Bidirectional Human–AI Collaboration for Equitable Student Performance Prediction via Deep Uncertainty Learning

Ruohan Zong[1] , **Yang Zhang**[1] , **Lanyu Shang**[2] , **Frank Stinar**[1] , **Nigel Bosch**[1] , **Dong Wang**[1]
[1]University of Illinois Urbana-Champaign
[2]Loyola Marymount University
{rzong2, yzhangnd}@illinois.edu, Lanyu.Shang@lmu.edu, {fstinar2, pnb, dwang24}@illinois.edu

Abstract

This paper studies a bidirectional human–AI collaborative student performance prediction problem to enhance equitable online education, aligning with the United Nations' Sustainable Development Goal (SDG) of ensuring inclusive and equitable quality education for all. The goal is to leverage collaborative intelligence to generate accurate and fair student outcome predictions from behavioral data, ensuring equitable estimation for underrepresented populations. Current fair AI solutions often fail to mitigate demographic bias in the absence of student demographic data, while human–AI collaborative approaches frequently overlook human cognitive biases, leading to inaccurate predictions. We develop CollabDebias, a novel bidirectional human–AI collaborative framework that utilizes the complementary strengths of AI and humans to mitigate the AI demographic bias and human cognitive bias. To address AI demographic bias, we propose an uncertainty learning-based bias identification method and a reliability-aware human–AI integration approach. To reduce human cognitive bias, we design uncertainty-aware visualization of AI decision area and attention mechanism. Experimental results on an online course demonstrate CollabDebias's effectiveness in improving student performance prediction accuracy and fairness.

1 Introduction

Emerging advances in online technologies, such as open educational platforms and AI-driven personalized learning systems, offer transformative opportunities to advance the United Nations' Sustainable Development Goal (SDG) of inclusive and equitable quality education [Ayeni *et al.*, 2024]. Integrating AI and human intelligence (HI) into online learning platforms is a pivotal approach to enhancing learning experiences and outcomes [Wang *et al.*, 2022]. This paper studies student performance prediction in online education [Waheed *et al.*, 2020], aiming to predict students' final grades (e.g., Fail, Pass, Distinction) using behavioral data (e.g., reading materials, completing quizzes). However, AI predictions often suffer from biases that disproportionately affect vulnerable populations, such as underrepresented older students who often face additional barriers in online learning. We propose a novel bidirectional human–AI collaborative framework in which diverse stakeholders—students, crowdsourced annotators, and AI fairness researchers—actively contribute to mitigating both AI demographic bias and human cognitive bias. This approach aims to generate accurate and fair AI predictions to assist students in developing accurate and unbiased self-assessment through AI-assisted self-reflection. By actively incorporating multi-disciplinary experts in both education and AI fairness domains into our team, this work advances state-of-the-art human–AI collaboration technologies in online education. This approach fosters equitable and lifelong learning opportunities for diverse populations, with a special focus on supporting underrepresented groups.

Emerging advances in AI have demonstrated impressive accuracy and scalability in student performance prediction [Waheed *et al.*, 2020; Qiu *et al.*, 2022]. However, current AI solutions often struggle to produce both accurate and fair predictions across demographic groups due to inherent biases in online education (referred to as **AI demographic bias**) [Kizilcec and Lee, 2022], particularly when sensitive demographic data is unavailable due to privacy concerns [Nguyen *et al.*, 2023]. The left side of Figure 1 illustrates AI demographic bias in age groups: the model incorrectly predicts an older student will achieve a Distinction grade, based on trends from younger, traditional students, while overlooking the additional efforts often required by older, non-traditional students. Such inaccurate predictions can mislead underrepresented students, causing them to either underprepare or overcommit efforts, exacerbating their existing disadvantages in online education [Kizilcec and Lee, 2022]. The fair AI solutions have been proposed to mitigate AI demographic bias by re-weighting underrepresented samples or adding fairness regularizations [Kini *et al.*, 2021]. However, these methods typically require access to student demographic information, which is often unavailable due to privacy concerns in educational applications [Kizilcec and Lee, 2022]. In this paper, we focus on the more challenging scenario of addressing AI demographic bias *without access to sensitive demographic data*, ensuring fair AI-driven feedback in privacy-aware settings [Nguyen *et al.*, 2023].

HI has been used to enhance AI predictions through high-

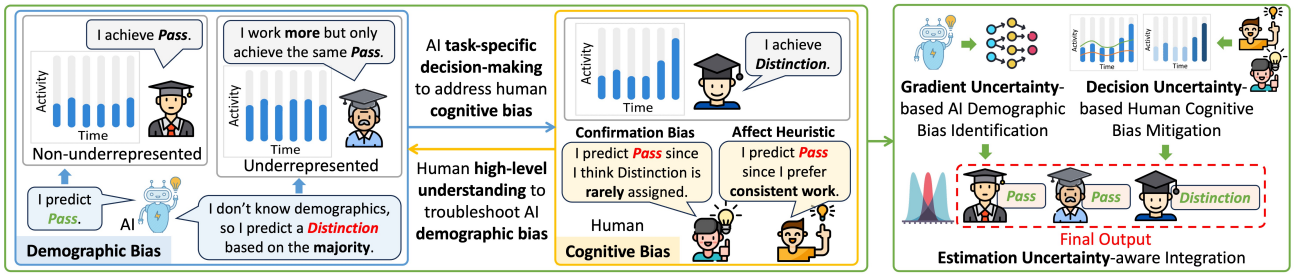


Figure 1: Illustration of addressing AI demographic bias and human cognitive bias by their complementary strengths.

level understanding (e.g., common-sense reasoning, contextual understanding) in tasks like language generation [OpenAI, 2023] and image classification [Budd *et al.*, 2021]. In student performance prediction, we propose leveraging HI to mitigate AI demographic bias by guiding humans to predict final grades based on general learning activity patterns that humans can easily recognize (e.g., consistent work, extra effort before the final), which are less affected by the lack of underrepresented training data [Zhang *et al.*, 2023]. However, current human–AI collaborative approaches often overlook **human cognitive bias**, leading to inaccurate predictions [Draws *et al.*, 2022]. Specifically, we identify two key cognitive biases in our problem [Draws *et al.*, 2021]: 1) Confirmation bias, where humans favor outcomes align with their preexisting beliefs (e.g., predicting more Pass grades and fewer Fail or Distinction grades), and 2) Affect heuristic, where predictions are influenced by immediate emotional reactions to study behaviors (e.g., preferences towards consistent work vs. extra work before the final), causing inconsistent/inaccurate outcome predictions, as illustrated in Figure 1.

Motivated by the limitations of current AI, fair AI, and human–AI methods, we propose a **bidirectional human–AI collaborative solution** to address both AI demographic bias and human cognitive bias by leveraging their complementary strengths. There are two key technical challenges:

The first challenge is effectively *addressing AI demographic bias without access to sensitive student demographic data*. Current fair AI solutions, such as sample re-weighting or fairness regularization, rely on demographic data to identify underrepresented groups. Without demographic data, alternative methods are needed to accurately detect and mitigate biased predictions for underrepresented students. First, while active learning methods [Shukla and Ahmed, 2021] can select hard-to-predict samples, they often prioritize accuracy over fairness, potentially neglecting underrepresented groups. Alternatively, human–AI collaborative approaches, such as ensemble strategies [Zhang *et al.*, 2021], often overlook human cognitive bias and assume equal reliability between AI and humans, leading to biased and inaccurate predictions.

The second challenge is *addressing different types of human cognitive biases across different demographic groups*. To mitigate confirmation bias, providing AI prediction criteria (e.g., estimated activities needed for a Distinction) can help humans re-evaluate their assumptions about the required activities to achieve a certain grade. However, these prediction criteria vary across demographic groups (e.g., older students may need more efforts than younger ones to achieve the same grade) and cannot be directly applied to all demographic groups. The affect heuristic further complicates predictions, as humans who favor consistent work over last-minute effort may incorrectly predict a Pass for the student shown in Figure 1, who actually achieve a Distinction by extra effort before the final. Addressing these cognitive biases requires a thorough consideration of diverse study behaviors across demographic groups.

To address these challenges, we propose CollabDebias, a framework that combines the complementary strengths of AI and HI to mitigate both AI demographic bias and human cognitive bias, enabling accurate and fair human–AI collaborative predictions, as illustrated in Figure 1. In particular, for AI demographic bias, it identifies underrepresented students by estimating uncertainty in training gradients, as gradient variability is a reliable indicator of prediction generalizability, especially for underrepresented groups with distinct characteristics. For human cognitive bias, CollabDebias designs novel visualization for AI decision areas and attention mechanism, providing comprehensive classification criteria that explain how AI model makes predictions across diverse demographic groups, effectively addressing cognitive bias. Additionally, a reliability-aware, estimation theory-based integration strategy combines debiased human and AI predictions, correcting bias without compromising accuracy. Experiments on Open University Learning Analytics Dataset (OULAD) [Kuzilek *et al.*, 2017] show that CollabDebias improves student performance prediction accuracy (e.g., +5.43% in F1) and fairness (e.g., -39.42% in Equalized Odds) without requiring sensitive demographic data.

2 Related Work

AI and HI in Online Education. There is a growing trend of leveraging AI and HI to enhance learning experiences and outcomes in online education [Ma *et al.*, 2022; Jiang *et al.*, 2024]. For instance, Abdi *et al.* [2020] incorporated crowdsourced contributions from students to enhance the efficiency of online education assessments and promote student engagement. Zhu *et al.* [2022] developed a metaverse-based online education platform that leverages human editor expertise to improve user experience of AI-driven educational modules. Pardamean *et al.* [2022] proposed an AI-based collaborative filtering algorithm to predict student learning preferences and recommend tailored learning materials. While both AI and HI have been utilized to facilitate

online education, they also suffer from certain biases, limiting their fairness of educational opportunities for all learners [Neal *et al.*, 2022].

AI Demographic Bias. There has been a surge in AI demographic bias analysis and mitigation across various fields [Caliskan, 2023]. For instance, Raji *et al.* [2019] audited commercial facial analysis algorithms and identified significant biases in the categorization across racial groups. Blodgett *et al.* [2020] conducted a comprehensive review of bias in language processing systems, demonstrating that models trained on biased datasets often reinforce stereotypes and proposed data curation to mitigate biases. In education, Baker *et al.* [2022] explored AI demographic bias, recommending fairness analysis during data collection and model evaluation. Similarly, Wongvorachan *et al.* [2024] investigated bias mitigation techniques, such as resampling, in student dropout rate prediction models. However, fair AI solutions often 1) rely on demographic group to re-weight data or impose fairness constraints, and 2) improve fairness at the expense of reduced accuracy [Wongvorachan *et al.*, 2024].

Cognitive Bias in Human–AI Collaboration. Human cognitive bias has been a critical issue in human–AI collaboration where human perceptions, expectations, and interpretations can introduce bias to undermine the potential synergies between human expertise and AI [Neal *et al.*, 2022]. Several solutions have been proposed to address human cognitive bias in human–AI collaboration [Kliegr *et al.*, 2021]. For example, Draws *et al.* [2021] introduced a checklist to mitigate cognitive bias in crowdsourcing data annotation. Gemalmaz *et al.* [2021] developed a bias-aware label aggregation method that examines cognitive bias to infer accurate labels to train AI models. Soleimani *et al.* [2022] investigated the collaboration and knowledge sharing between HR managers and AI developers in mitigating cognitive bias in AI-assisted recruitment systems. However, none of the solutions have leveraged the decision making information from AI model itself to reduce cognitive bias in human–AI collaboration.

3 Problem Formulation

In this section, we formally introduce our problem of human–AI collaborative student performance prediction in online education. We define the input **activity data** $\boldsymbol{A} = \{A_1, ..., A_N\}$ as a diverse range of activities (e.g., reviewing course materials, taking quizzes, engaging in discussion and collaboration) carried out by students in an online course. Here, A_i denotes the activity data for the i^{th} student, where N stands for the total number of students enrolled in the course. In our problem, we measure activities using the clickstream data generated from an online learning platform each day throughout a semester (e.g., numbers of clicks on different types of course activities per day), which aligns with established common practices in student performance prediction frameworks [Qiu *et al.*, 2022]. Clickstream data serve as a proxy for student engagement, reflecting behaviors such as content review, assessment participation, and collaboration [Kuzilek *et al.*, 2017]. We define the output **student final performance** $\boldsymbol{P} = \{P_1, ..., P_N\}$ as a set denoting the final performance grade (*Fail*, *Pass*, or *Distinction*) of all students enrolled in a course. P_i represents the final performance assigned by the course instructor to the i^{th} student.

We introduce $\boldsymbol{D} = \{D_1, ..., D_M\}$ to represent a **demographic attribute** of students (e.g., gender, age, highest education), where M denotes the number of categories for the demographic attribute. Across different demographic attribute categories, we define **underrepresented groups ($\boldsymbol{U}$)** to indicate traditionally underrepresented groups of students, such as female students in STEM courses. In our problem, demographic information is used solely for evaluation purposes and is not available during model training to protect private information from students [Nguyen *et al.*, 2023]. We note that achieving demographic fairness in the absence of demographic information is particularly challenging because it is impractical to tune the AI model or augment the input data for particular demographic groups without knowing the their demographic labels [Zhang, 2024]. Therefore, we take a different approach by leveraging complementary HI to address the demographic bias of AI. In particular, we introduce the **collaborative prediction ($\widehat{\boldsymbol{P}}$)** that represents the overall human–AI collaborative prediction of our student performance prediction framework. We define $\widehat{\boldsymbol{P}} = \{\widehat{P_1}, ..., \widehat{P_N}\}$ as a set comprising predictions of our framework, with each $\widehat{P}_i$ representing the prediction of CollabDebias for the i^{th} student.

The overall objective of CollabDebias is to explore AI and HI to achieve accurate and fair predictions of student performance. This entails maximizing prediction accuracy while minimizing demographic bias as follows:

$$\arg\max_{\widehat{P_i}} \Pr(\widehat{P}_i = P_i \mid \boldsymbol{A}) \quad \& \quad \arg\min_{\widehat{P_i}} \mathrm{F}(\widehat{P}_i, P_i \mid \boldsymbol{A}) \tag{1}$$

for $\forall 1 \leq i \leq N$, where $\mathrm{F}(\cdot, \cdot)$ represents the fairness metric to assess performance disparities across various demographic groups given demographic attribute $\boldsymbol{D}$. We note that the fairness metric is minimized since a smaller fairness metric value indicates better fairness performance.

4 Solution

CollabDebias is a bidirectional framework leveraging the complementary strengths of AI and HI to mitigate AI demographic bias and human cognitive bias in student performance prediction. It comprises three key modules: 1) *Gradient Uncertainty-based AI Demographic Bias Identification*, which uses gradient-based deep uncertainty learning to detect biased samples for crowdsourced human predictions to mitigate the bias; 2) *Decision Uncertainty-based Human Cognitive Bias Mitigation*, which designs an AI decision area and attention visualization scheme to address human cognitive bias; and 3) *Estimation Uncertainty-aware Integration*, which integrates AI and human predictions using a reliability-aware, estimation theory-based model to ensure accurate and fair collaborative predictions.

4.1 Gradient Uncertainty-based AI Demographic Bias Identification

To predict students' final performance $\boldsymbol{P}$ from activity data $\boldsymbol{A}$, we build a **performance prediction model** ($m(\cdot)$):

$$\widehat{P_j^{AI}} = m(A_j), \quad \forall 1 \leq j \leq J \tag{2}$$

where $\widehat{P_j^{AI}}$ is the AI prediction for the j^{th} student's final performance in the *training set*. We utilize the long short-term memory (LSTM) model followed by fully connected layers as our model $m(\cdot)$, where the LSTM model is well-suited for extracting meaningful patterns from sequential behavioral data, as demonstrated in prior educational research [Li *et al.*, 2020]. To further enhance the model's ability to focus on the most relevant sections of the input behavioral data, we incorporate an attention mechanism [Vaswani *et al.*, 2017] after the LSTM model, which learns attention weights to prioritize key temporal features for each student's performance prediction. While our input data consists of structured, sequential time-series activity records (e.g., daily counts of course activity accesses), its explainable nature differs significantly from unstructured data like images or texts, which often require highly complex architectures such as large-scale transformers. The structure simplicity and explainability of our data allows us to employ a streamlined model architecture with fewer parameters, ensuring computational efficiency and reducing the risk of overfitting. Importantly, this design choice enhances model explainability, enabling clear visualization of decision areas and attention mechanisms, which is critical for the development of our second key module in Section 4.2.

To identify biased AI predictions and improve fairness, we leverage deep uncertainty learning to select a subset of students from the testing set where the model exhibits high prediction uncertainty, referred to as the **biased AI subset ($\boldsymbol{S}$)**. Formally, $\boldsymbol{S} = \{A_1, ..., A_K\}$, where $K = \alpha \cdot N$. This subset aims to prioritize underrepresented groups $\boldsymbol{U}$ without requiring demographic information, as these students are more likely to receive uncertain predictions due to insufficient training data. Please note that we do not use any labels of the samples in $\boldsymbol{S}$ by following the common practices in using the testing data [Ren *et al.*, 2021].

AI models tend to produce uncertain predictions for samples with high gradient uncertainty during training, as gradient variability is a known indicator of generalization difficulty, particularly when learning from diverse or underrepresented distributions [Ren *et al.*, 2018]. Samples from underrepresented groups often exhibit distinct input characteristics (e.g., older students often require more efforts to achieve the same grade as younger ones), making them inherently more challenging for the model to learn effectively. We propose a gradient uncertainty-based method to identify such samples *without accessing sensitive demographic attributes*. To quantify gradient uncertainty, we compute the variance of gradients for each student across all training epochs. For a given student, we calculate the gradient at each epoch and then estimate the variance of these gradients over time. This variance serves as a proxy for uncertainty, as higher variance indicates greater instability in the model's learning process for that sample. Formally, gradient uncertainty is computed as:

$$V_j = \mathrm{Var}\left[\frac{\partial \mathcal{L}(\widehat{P_j^{AI}}, P_j)}{\partial A_j}\right] = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial \mathcal{L}_t(\widehat{P_{j,t}^{AI}}, P_j)}{\partial A_j} - g_j\right)^2 \quad (3)$$

for $\forall 1 \leq j \leq J$, where T is the total number of training epochs, $\mathcal{L}_t$ is the loss at the t^{th} epoch, and g_j is the average of gradient over all epochs for the j^{th} sample, computed as:

$$g_j = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathcal{L}_t(\widehat{P_{j,t}^{AI}}, P_j)}{\partial A_j} \quad (4)$$

We select a subset of training samples $\boldsymbol{S^T}$, comprising the top α largest gradient variances V_j in the training set:

$$\boldsymbol{S^T} = \{A_j \mid V_j \geq V_\alpha, \forall 1 \leq j \leq J\} \quad (5)$$

where V_α is the α largest gradient variance. The choice of α is made empirically, considering the trade-off between algorithmic fairness and the available crowdsourcing budget.

Identifying a subset of *testing* samples with significant gradient uncertainty is non-trivial due to the lack of ground truth annotations required for gradient computation. To address this limitation, we select the top α testing samples exhibiting similar characteristics as the samples selected from the training set to be incorporated into the biased AI subset $\boldsymbol{S}$ for human intervention. This strategy is motivated by the observation that an AI model generates similar predictions and gradients for input samples that share similar characteristics [Charpiat *et al.*, 2019]. To quantify similarity, we compute the Euclidean distance between vectors representing each test sample and the selected training samples in $\boldsymbol{S^T}$:

$$E_i = \sum_{A_j \in \boldsymbol{S^T}} \|A_j - A_i\|_2, \forall 1 \leq i \leq N \quad (6)$$

Then, we select the top α testing samples exhibiting similar characteristics as selected training samples:

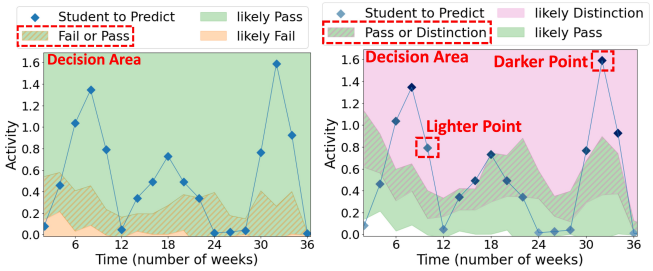
$$\boldsymbol{S} = \{A_i \mid E_i \leq E_\alpha, \forall 1 \leq i \leq N\} \quad (7)$$

where a smaller value of Euclidean distance indicates a higher similarity between two samples. E_α is the α smallest distance value. Given the selected biased AI subset $\boldsymbol{S}$, we then leverage HI to troubleshoot these demographically biased samples. Our bias identification design effectively selects demographically biased samples, increasing the percent of underrepresented age group from 24.4% in the whole population to 56% in the biased AI subset.

4.2 Decision Uncertainty-based Human Cognitive Bias Mitigation

Our visualization designs to address human cognitive bias is shown in Figure 2. We first present the activity data of a student using a line chart with blue data points. The activity data A_i for the i^{th} student is the completed activities on the online learning platform every two weeks throughout the semester measured by clickstream data. We consider bi-weekly activity data since aggregating activities within a certain period is observed to be both visually clear and informative enough to assist humans in accurate prediction. We rescale all numbers of activities for visualization clarity.

To address **confirmation bias**, where humans favor predictions aligning with their preexisting beliefs, we leverage AI to generate classification criteria that help recalibrate human hypotheses. To ensure these criteria are free from AI demographic bias, we introduce a novel **decision area** design based on the AI model's decision boundary. While decision boundaries separate prediction categories [Karimi *et al.*,



(a) Visualization of **decision area** in *step 1* of our prediction task (i.e., selecting Fail or Pass).

(b) Visualization of **attention and decision area** in *step 2* of our task, where attention is presented by varying color darkness.

Figure 2: Visualization of our decision area and attention designs for a sample student who attains a Distinction.

2019], they often fail to account for intrinsic uncertainty in AI predictions, which can vary across demographic groups (Section 1). To address this limitation, we extend the concept of decision boundaries to decision areas, which represent an uncertain range of decision criteria rather than a fixed boundary. Specifically, we construct decision areas by visualizing a region within one standard deviation (σ) of the decision boundary. This design captures the variability in prediction criteria across demographic groups, providing comprehensive and unbiased guidance for human annotators. By visualizing decision areas, we ensure that humans are exposed to a broader and more representative range of criteria, mitigating confirmation bias and improving collaborative prediction accuracy.

Decision boundaries in the original input feature space are often challenging to visualize, as they often exist in high-dimensional spaces and are typically intertwined. Thus, we propose to use **borderline samples**—training samples with final grades near the decision boundary between two adjacent grade levels (e.g., Pass and Distinction)—to generate and visualize decision areas in the sample space. In particular, we identify training samples whose predictions keep changing between two adjacent grade levels across epochs during the training process [Chang *et al.*, 2017]. These uncertain predictions between two adjacent grade levels indicate a high probability of the samples belonging to either of these two grade levels, typically because these samples share characteristics with two grade levels. We focus on such samples whose predictions exhibit high uncertainty during training by employing Shannon entropy as a quantitative measure as follows:

$$H(A_j) = -\sum_{t=1}^{T} \widehat{P^{AI}_{j,t}} \cdot \log\left(\widehat{P^{AI}_{j,t}}\right), \forall 1 \leq j \leq J \tag{8}$$

where T is the total number of training epochs. We focus on samples A_j where $H(A_j)$ exceeds a predefined threshold β, capturing those borderline samples with high uncertainty:

$$\boldsymbol{S^B} = \{A_j \mid H(A_j) > \beta, \forall 1 \leq j \leq J\} \tag{9}$$

Our decision area design is shown in Figure 2, where we separate the prediction task into two steps for visual clarity. For each student, we plot the decision areas that represent the uncertain areas between grade levels (i.e., the "Fail or Pass" and "Pass or Distinction" areas with hatched lines). The decision area visualization mitigates human confirmation bias (e.g., the preconception that a Distinction grade is rarely assigned) by presenting AI classification criteria, such as the level of activity required for a student to achieve a Distinction.

The **affect heuristic**—people tend to select the answer based on their initial emotional reactions—primarily appears in students whose final performance is difficult to predict. For instance, a student with average prior effort but significantly increased effort before the final exam may still achieve a Distinction in our problem. A person favoring late-stage effort may predict this correctly, while another who prefers consistent work may incorrectly predict a Pass. Such incorrect predictions arise from differing importance assigned to different periods during the semester. Thus, to identify samples where the affect heuristic is prevalent, we measure inter-annotator inconsistency using Shannon entropy for each student:

$$G(A_i) = -\sum_{a=1}^{A} \widehat{P_i^{Crowd_a}} \cdot \log\left(\widehat{P_i^{Crowd_a}}\right), \forall 1 \leq i \leq N \tag{10}$$

where A is the number of annotators predicting for a student.

To tackle such challenge, we leverage the **attention mechanism**, widely recognized for its ability to enhance performance and explainability [Vaswani *et al.*, 2017]. We present our visualization design in Figure 2b, where we plot different attention weights for each activity data point using varying darkness of the blue color. A darker color indicates a larger attention weight computed by the model. Our color darkness visualization design is motivated by the characteristic of human visual attention that humans tend to pay more attention to darker color compared to lighter color [Sun *et al.*, 2016]. For the student in Figure 2 who achieves a Distinction, the first step is relatively simple given the relatively high activities in Figure 2a. However, the second step shown in Figure 2b is challenging only with decision areas. There are some weeks where the activity clearly lies in the Distinction area and other weeks where the activity obviously lies in the Pass area. Our attention visualization of varying color darkness assists humans in predicting the correct answer of Distinction by highlighting the extra work before the final.

4.3 Estimation Uncertainty-aware Integration

We collect human predictions from the crowdsourcing platform using our visualization design to address human cognitive bias as explained in Subsection 4.2. Specifically, human annotators are tasked with predicting only the samples in the biased AI subset $\boldsymbol{S}$, which are more likely to belong to underrepresented groups and thus receive uncertain predictions from AI models. In addition to bias, we observe another dimension of uncertainty: the AI model and different human annotators may exhibit different levels of reliability when it comes to prediction accuracy. In these scenarios with varying source reliability, directly employing the majority voting to aggregate AI and human predictions is known to be suboptimal. In particular, majority voting assumes the same reliability among all different sources including the AI model and all human annotators and thus assigns the same weight to all of them in prediction aggregation. In such a case, a biased and less reliable source is treated as the same as an unbiased

and reliable source in voting, leading to inaccurate and unfair collaborative predictions.

Therefore, we design a maximum likelihood estimation model, which enables us to jointly derive accurate human–AI collaborative predictions $\widehat{\boldsymbol{P}}$ while quantifying the reliability of the AI model and each human annotator. The model iterates between two steps: 1) estimating the reliability of each source (AI and annotators) given current predictions, and 2) updating collaborative predictions based on estimated reliability. This process maximizes the following likelihood function, which minimizes collaborative prediction errors:

$$L(P,\hat{P}) = \prod_{i=1}^{I}\prod_{c=1}^{C}\Big(\Pr\left(\widehat{P_i^{AI}}=c\middle|P_i=c\right)^{\mathbb{1}(P_i=c)} \cdot \prod_{a=1}^{A}\Pr\left(\widehat{P_i^{Crowd_a}}=c\middle|P_i=c\right)^{\mathbb{1}(P_i=c)}\Big) \tag{11}$$

where C represents the number of categories of student final performance and A is the number of crowd workers making predictions for each student. $\Pr\left(\widehat{P_i^s}=c|P_i=c\right)^{\mathbb{1}(P_i=c)}$ represents the probability that a source s predicts $\widehat{P_i^s}=c$ given the true category $P_i=c$. $\mathbb{1}(P_i=c)$ is a indicator function that equals 1 if $P_i=c$ is true, and 0 otherwise.

5 Experiments

We leverage the Open University Learning Analytics Dataset [Kuzilek *et al.*, 2017], focusing on a STEM course with 1,938 students. The dataset comprises demographic information (age), activity data (daily clickstream interactions), and performance outcomes (grades: Fail, Pass, Distinction). Following established fairness practices [Hardt *et al.*, 2016], the students were classified into two age groups: underrepresented ($\geq$35, 24.4%) and non-underrepresented ($<$35, 75.6%). Detailed description and analysis of the dataset are provided in Appendix. We collect human predictions through Amazon Mechanical Turk (*MTurk*), a leading crowdsourcing platform that grants access to a vast global workforce at reasonable costs. We set the percentage α of the selected biased AI samples for crowdsourcing as 15% and recruited 5 people to work on each prediction task. Our crowdsourcing interface design and settings are in Appendix.

We compare our CollabDebias with state-of-the-art AI, fair AI, and human–AI baselines: **ANN** [Waheed *et al.*, 2020], **BCEP** [Qiu *et al.*, 2022], **SPDN** [Li *et al.*, 2020], **VS** [Kini *et al.*, 2021], **JMLR19** [Zafar *et al.*, 2019], **NeurIPS21** [Bendekgey and Sudderth, 2021], **StreamCollab** [Zhang *et al.*, 2021], **DeepActive** [Sener and Savarese, 2018], **LearningLoss** [Shukla and Ahmed, 2021], **DebiasEdu** [Zong *et al.*, 2023]. Detailed baseline descriptions are in Appendix. To evaluate accuracy, we use four multi-class classification metrics—F1, Accuracy (Acc), Cohen's Kappa Score (Kappa), and Matthews Correlation Coefficient (MCC) [Chicco and Jurman, 2020]. For fairness, we consider four metrics: True Positive Parity (TP Par.), False Positive Parity (FP Par.), Equalized Odds (Eq. Odds), and Accuracy Parity (Acc Par.) [Hardt *et al.*, 2016; Yan *et al.*, 2020]. Fairness metrics quantify *unfairness* as differences across demographic groups, where lower values indicate better fairness. Since there are no widely accepted human cognitive bias metrics, we propose two metrics tailored to our problem: 1) Confirmation Bias Parity (Conf. Par.): Sum of absolute differences in prediction accuracy across three grade levels. 2) Affect Heuristic Parity (Affect Par.): Sum of absolute differences in prediction accuracy across three observed study patterns: consistent effort, last-minute effort, and early disengagement. Similar to fairness, lower values indicate less bias. Detailed experiment settings are in Appendix.

Category	Algorithm	F1	Acc	Kappa	MCC
AI	ANN	0.6123	0.5971	0.3136	0.3183
	BCEP	0.6931	0.7208	0.4523	0.4572
	SPDN	0.7126	0.7059	0.4780	0.4793
Fair AI	VS	0.6136	0.6000	0.3834	0.4189
	JMLR19	0.6778	0.6647	0.4319	0.4397
	NeurIPS21	0.7509	0.7500	0.5442	0.5443
Human–AI	StreamCollab	0.7003	0.6735	0.4451	0.4514
	DeepActive	0.7214	0.7177	0.4927	0.4932
	LearningLoss	0.7206	0.7176	0.4920	0.4924
	DebiasEdu	0.7861	0.7882	0.6046	0.6065
Ours	**CollabDebias**	**0.8288**	**0.8294**	**0.6862**	**0.6864**

Table 1: Experiment results on prediction *accuracy*.

Collaborative Prediction Accuracy Comparison. We first compare the student performance prediction accuracy of all approaches on the online STEM course dataset. The results presented in Table 1 demonstrate that our CollabDebias achieves consistent performance gains compared to all baselines on all metrics. For instance, the performance gains of CollabDebias compared to the best-performing baseline DebiasEdu on F1, Acc, Kappa, and MCC are 5.43%, 5.23%, 13.50%, and 13.17%, respectively. Such performance improvements verify that our CollabDebias successfully leverages the complementary strengths of humans and AI to address demographic bias and cognitive bias, thus improving overall student performance prediction accuracy. Besides, we observe that human–AI baselines do not necessarily always outperform AI-only baselines. Equal/underperformance may perhaps be attributed to 1) the AI model designs are different in the AI and human–AI baselines and 2) the sample subset selection methods in human–AI baselines may fail to select samples that receive biased and inaccurate AI predictions in our student performance prediction problem (i.e., if the selected AI predictions are reasonably accurate, it is difficult for crowd annotators to further improve prediction performance). Moreover, a key distinction between our work and DebiasEdu is that DebiasEdu relies on demographic data, while CollabDebias operates without such data due to privacy constraints. In our setting, DebiasEdu's performance deteriorates significantly (see Tables 1 and 2), highlighting its reliance on demographic data. In contrast, CollabDebias remains effective by leveraging uncertainty-driven bias identification and human-guided corrections.

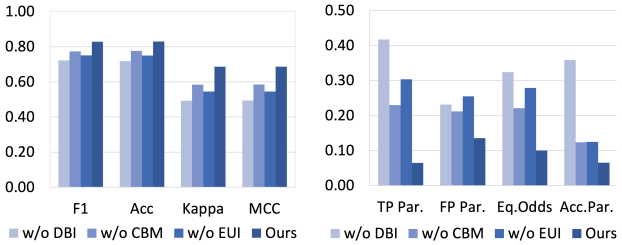
Collaborative Prediction Demographic Fairness Comparison. We evaluate the fairness of our CollabDebias and all compared baselines on our dataset. Results in Table 2 show that CollabDebias consistently outperforms all baselines by

Category	Algorithm	TP Par.	FP Par.	Eq. Odds	Acc Par.
AI	ANN	0.2221	0.2079	0.2189	0.2117
	BCEP	0.2851	0.3346	0.3073	0.1463
	SPDN	0.4671	0.2760	0.3780	0.4327
Fair AI	VS	0.3134	0.3843	0.3522	0.3075
	JMLR19	0.4868	0.3188	0.4044	0.4625
	NeurIPS21	0.3032	0.2546	0.2823	0.1247
Human–AI	StreamCollab	0.3286	0.3001	0.3151	0.3760
	DeepActive	0.4168	0.2313	0.3306	0.3588
	LearningLoss	0.4169	0.2425	0.3363	0.3589
	DebiasEdu	0.2847	0.2115	0.2457	0.3315
Ours	**CollabDebias**	**0.1309**	**0.1352**	**0.1326**	**0.0523**

Table 2: Experiment results on *demographic fairness*.

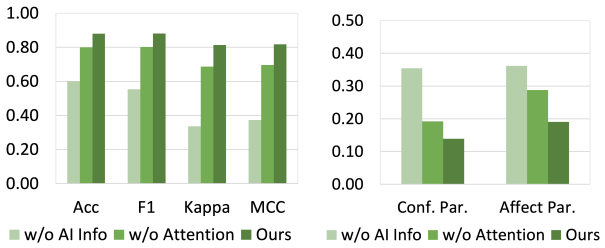
achieving the lowest values in all fairness metrics (i.e., lowest prediction differences between groups). For example, the decreases—that is, improvement—in TP Par., FP Par., Eq. Odds, and Acc Par. of our CollabDebias compared to the best-performing baseline ANN are 41.06%, 34.97%, 39.42%, and 75.30%, respectively. We note that ANN achieves the lowest bias compared to other baselines simply because it predicts similarly inaccurately on all demographic groups, which is not useful in terms of the objective of fairness design (i.e., achieving accurate predictions among all demographic groups). The substantial performance gains in fairness can be attributed to the fact that our CollabDebias approach successfully addresses demographic bias in student performance prediction by leveraging the common-sense reasoning and contextual understanding of HI while mitigating human cognitive bias in this process. In addition, we observe that fair AI baselines fail to achieve better fairness performance across all metrics, which can be related to the fact that 1) these baselines can be sensitive to the imperfect crowdsourced labels in the selected biased AI subset, 2) they often only focus on improving one specific fairness metric, and 3) human–AI baselines in our experiments also include designs to improve inaccurately predicted samples, where the accuracy improvement may cause bias reduction to some extent if the improvement happens to be more relevant to underrepresented samples.

Ablation Study. We first evaluate the contributions of the three core components of CollabDebias: Gradient Uncertainty-based AI Demographic Bias Identification (DBI), Decision Uncertainty-based Human Cognitive Bias Mitigation (CBM), and Estimation Uncertainty-aware Integration (EUI). To exclude the DBI module, we uniformly sample 15% of samples from the test set for human prediction and integration, ensuring the sampling rate matches that of our framework for a fair comparison. To disable the CBM module, we modify the crowdsourcing interface by removing the decision area and attention visualizations, instead providing a display of average student activity to assist human predictions. To remove the EUI module, we use the crowd predictions from the selected subset exclusively to retrain the AI model. The accuracy and fairness evaluation results are presented in Figure 3. These results demonstrate the significant contributions of all three modules—DBI, CBM, and EUI—in improving both the accuracy and demographic fairness of the CollabDebias framework.



(a) Accuracy. (b) Demographic Fairness.

Figure 3: Ablation study of collaborative prediction.



(a) Accuracy. (b) Cognitive Bias.

Figure 4: Ablation study of human prediction on the biased subset.

To analyze the effectiveness of CollabDebias in improving human prediction accuracy and reducing cognitive bias, we conduct an ablation study on the selected biased AI subset $\boldsymbol{S}$ by removing our two key design modules of decision area and attention mechanism hierarchically. In particular, we first remove the attention design by presenting the visualization with only decision areas. We then further remove the decision areas by showing the visualization with the average activity of all students. The results, presented in Figure 4, demonstrate that each of our two modules contributes significantly to improving annotators' prediction accuracy and reducing cognitive bias. Furthermore, 82.6% of annotators reported increased confidence in predicting student performance when provided with AI decision area and attention information, validating the effectiveness of our design.

6 Conclusion

In this paper, we develop CollabDebias to address AI demographic bias and human cognitive bias in human–AI collaborative student performance prediction. We design a bidirectional framework that incorporates uncertainty learning-based bias identification and collaborative aggregation methods to mitigate AI demographic bias, while generating uncertainty-aware AI decision area and attention visualization to reduce human cognitive bias. Experiment results on an online course demonstrate that CollabDebias achieves consistent performance gains compared to all state-of-the-art baselines in both prediction accuracy and fairness. We believe our CollabDebias provides useful insights to address AI demographic bias and human cognitive bias in other privacy-aware human–AI collaborative applications (e.g., financial fraud detection, medical image classification).

Acknowledgments

This research is supported in part by the National Science Foundation under Grant No. CNS-2427070, IIS-2202481. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[Abdi *et al.*, 2020] Solmaz Abdi, Hassan Khosravi, and Shazia Sadiq. Modelling learners in crowdsourcing educational systems. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, pages 3–9. Springer, 2020.

[Ayeni *et al.*, 2024] Oyebola Olusola Ayeni, Nancy Mohd Al Hamad, Onyebuchi Nneamaka Chisom, Blessing Osawaru, and Ololade Elizabeth Adewusi. Ai in education: A review of personalized learning and educational technology. *GSC Advanced Research and Reviews*, 18(2):261–271, 2024.

[Baker and Hawn, 2022] Ryan S Baker and Aaron Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pages 1–41, 2022.

[Bendekgey and Sudderth, 2021] Henry C Bendekgey and Erik Sudderth. Scalable and stable surrogates for flexible classifiers with fairness constraints. *Advances in Neural Information Processing Systems*, 34:30023–30036, 2021.

[Blodgett *et al.*, 2020] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

[Budd *et al.*, 2021] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, 2021.

[Caliskan, 2023] Aylin Caliskan. Artificial intelligence, bias, and ethics. In *IJCAI*, pages 7007–7013, 2023.

[Chang *et al.*, 2017] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30, 2017.

[Charpiat *et al.*, 2019] Guillaume Charpiat, Nicolas Girard, Loris Felardos, and Yuliya Tarabalka. Input similarity from the neural network perspective. *Advances in Neural Information Processing Systems*, 32, 2019.

[Chicco and Jurman, 2020] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

[Draws *et al.*, 2021] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 48–59, 2021.

[Draws *et al.*, 2022] Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. The effects of crowd worker biases in fact-checking tasks. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2114–2124, 2022.

[Gemalmaz and Yin, 2021] Meric Altug Gemalmaz and Ming Yin. Accounting for confirmation bias in crowdsourced label aggregation. In *IJCAI*, pages 1729–1735, 2021.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[Jiang *et al.*, 2024] Lu Jiang, Yanan Xiao, Xinxin Zhao, Yuanbo Xu, Shuli Hu, Pengyang Wang, and Minghao Yin. Hierarchical reinforcement learning on multi-channel hypergraph neural network for course recommendation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, 2024.

[Karimi *et al.*, 2019] Hamid Karimi, Tyler Derr, and Jiliang Tang. Characterizing the decision boundary of deep neural networks. *arXiv preprint arXiv:1912.11460*, 2019.

[Kini *et al.*, 2021] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34:18970–18983, 2021.

[Kizilcec and Lee, 2022] René F Kizilcec and Hansol Lee. Algorithmic fairness in education. In *The ethics of artificial intelligence in education*, pages 174–202. Routledge, 2022.

[Kliegr *et al.*, 2021] Tomáš Kliegr, Štěpán Bahník, and Johannes Fürnkranz. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 295:103458, 2021.

[Kuzilek *et al.*, 2017] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset. *Scientific data*, 4(1):1–8, 2017.

[Li *et al.*, 2020] Xiang Li, Xinning Zhu, Xiaoying Zhu, Yang Ji, and Xiaosheng Tang. Student academic performance prediction using deep multi-source behavior sequential network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 567–579. Springer, 2020.

[Ma *et al.*, 2022] Haiping Ma, Jingyuan Wang, Hengshu Zhu, Xin Xia, Haifeng Zhang, Xingyi Zhang, and Lei Zhang. Reconciling cognitive modeling with knowledge forgetting: A continuous time-aware neural network approach. In *IJCAI*, pages 2174–2181, 2022.

[Neal *et al.*, 2022] Tess Neal, Pascal Lienert, Emily Denne, and Jay P Singh. A general model of cognitive bias in human judgment and systematic review specific to forensic mental health. *Law and human behavior*, 46(2):99, 2022.

[Nguyen *et al.*, 2023] Andy Nguyen, Ha Ngan Ngo, Yvonne Hong, Belle Dang, and Bich-Phuong Thi Nguyen. Ethical principles for artificial intelligence in education. *Education and Information Technologies*, 28(4):4221–4241, 2023.

[OpenAI, 2023] OpenAI. Chatgpt, 2023.

[Pardamean *et al.*, 2022] Bens Pardamean, Teddy Suparyanto, Tjeng Wawan Cenggoro, Digdo Sudigyo, and Andri Anugrahana. Ai-based learning style prediction in online learning for primary education. *IEEE Access*, 10:35725–35735, 2022.

[Qiu *et al.*, 2022] Feiyue Qiu, Guodao Zhang, Xin Sheng, Lei Jiang, Lijia Zhu, Qifeng Xiang, Bo Jiang, and Pingkuo Chen. Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*, 12(1):453, 2022.

[Raji and Buolamwini, 2019] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

[Ren *et al.*, 2018] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR, 2018.

[Ren *et al.*, 2021] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

[Sener and Savarese, 2018] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[Shukla and Ahmed, 2021] Megh Shukla and Shuaib Ahmed. A mathematical analysis of learning loss for active learning in regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3328, 2021.

[Soleimani *et al.*, 2022] Melika Soleimani, Ali Intezari, and David J Pauleen. Mitigating cognitive biases in developing ai-assisted recruitment systems: A knowledge-sharing approach. *International Journal of Knowledge Management (IJKM)*, 18(1):1–18, 2022.

[Sun *et al.*, 2016] Peng Sun, Charles Chubb, Charles E Wright, and George Sperling. Human attention filters for single colors. *Proceedings of the National Academy of Sciences*, 113(43):E6712–E6720, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Waheed *et al.*, 2020] Hajra Waheed, Saeed-Ul Hassan, Naif Radi Aljohani, Julie Hardman, Salem Alelyani, and Raheel Nawaz. Predicting academic performance of students from vle big data using deep learning models. *Computers in Human behavior*, 104:106189, 2020.

[Wang *et al.*, 2022] Qiaosi Wang, Ida Camacho, Shan Jing, and Ashok K Goel. Understanding the design space of ai-mediated social interaction in online learning: Challenges and opportunities. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–26, 2022.

[Wongvorachan *et al.*, 2024] Tarid Wongvorachan, Okan Bulut, Joyce Xinle Liu, and Elisabetta Mazzullo. A comparison of bias mitigation techniques for educational classification tasks using supervised machine learning. *Information*, 15(6):326, 2024.

[Yan *et al.*, 2020] Shen Yan, Hsien-te Kao, and Emilio Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724, 2020.

[Zafar *et al.*, 2019] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778, 2019.

[Zhang *et al.*, 2021] Yang Zhang, Lanyu Shang, Ruohan Zong, Zeng Wang, Ziyi Kou, and Dong Wang. Streamcollab: A streaming crowd-ai collaborative system to smart urban infrastructure monitoring in social sensing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 179–190, 2021.

[Zhang *et al.*, 2023] Yang Zhang, Lanyu Shang, Ruohan Zong, Huimin Zeng, Zhenrui Yue, and Dong Wang. Collabequality: A crowd-ai collaborative learning framework to address class-wise inequality in web-based disaster response. In *Proceedings of the ACM Web Conference 2023*, pages 4050–4059, 2023.

[Zhang, 2024] Wenbin Zhang. Ai fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine*, 45(3):386–395, 2024.

[Zhu, 2022] Hongyin Zhu. Metaaid: A flexible framework for developing metaverse applications via ai technology and human editing. *arXiv preprint arXiv:2204.01614*, 2022.

[Zong *et al.*, 2023] Ruohan Zong, Yang Zhang, Frank Stinar, Lanyu Shang, Huimin Zeng, Nigel Bosch, and Dong Wang. A crowd–ai collaborative approach to address demographic bias for student performance prediction in online education. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 198–210, 2023.