financial features creates realistic feature trade-offs. Participants were tasked with building the best-performing loan approval prediction model within a 10-minute time limit, reflecting rapid prototyping common in AutoML environments. Only the feedback mechanisms differed between conditions; all participants used the same dataset, algorithms, and performance metrics.

Experimental Manipulation

The study employed a single between–subjects independent variable where participants were randomly assigned to one of two conditions:

- **Baseline** The interface displayed standard performance feedback after training and validation (overall accuracy and F1 score).
- **Intervention Group** In addition to the baseline performance feedback, the interface provided (1) a just-in-time warning when users selected sensitive attributes and (2) a post-training visualization intended to highlight model impact.

Interactive Model-Building Tool

We developed a custom web-based interface using React (front end) and Python/Flask (back end). The tool was designed to resemble simplified AutoML platforms while maintaining experimental control. The workflow included:

Feature Selection Panel: Participants selected features via checkboxes from a list including non-sensitive attributes (e.g., loan amount, income, debt-to-income ratio, loan-to-value ratio, dwelling category) and sensitive attributes (race, gender, ethnicity). In the intervention condition, selecting a sensitive attribute triggered a non-blocking warning displayed beneath the selection area: "Caution: You have selected training fields that can potentially introduce

bias into the model." Participants could proceed without responding to the warning.

Algorithm Selection and Training: Participants chose one of three scikit-learn classifiers: Logistic Regression, Random Forest (100 estimators, max depth 10), or Gradient Boosting (100 estimators, learning rate 0.1). Participants trained models by clicking the "Train Model" button.

Performance Display: After training, the system performed 5- fold cross-validation and displayed accuracy and F1 score. The baseline condition displayed only performance metrics and example predictions. The intervention condition additionally displayed an a simple disparity chart showing true positive rate for each demographic group (race/gender/ethnicity), presented separately from the performance overview.

Measurements

Number of sensitive features chosen: We recorded the number of sensitive attributes included as training features in each participant's submitted model. This count ranged from 0 to 3 (race, gender, ethnicity).

Equal Opportunity Gap: We measured equal opportunity gap (EOG) as differences in true positive rates (TPR) across demographic groups. For each sensitive attribute, we calculated group TPRs on the test set using a 0.5 threshold. For binary attributes, the attribute EOG was the signed difference in TPR. For multi-category attributes, we computed the maximum pairwise absolute difference in TPR across observed groups. To obtain a single fairness value per participant model, we took the maximum absolute EOG across the three sensitive attributes, producing one absolute EOG per model.

Moral Acceptability: We measured perceived ethical acceptability using a four-item moral acceptability questionnaire assessing whether participants found their submitted model acceptable for real-world loan decisions. Items were rated on a 7-point Likert scale (1 = Strongly disagree, 7 = Strongly agree) and averaged into a composite score.

System Usability Scale: We assessed perceived usability using the System Usability Scale (SUS). Following standard scoring (reverse-coding even-numbered items, summing adjusted responses, and multiplying by 2.5), we obtained a single usability score from 0 to 100.

Procedure

The study was conducted on Prolific and approved by the Institutional Review Board. After informed consent, participants completed a demographics survey and were randomly assigned to condition. Participants were instructed to build the best-performing model (highest accuracy and F1) by iterating over features and model types within a 10-minute time limit. After the task, participants completed post-task questions assessing usability and moral acceptability and answered two open-ended questions. The system automatically recorded sensitive-feature count and EOG. The study took approximately 20 minutes. We recruited 48 participants from the United States and assigned them equally across conditions. Participants were

required to have at least some college education to support comprehension of the task. Fourteen participants were excluded for failing attention checks or not following task instructions, leaving 34 participants for analysis.

RESULTS

For the number of the three sensitive features included in models across both conditions, participants in the baseline condition (n = 17) included significantly more sensitive features (M = 2.35, Mdn = 3, IQR = 2–3) compared to those in the intervention condition (n = 17, M = 1.12, Mdn = 1, IQR = 0–2). A Mann–Whitney U test was used to confirmed that the difference is statstically significant, U = 57.0, z = –3.11, p = 0.0018. Showing that for the final model submission, the ones in the base condition included significantly more sensitive attributes.

We compared the *equal opportunity gap* produced by participants' final models across both conditions. Participants in the baseline condition (n = 17) yielded markedly larger gaps (M = 0.142, Mdn = 0.120, IQR = 0.090–0.180) than those in the intervention condition (n = 17, M = 0.047, Mdn = 0.040, IQR = 0.020–0.070). A Mann–Whitney U test confirmed that this difference was statistically significant, U = 58.0, z = –2.98, p = 0.0029, Cliff's δ = –0.60 (large). These results indicate that models built with intervention shows significantly smaller equal opportunity gaps than those built under the baseline condition.

For SUS scores across conditions. Participants in the baseline condition (n = 17) reported slightly higher usability (SUS: M = 72.1, Mdn = 74, IQR = 65–80) than those in the intervention condition (n = 17; M = 68.0, Mdn = 66, IQR = 60–75). An independent-samples t-test showed that this difference was not statistically significant, $t(32)$ = 1.12, p = 0.27, Cohen's d = 0.38. Levene's test indicated equal variances, F (1, 32) = 0.45, p = 0.51. For the Moral Acceptability score, participants in the baseline condition (n = 17) reported moderately low acceptability for their final models (M = 3.60, Mdn = 3.50, IQR = 3.00–4.20). Participants in the intervention condition (n = 17) judged their models as slightly *less* acceptable (M = 3.05, Mdn = 3.00, IQR = 2.50–3.60). An independent-samples t-test showed that this difference approached significance, $t(32)$ = 1.74, p = 0.09, Cohen's d = 0.60 (medium). Levene's test indicated equal variances, F (1, 32) = 0.18, p = .68. Thus, although moral acceptability tended to be lower under the transparency intervention, the effect did not reach significance.

DISCUSSION

Despite improved fairness metrics, participants in the intervention group did not report higher moral acceptability for their models. Instead, moral acceptability ratings trended lower in the intervention condition (though this difference was not statistically significant at α = .05). One possible explanation is that increased visibility into model disparities heightened users' uncertainty about their final model choice. This aligns with prior findings that explanations or fairness feedback which surface bias can lower

perceived fairness and trust by drawing attention to imperfections rather than reassuring users (Goyal et al., 2024; Gaba et al., 2023).

Addressing Research Questions and Hypotheses

Regarding RQ1, our findings strongly demonstrate that minimal fairness interventions reduce the inclusion of sensitive features in models built by non-experts. Participants in the intervention condition included significantly fewer sensitive attributes (M = 1.12) compared to baseline participants (M = 2.35), with a large effect size. This supports H1 and suggests that even simple warnings at the point of feature selection can prompt users to reconsider their inclusion of demographic attributes. Qualitative data reinforce this interpretation, with several intervention participants explicitly mentioning that warnings influenced their feature selection decisions. This finding is particularly encouraging given the time-constrained nature of the task, indicating that fairness considerations can be activated even under pressure for rapid model development, consistent with research on time- pressured ethical decision-making (Shalvi et al., 2012).

For RQ2, models built with fairness interventions exhibited substantially smaller equal opportunity gaps (M = 0.047) compared to baseline models (M = 0.142), confirming H2. The large effect size (Cliff's $\delta = -0.60$) suggests that the behavioral changes induced by the interventions translated into meaningful improvements in model fairness (Romano et al., 2006).

RQ3 asked how fairness interventions affect user experience. Consistent with H3, we found no significant difference in system usability between conditions, with both groups reporting SUS scores in the acceptable range (baseline: M = 72.1; intervention: M = 68.0). This suggests that fairness interventions need not compromise the streamlined user experience that makes AutoML platforms attractive to non-experts (Wang et al., 2019).

H4, however, was not supported. Participants exposed to fairness information reported lower moral acceptability for their models (M = 3.05) compared to baseline participants (M = 3.60), but this difference was not statistically significant ($p = .09$). Rather than treating this as a simple null result, we interpret it as evidence of a shift in ethical stance. The pattern suggests that fairness awareness may increase users' moral scrutiny of their models, making them more cautious rather than more confident.

The Paradox of Moral Acceptability

The finding that users in the intervention condition produced fairer models but reported lower moral acceptability deserves careful consideration. This apparent paradox may reflect a desirable increase in moral sensitivity, where exposure to fairness information makes users more aware of the ethical complexities inherent in automated decision-making. Rather than indicating intervention failure, lower or unchanged moral acceptability scores may suggest that users are developing a more nuanced understanding of fairness challenges.

This interpretation aligns with research on moral dumbfounding (Haidt, 2001), where increased awareness of ethical issues can reduce confidence in moral judgments even when behavior improves. The intervention may have shifted users from unreflective confidence to reflective uncertainty, a potentially valuable outcome for responsible AI development. Future work should explore whether this uncertainty motivates continued learning about fairness and prompts further refinement of models, or whether it risks leading to decision paralysis or disengagement.

Design Implications

Our findings reveal important considerations for integrating fairness interventions into machine learning development tools. The effectiveness of minimal interventions in reducing bias, combined with the complex relationship between fairness awareness and user confidence, suggests several design principles for future systems.

Timing and Placement of Interventions

The success of just-in-time warnings during feature selection indicates that fairness considerations are most effective when embedded at decision points rather than presented as separate documentation or post-hoc analysis. Designers should identify critical junctures in the ML workflow—such as feature engineering, algorithm selection, and threshold setting—and surface relevant fairness information precisely when users make these choices. However, our qualitative findings reveal that performance metrics remain the primary decision driver, with fairness serving as a secondary consideration. This suggests that interventions should complement rather than replace performance feedback, presenting fairness and accuracy as parallel objectives that users can explicitly balance.

Managing the Awareness–Confidence Paradox

The decreased moral-acceptability ratings among intervention participants, despite producing fairer models, reveal an important factor in fairness-aware design. Making bias and disparities visible appears to appropriately increase users' uncertainty about algorithmic decision-making, shifting them from unreflective confidence to critical engagement. Rather than viewing this as a failure, designers should recognize productive uncertainty as a desirable outcome. Interfaces might explicitly acknowledge the inherent tensions in fairness, presenting trade-offs transparently rather than suggesting that any model can be perfectly fair. This could include confidence intervals for fairness metrics, comparisons to baseline disparities in historical data, or explicit statements about which fairness criteria cannot be simultaneously satisfied.

Limitation

There are several important limitations to consider regarding our findings. First, our experiment involved a small sample of non-expert participants

(N = 34) working on a short, time-constrained modeling task using a subset of the HMDA dataset. The five-minute task was chosen to simulate rapid prototyping scenarios typical of AutoML workflows, but this artificial time pressure can place extra constrains outside of a normal work environment. Our prototype interface and laboratory setting also simplifies real AutoML platforms and organizational practices. The intervention delivered lightweight, in-context cues rather than complex systems. In practice, adoption will depend on how interventions interact with workflow and team practices. Field deployments and longitudinal studies are needed to show whether this type of intervention can produce sustained behavior change.

CONCLUSION

This paper examined whether lightweight, embedded fairness interventions can steer non-expert model builders toward fairer outcomes in rapid AutoML-style workflows. In a controlled loan-modeling task using HMDA data (N = 34), participants who received a sensitive-attribute warning and post-training disparity visualization selected fewer sensitive features and produced models with significantly smaller equal opportunity gaps, without a significant decrease in usability. Moral acceptability did not significantly differ between conditions, though it trended lower under the intervention, suggesting that making disparities visible may increase ethical caution. Overall, our results indicate that minimal, well-timed fairness feedback can reduce bias in time-constrained prototyping and motivate design patterns that support fairness awareness without derailing workflow efficiency.

REFERENCES

Ahn, Y. and Lin, Y.-R. (2019), FairSight: Visual analytics for fairness in decision making, *in* 'IEEE Transactions on Visualization and Computer Graphics', Vol. 26, IEEE, pp. 1086–1095.

Amershi, S., Cakmak, M., Knox, W. B. and Kulesza, T. (2014), 'Power to the people: The role of humans in interactive machine learning', *AI Magazine* **35**(4), 105–120.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016), 'Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks', https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, S. and Selbst, A. D. (2016), 'Big data's disparate impact', *Calif. L. Rev.* **104**, 671.

Bartlett, R., Morse, A., Stanton, R. and Wallace, N. (2022), 'Consumer-lending discrimination in the FinTech era', *Journal of Financial Economics* **143**(1), 30– 56.

Buçinca, Z., Malaya, M. B. and Gajos, K. Z. (2021), 'To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making', *Proceedings of the ACM on Human-computer Interaction* **5**(CSCW1), 1–21.

Cabrera, Á. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J. and Chau, D. H. (2019), FairVis: Visual analytics for discovering intersectional bias in machine learning, *in* '2019 IEEE Conference on Visual Analytics Science and Technology (VAST)', IEEE, pp. 46–56.

Gaba, A., Kaufman, Z., Cheung, J., Shvakel, M., Hall, K. W., Brun, Y. and Bearfield, C. X. (2023), 'My model is unfair, do people even care? visual design affects trust

and perceived bias in machine learning', *IEEE transactions on visualization and computer graphics* **30**(1), 327–337.

Goyal, N., Baumler, C., Nguyen, T. and Daum´e III, H. (2024), The impact of explanations on fairness in human-ai decision-making: Protected vs proxy features, *in* 'Proceedings of the 29th International Conference on Intelligent User Interfaces', pp. 155–180.

Haidt, J. (2001), 'The emotional dog and its rational tail: A social intuitionist approach to moral judgment', *Psychological Review* **108**(4), 814–834.

Holstein, K., Wortman Vaughan, J., Daum´e III, H., Dud´ık, M. and Wallach, H. (2019), Improving fairness in machine learning systems: What do industry practitioners need?, *in* 'Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems', pp. 1–16.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H. and Wortman Vaughan, J. (2020), Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning, *in* 'Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems', pp. 1–14.

Kim, S. S., Liao, Q. V., Vorvoreanu, M., Ballard, S. and Vaughan, J. W. (2024), " i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust, *in* 'Proceedings of the 2024 ACM conference on fairness, accountability, and transparency', pp. 822–835.

Romano, J., Kromrey, J. D., Coraggio, J. and Skowronek, J. (2006), 'Appropriate statistics for ordinal level data: Should we really be using t-test and cohen's d for evaluating group differences on the nsse and other surveys', *Annual Meeting of the Florida Association of Institutional Research* **177**, 34.

Shalvi, S., Eldar, O. and Bereby-Meyer, Y. (2012), 'Ethical maneuvering: Why people avoid both major and minor lies', *British Journal of Management* **23**, S65–S79.

Wang, D., Yang, Q., Abdul, A. and Lim, B. Y. (2019), Designing theory-driven user-centric explainable ai, *in* 'Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems', pp. 1–15.

Yan, J. N., Gu, Z., Lin, H. and Rzeszotarski, J. M. (2020), Silva: Interactively assessing machine learning fairness using causality, *in* 'Proceedings of the 2020 chi conference on human factors in computing systems', pp. 1–13.