What Constitutes AI Harms and/or Unfairness? An Empirical Analysis of Teacher Deliberation with a Fairness Elicitation Scaffold

Shamya Karumbaiah[1 [0000−0002−7920−4510]], Yaxuan Yin[1 [0009−0006−6661−2183]], and Harry Brighouse[1 [0009−0007−3378−5704]]

University of Wisconsin–Madison, Madison, WI 53706, USA
{shamya.karumbaiah, yaxuan.yin, mhbrigho}@wisc.edu

Abstract. Which values matter, when, and why are highly contextual. Yet, conversations on values such as AI fairness often exclude stakeholder expertise. Consequent conceptualization and operationalization suffer from value misalignment. To bridge this gap between AI development and downstream use in education, we explore a method (reflective equilibrium) and a practical tool (fairness elicitation scaffold) that supports stakeholders such as educators to construct fairness conceptions ground up (e.g., what is (un)fair, for whom, why?). In this paper, we present empirical evidence by analyzing data collected from 29 teachers in two distinct educational contexts: United States and the Philippines. Our analysis revealed that there is a statistically significant increase in teachers' recognition of AI's potential to reinforce bias or unfair treatment with the scaffold. Teachers in this study understood fairness in educational AI as conditional and context-sensitive rather than rule-based, drawing on their knowledge of students, classroom dynamics, and instructional goals to weigh learning benefits against potential harms. Unequal learning outcomes were sometimes viewed as fair—particularly when they supported equity goals—but became contested when paired with additional harms or competitive stakes. These findings suggest that fairness in educational AI must account for how benefits and harms are distributed in specific contexts. Overall, our study underscores the need for bottom-up, participatory approaches that treat fairness as an ongoing reasoning process led by educational stakeholders. Future studies will investigate how stakeholders prioritize competing fairness values across contexts and how these priorities evolve through reflective deliberation.

Keywords: Fairness · Harms · Deliberation · Teachers · Context.

1 Introduction

As artificial intelligence (AI) and other technological interventions in education become more sophisticated, evidence suggests that some tools—particularly those designed around human-AI partnerships—can improve learning outcomes [8]. This has sparked interest in large-scale AI implementation in education,

which in turn raises critical questions about fairness and equity: are the harms and benefits of using AI for teaching and learning distributed equitably? These questions are deeply contextual, with answers varying by student populations, learning goals, sociopolitical settings, and the specific values that teachers and other stakeholders bring to AI adoption [24].

Yet, dominant frameworks that shape the development of responsible and fair AI systems rarely consider the unique ethical concerns in education [7], be it those from the tech industry [15], federal agencies (AI Risk Management Framework by the National Institute of Standards and Technology (NIST); [22]) or multilateral organizations (Organisation for Economic Co-operation and Development (OECD); [16]). These frameworks operate at a level of generality that makes direct application to education difficult: education involves distinctive relationships of power and trust (between teachers, students, parents, and institutions), developmental considerations, and equity stakes that generic principles do not adequately capture [20]. Even ethical, responsible, and human-centered principles can fail when applied to problematic use cases to begin with—an issue often referred to as "putting lipstick on a pig" [12].

Gross generalizations of deeply contextual fairness value and a lack of rigorous conceptualization [2] in these frameworks lead to invalid operationalizations [10]. For instance, questions such as "What system behaviors are fair? For whom? Why?" are often overlooked in favor of oversimplified technical notions of fairness (e.g., assuming homogeneity in student groups based on flawed, static demographic labels; [9]). Consequently, even well-intentioned AI bias audits and mitigation efforts may produce outputs that are misaligned with the social justice goals of real-world educational stakeholders. Moreover, different responsible AI values often conflict in non-trivial ways [23]. For instance, ensuring fairness may raise privacy concerns in sharing individual student demographics such as race, gender, special education, and English learner status. Likewise, fairer AI models may sacrifice performance, or better-performing AI models may be less transparent [11].

In sum, dominant frameworks on responsible AI offer a high-level, conceptual scaffolding for considering values; however, their application is limited in situated contexts like education and we lack the tools and methods for operationalizing these values in practice [6].

To bridge this gap, in this paper, we present empirical evidence for the efficacy of a *fairness elicitation scaffold* (elaborated in Section 2.1) in supporting teacher deliberation (Section 2.2) on different fairness notions. There are better, and worse, ways of conceptualizing fairness. The aim of the scaffold is not to settle which are the better ways. Instead, it aims to make it easier for the deliberator to arrive at their own judgments, and to avoid misunderstanding of others who make different judgments. We analyze how teachers use the scaffold to make sense of different conceptions of fairness, in the context of thinking about AI for teaching and learning. We discuss how the scaffold could also be used by researchers to elicit fairness notions from educational stakeholders such as teachers when designing fairness measures and auditing frameworks. Our analysis is

guided by the following research questions: *How do teachers deliberate on AI harms and/or unfairness using the fairness elicitation scaffold?*

2 AI Fairness Deliberation: A Method and Scaffold

2.1 The AI Fairness Elicitation Scaffold to Disambiguate Conceptions of AI Fairness

We developed an *AI fairness elicitation scaffold* to support educational stakeholders in deliberating and articulating how they understand fairness in the context of AI-supported learning. Rather than prescribing a definition of fairness, the scaffold presents a set of structured scenarios that vary how learning benefits, harms, risks, and time horizons are distributed across groups. Participants are prompted to reflect on which configurations they view as problematic, for whom, and why.

The scaffold assumes a working definition of learning that is specified within each scenario (e.g., performance on a knowledge assessment) and asks participants to reason about how anticipated benefits and harms are distributed across groups. In our study, participants considered an AI intervention for automated feedback on student essays that was expected to improve learning outcomes. They were invited to reflect on whether their judgments depended on which groups were compared (e.g., students from different regions or linguistic backgrounds), the magnitude of differences in benefit or harm, or whether certain thresholds or conditions shaped their evaluation.

To support this deliberation, the scaffold consists of six core cases that capture common benefit–fairness tensions in educational contexts:

1. *[Benefit, No Harm]* Using an AI intervention only makes one group's learning better and not the other, without reducing the learning of the other.
2. *[Benefit, Harm]* Using an AI intervention makes one group's learning better and makes the other group's learning worse compared to a feasible alternative intervention.
3. *[Unequal Benefit]* Using an AI intervention makes everyone's learning better off than before but improves learning more for one group than the other and
   (a) *[absolute value]* actual learning is the outcome of interest.
   (b) *[relative value]* attaining positional goods as a result of learning (e.g., college admission) is the outcome of interest i.e., relative or competitive value of the learning.
4. *[All-Things-Considered Benefit]* Using an AI intervention makes a group's learning worse in the short-term but better in the long-term.
5. *[Unequal Benefit to Redress Historical Bias]* Using an AI intervention makes everyone's learning better off than before but improves learning more for the historically disadvantaged group than the other.
6. *[Actual vs. Risk of Harm]* Using an AI intervention has a very high probability to make learning better off than before but:

(a) *[slight risk of failure]* there's also a low probability that learning will be harmed.
(b) *[slight risk of exposure to toxicity]* there is some small risk that students are exposed to racist, sexist, or xenophobic content.

Across these cases, benefits and harms are primarily defined in terms of academic learning outcomes, reflecting the central goals of schools and classrooms. However, participants were also encouraged to raise concerns that extend beyond learning performance when relevant to their reasoning. In particular, some discussions surfaced potential non-learning harms, including concerns related to students' dignity, procedural treatment, or recognition as epistemic agents. These considerations were not introduced as additional cases, but rather emerged as secondary concerns that participants invoked when interpreting the core scenarios [4, 5]. Together, the scaffold supports comparative reasoning across cases while leaving space for participants to surface broader conceptions of unfairness and harm.

2.2 Reflective Equilibrium as a Method to Elicit AI Fairness Conceptions

"Fairness" is a concept about how benefits and burdens should be distributed, so it should not be surprising that people differ in their conceptions of it, which are liable either to generate, or reflect, different views about what the right distribution of burdens and benefits is, about what the right method is for determining that distribution, and/or whether they should be distributed at all. Simultaneously, people may also differ in their conceptions of what counts as benefits or burdens. We propose a method (reflective equilibrium) and a practical scaffold readers can use (ideally with others with whom they disagree) in attempting to make progress in determining what the correct (or at worst a good) conception—is, in their particular contexts.

Reflective equilibrium [1] is a method for moral reasoning that helps agents deliberate about questions of value when no fixed algorithm or authoritative ranking of values is available. Rather than prescribing particular moral conclusions, it provides resources for responsible agents—such as individuals, institutions, or governments—to reason for themselves. It begins with two fallible kinds of judgments: judgments about particular cases and judgments of general principles. When enough of these judgments are considered together, inconsistencies typically emerge. Since inconsistency implies falsehood, uncovering and resolving such conflicts is a central task of moral philosophy.

The method is not merely logical but evaluative. Once contradictions are identified, reasoners attempt to determine which judgments are least reliable and should therefore be revised or rejected. Judgments are considered less reliable when they align with self-interest, are inherited uncritically from social environments, are reasonably disputed by others, or are unstable over time. None of these conditions is decisive on its own, but they provide comparative reasons

to favor some judgments over others when resolving inconsistencies. Crucially, reflective equilibrium allows for the temporary suspension of judgment rather than outright rejection. Further reflection on the reasons involved may justify retaining or reinstating a belief once it is better supported.

Although reflective equilibrium can be described as an individual reasoning process (akin to other fairness elicitation methods; [3]), it is ideally collaborative [21, 25]. Our thought is that dialogue can be more productive when, rather than insisting that their conception is 'the correct one' or, worse, not even recognizing that their conceptions are different, participants start out with an understanding of how the conceptions they bring to the discussion differ from each other's. Everyone has insights that are due to their background, upbringing, environment, cultures, and particular experiences. And everyone has blindspots that are due to the same things. So the ideal reflective equilibrium involves multiple participants, as different in environment, culture, outlook, experience, and ideology as is compatible with effective communication, but all of whom are committed to honest and self-critical collective deliberation. The expectation is that in such a process it will be more common for better reasons and better ideas than for worse reasons and worse ideas to prevail.

3 Methods

3.1 Participants and Contexts

The study recruited 29 teachers from two educational contexts, the United States (n = 14) and the Philippines (n = 15). Participants were recruited through partner institutions and participated voluntarily. Teachers represented a range of subject areas and instructional settings, as shown in Table 1. Nearly all participants reported prior experience using AI-powered tools, with only one participant in the United States and one in the Philippines indicating no prior exposure. The most common uses involved generative text and chat tools ($N = 27$) for brainstorming, drafting, or refining materials, as well as writing support and lesson planning or curriculum development.

3.2 Procedure

Each workshop followed the same structured procedure across locations. Participants first completed a pre-survey that captured their initial judgments about fairness-related concerns in AI-supported instructional contexts. They were then organized into small groups of three to four teachers. Within these groups, participants discussed the scenarios from the *AI fairness elicitation scaffold*. Discussions centered on whether and why particular scenarios raised fairness-related concerns, and on how contextual factors shaped participants' reasoning. This phase provided a focused setting for exchanging perspectives and surfacing points of convergence, divergence, or uncertainty prior to broader deliberation.

The workshop then transitioned to a whole-group discussion structured as a facilitated fishbowl-style activity. Representatives from each small group shared

Table 1. Participant demographics by geographic context

Characteristic	US (n=14)	Philippines (n=15)
Primary Role		
K–12 Classroom Teacher	11	5
Higher Education Faculty / Instructor	2	9
Administrative / Support Role	1	1
Years of Teaching Experience		
1–3 years	1	1
4–6 years	1	1
7–10 years	0	3
11–15 years	2	4
16–20 years	4	2
More than 20 years	6	4

salient perspectives with the full workshop, while facilitators prompted comparison across scenarios and encouraged reflection on similarities, differences, and unresolved tensions in participants' reasoning. The fishbowl format was designed to surface a range of viewpoints and reasoning strategies without requiring consensus or privileging particular positions.

Participants completed the post-survey at the conclusion of the workshop discussion. All sessions were audio recorded and transcribed for subsequent qualitative analysis.

3.3 Quantitative Measures and Qualitative Data

The pre–post surveys consisted of ordinal Likert-scale items designed to capture teachers' fairness-related judgments about AI interventions. Specifically, participants responded to the statement: *"Using AI tools can reinforce bias and/or unfair treatment of students in the following aspects of my teaching."* Responses were collected across six instructional aspects: Assessment, Classroom Management, Discussion Facilitation, Lesson Planning, Personalized or Adaptive Instruction, and Teacher Reflection. Each item was rated on a four-point Likert scale ranging from *Strongly Disagree* to *Strongly Agree*.

Qualitative data consisted of transcripts from the facilitated workshop discussions. We analyzed the workshop transcripts using thematic analysis to identify recurring patterns in how teachers reasoned about unfairness and harm when evaluating AI interventions. The analysis focused on characterizing forms of reasoning and sensemaking rather than categorizing participants according to fixed fairness positions.

Members of the research team independently reviewed the transcripts to identify segments in which participants articulated fairness-related concerns. Through iterative discussion, the team refined a set of themes that captured common reasoning patterns observed across scenarios. Quotes presented in Section 4 are labeled with anonymized participant identifiers and geographic context.

4 Results

This section examines how teachers reasoned about unfairness and harm when evaluating AI interventions across structured classroom scenarios, drawing on both survey responses and workshop discussions. We report quantitative pre–post changes in survey responses (Section 4.1) and then present qualitative findings that trace recurring patterns in teachers' reasoning.

4.1 Pre–Post Changes in Fairness-Related Judgments Across Aspects and Contexts

We first examined overall pre–post changes in teachers' fairness-related judgments using a Wilcoxon signed-rank test. The results indicated a significant positive shift from pre to post ($n = 29$, $p = 0.0039$), with a large effect size ($r = 0.62$). This effect reflected an imbalance in the direction of individual change, with more participants shifting toward stronger agreement with the survey statement regarding AI's potential to reinforce bias or unfair treatment. We next examined whether these response patterns differed by geographic location (US vs. Manila). Because responses were measured on an ordinal Likert scale and group sizes were modest, we used Mann–Whitney U tests to compare independent groups. We found no significant difference between US and Manila participants at a single time point ($p = 0.30$). We further compared individual pre–post change scores across locations and again found no significant difference in either the magnitude or direction of change ($U = 110$, $p = 0.82$). Together, these results indicate that the observed pattern of pre–post change was similar across geographic contexts, suggesting that increased articulation of fairness-related concerns was not specific to one location.

To examine whether this pattern held across different instructional aspects, we conducted Wilcoxon signed-rank tests separately for each aspect. Across all six aspects, median responses remained unchanged from pre to post ($\text{Median}_{\text{pre}} = 3$, $\text{Median}_{\text{post}} = 3$), indicating that the central tendency of responses did not shift. However, Five of the six aspects showed statistically significant positive pre–post changes: Assessment ($p = 0.0039$), Classroom Management ($p = 0.0124$), Discussion Facilitation ($p = 0.0075$), Lesson Planning ($p = 0.0209$), and Personalized or Adaptive Instruction ($p = 0.0075$). Teacher Reflection exhibited a marginal effect that did not reach conventional significance ($p = 0.0522$). As shown in Figure 1, these effects were not driven by uniform shifts across all participants, but by a consistent directional pattern: across instructional aspects, more participants increased their responses from pre to post than decreased them, while some reported no change. Downward shifts were rare. This pattern indicates that participation in the workshop was associated with a greater tendency for teachers to recognize or articulate potential fairness-related concerns across multiple areas of teaching.

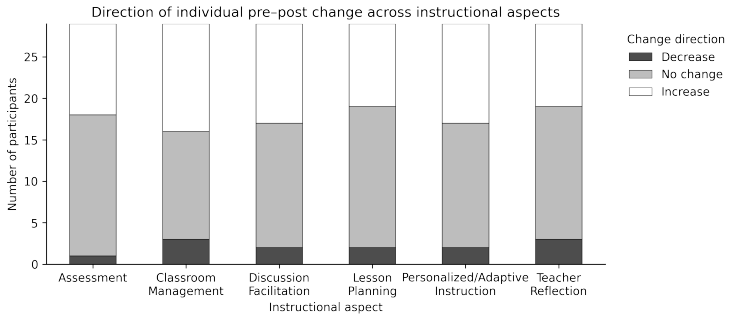

Fig. 1. Direction of individual pre–post changes across instructional aspects. Bars indicate the number of participants whose responses increased, decreased, or remained unchanged from pre to post. Across all aspects, upward changes were more common than downward changes, despite stable median responses.

4.2 Fairness Concerns Are Minimal When Learning Benefits Do Not Create Harm

In scenarios where an AI intervention was described as improving learning outcomes without disadvantaging any group, participants largely expressed agreement that the intervention did not raise substantial fairness concerns. Our results show that teachers tended to treat the absence of harm or loss as a key condition under which learning benefits were seen as compatible with fairness.

For example, in cases where one group's learning improved while other groups' learning outcomes remained unchanged, participants often framed the intervention as acceptable or unproblematic. One teacher noted: *"If no one is actually losing anything and some students are benefiting, I don't really see where the fairness issue is."* (P04)

Similarly, participants frequently distinguished these scenarios from others involving explicit tradeoffs, emphasizing that fairness concerns arose primarily when improvements came at a cost to others: *"This feels different from the other examples, because here nobody is worse off. It's just extra support for some students."* (P12)

Cases involving unequal benefit intended to redress historical disadvantage also tended to elicit broad agreement, particularly when participants viewed the intervention as compensatory rather than exclusionary. Several teachers explicitly connected these scenarios to existing equity-oriented practices in education. As one participant explained: *"We already do things like this all the time—extra resources for students who've been underserved. So this doesn't feel unfair to me."* (P07)

Rather than interpreting unequal improvement as inherently problematic, participants in these cases often evaluated fairness in relation to broader educational goals, such as reducing achievement gaps or providing targeted support. Another teacher described this reasoning as follows: *"If the goal is to help stu-*

dents who are behind catch up, then it makes sense that they might benefit more. That's kind of the point." (P09)

Across these scenarios, participants rarely expressed uncertainty or hesitation. Instead, they articulated fairness judgments with confidence, often contrasting these cases with others that involved harm, risk, or competitive disadvantage. Taken together, these findings suggest that when learning benefits do not introduce harm or worsen outcomes for any group, teachers' fairness judgments tend to converge, treating benefit and fairness as aligned rather than in tension.

4.3 Learning Benefits Alone Do Not Resolve Fairness Disagreements

Across several scenarios, participants diverged in their fairness judgments when learning benefits were accompanied by harm, risk, or unequal consequences. In these cases, improvements in overall learning outcomes were not sufficient to produce consensus, and participants articulated competing interpretations of what fairness required.

In scenarios where an AI intervention improved learning for one group while disadvantaging another relative to a feasible alternative, some participants emphasized aggregate learning gains as the primary consideration. For example, one participant argued that increased overall performance justified adoption despite unequal effects: *"If most students are learning more, that still feels like a positive outcome overall, even if it's not perfect for everyone."* (P01)

In contrast, others rejected this reasoning, foregrounding harm to the worse-off group as a decisive fairness concern: *"I'm uncomfortable with the idea that one group is learning less just so another can learn more. That feels like we're sacrificing some students for efficiency."* (P08)

Disagreement was particularly pronounced in cases involving unequal benefit tied to relative or positional outcomes, such as college admission or competitive advancement. While some participants focused on absolute learning gains, others emphasized that relative advantages could exacerbate existing inequities. One teacher (P02) noted: *"Even if everyone improves, if one group gets a bigger boost, that's what matters in competitive settings. That advantage carries forward."*

Similarly, scenarios that introduced risk of harm, such as a small probability of exposure to biased or toxic content, surfaced contrasting risk tolerances. Some participants viewed low-probability harms as acceptable tradeoffs: *"There's always some risk with new tools. If the likelihood is small and the learning benefit is real, I'd still consider using it."* (P13) Others treated even minimal risks as unacceptable, particularly when harms implicated dignity or identity: *"Even a small chance of exposing students to racist content feels serious to me. That's not something I'm willing to gamble with."* (P08)

Across these cases, participants' disagreements did not stem from misunderstanding the scenarios but from different prioritizations of benefit, harm, and risk. While some framed fairness in terms of maximizing learning outcomes, others emphasized non-maleficence, protection of vulnerable students, or the broader social consequences of unequal or risky interventions. Together, these

findings indicate that learning benefits alone do not resolve fairness questions when interventions introduce harm, tradeoffs, or unequal consequences.

4.4 Teachers Reason About Fairness Through Conditional Tradeoffs Rather Than Binary Judgments

Across multiple scenarios, participants rarely framed fairness judgments as simple decisions of agreement or disagreement. Instead, they engaged in conditional reasoning, articulating the circumstances under which an AI intervention might be acceptable, problematic, or require additional safeguards. These patterns were particularly salient in cases involving unequal benefit, delayed outcomes, or probabilistic risks.

In scenarios where an AI intervention produced unequal benefits but improved learning outcomes overall, participants frequently distinguished between absolute learning gains and relative or competitive consequences. Rather than treating these cases as uniformly fair or unfair, teachers reasoned through how outcomes would be interpreted in different institutional contexts. One participant explained: *"If we're just talking about students understanding the material better, then the unequal improvement might be okay. But if this affects who gets into college, that changes how I think about it."* (P11)

Similarly, in cases involving short-term harm in exchange for long-term benefit, participants emphasized temporal tradeoffs. Some expressed openness to temporary setbacks if longer-term learning gains were expected, while others questioned whether such tradeoffs were justifiable without strong evidence. As one teacher (P04) noted: *"I could maybe accept a short-term dip if I knew it really helped them later, but that's a big 'if.' I'd want to be sure before going down that road."*

Moreover, scenarios that introduced probabilistic risks, such as a small chance of exposing students to biased or harmful content, further illustrated participants' conditional reasoning. Rather than responding uniformly to the presence of risk, teachers weighed likelihood, severity, and the availability of mitigation strategies. One participant (P09) articulated this balancing process explicitly: *"It's not just whether there's a risk, but how likely it is and what kind of harm we're talking about. Those things matter a lot."*

Across these cases, expressions of uncertainty were common, but they did not signal indecision or confusion. Instead, participants used uncertainty to surface underlying concerns and clarify what information or conditions would be necessary to reach a judgment. As one teacher (P04) reflected during discussion: *"I don't have a clear yes or no here, but talking it through makes me realize what I'd need to know before feeling comfortable."*

Taken together, these findings indicate that teachers' fairness reasoning is fundamentally contextual and conditional, shaped by considerations such as time horizon, competitive stakes, and risk mitigation. Rather than applying fixed rules, participants treated fairness judgments as contingent on how benefits and harms unfold in specific educational settings. This pattern helps explain why

median survey responses remained stable even as individual-level shifts were observed, reflecting refinement and articulation of reasoning rather than categorical changes in stance.

4.5 Comparing Benefit–Fairness Tradeoffs Helps Teachers Articulate Fairness Boundaries

As participants worked through multiple scenarios, they frequently referenced earlier cases to refine, qualify, or revise their fairness judgments. Rather than evaluating each scenario in isolation, teachers used comparison across cases to articulate the boundaries of what they considered acceptable or problematic. This comparative reasoning allowed participants to clarify which dimensions of benefit, harm, and risk were most salient to their fairness judgments.

Several participants explicitly contrasted cases where learning benefits were unaccompanied by harm with those involving tradeoffs earlier agreement as a reference point for later disagreement. One teacher (P14) noted: *"I was okay with the first example because no one was losing anything. But this one feels different, because now someone actually ends up worse off."*

Others used comparison to distinguish between different kinds of unequal benefit, particularly when reflecting on absolute versus relative outcomes. As one participant (P11) explained while revisiting an earlier case: *"At first I thought unequal improvement was fine, but then when I compared it to the college admissions example, I realized that context really changes how unfair it feels."*

In some instances, participants described how moving across scenarios helped them surface concerns that were initially implicit. One teacher (P08) reflected: *"I didn't think much about risk in the beginning, but after seeing the later cases, I realized that even a small chance of harm matters more to me than I expected."*

These cross-case reflections were not limited to shifts in agreement or disagreement. Instead, participants often used comparison to articulate conditional criteria, such as the role of safeguards, institutional context, or the reversibility of harm. As another participant (P03) summarized during discussion: *"Seeing all of these together makes me realize I don't have one rule for fairness. It depends on what kind of benefit it is and who's affected."*

Overall, comparison across cases functioned as a mechanism for sensemaking, enabling participants to move beyond initial reactions toward more articulated and differentiated fairness reasoning. Rather than converging on a single definition of fairness, teachers used the contrast between scenarios to clarify the conditions under which learning benefits aligned with or conflicted with their fairness concerns.

5 Discussion

We conducted a mixed-methods study combining surveys and workshops to examine how teachers reason about unfairness and harm when evaluating AI interventions in educational contexts. The study is analytically informed by reflective equilibrium, which we adopt as an orientation for eliciting and analyzing

teachers' reasoning rather than as a normative endpoint. Reflective equilibrium highlights the role of comparison across cases in surfacing and refining judgments about complex ethical questions. In this study, this orientation informed both the design of the *AI fairness elicitation scaffold* (see Section 2.1) and our analytic focus on how participants compared scenarios to reason about whether and why particular outcomes raised fairness-related concerns.

Rather than evaluating the correctness of specific fairness principles or the fairness of particular AI systems, our approach focuses on empirically characterizing how teachers articulate and negotiate fairness-related concerns across structured scenarios. These scenarios systematically vary the configuration of learning benefits, risks, and time horizons, allowing us to examine how teachers' reasoning shifts across contexts and tradeoffs.

5.1 Fairness Reasoning as Conditional and Context-Sensitive, Not Rule-Based

Teachers in this study did not treat fairness as a fixed rule that could be applied uniformly across AI interventions. Instead, fairness reasoning emerged as conditional and context-sensitive, shaped by how anticipated learning benefits interacted with potential harms, risks, and burdens in particular classroom settings. This finding reinforces the need for bottom-up approaches to AI fairness that prioritize stakeholders' expertise rather than relying on decontextualized definitions of responsible AI values.

Educators are not passive recipients of AI technologies but active mediators of how AI systems are interpreted, implemented, and experienced in classrooms [17]. Prior work has shown that teachers' practices significantly shape AI effectiveness and student outcomes [8, 24]. Our findings extend this work by showing that teachers also play a critical role in reasoning about fairness, drawing on their knowledge of students, instructional goals, and classroom dynamics to evaluate whether learning benefits are achieved at an acceptable cost.

Our results show that participants did not converge on a single fairness judgment. Instead, they refined the conditions under which learning gains should or should not count as fair. Fairness concerns were articulated relative to classroom context, student population, and instructional purpose, consistent with prior findings that answers to questions about fair AI use vary across learning goals and sociopolitical contexts [23]. These patterns suggest that fairness in educational AI cannot be meaningfully assessed without attending to the situated expertise of those who enact AI systems in practice.

5.2 When Benefits Legitimize Inequality, and When They Do Not

Our study shows that unequal learning outcomes were not inherently viewed as unfair. In several scenarios, teachers treated unequal benefit as legitimate or even desirable when it aligned with broader educational aims, such as remediation or addressing historical disadvantage. This reasoning reflects well-established distinctions in education and political philosophy between equality and equity,

where differential treatment is justified to promote fair opportunity or redress prior injustice [18].

However, unequal benefit became contested when it was coupled with additional burdens or when the meaning of learning outcomes shifted. Scenarios involving direct harm to another group, increased exposure to risk, or competitive and positional stakes generated more disagreement. In these cases, teachers questioned whether learning gains for some could justify losses, risks, or diminished opportunities for others.

Teachers also distinguished between different interpretations of learning value. When learning was framed in absolute terms, unequal improvement was often tolerated or endorsed. When learning outcomes were tied to relative or competitive advantages, participants were more likely to view unequal benefit as problematic. This implication echoes work in education policy and sociology that highlights how the stakes attached to learning outcomes shape perceptions of fairness and harm [19].

5.3 Implications for Designing and Evaluating AI in Education

These findings suggest that evaluating fairness in educational AI systems requires attention beyond aggregate learning gains. Even when AI interventions improve outcomes, teachers attend closely to how benefits and burdens are distributed, who bears risk, and whether tradeoffs align with pedagogical goals. This reinforces critiques of performance-driven approaches to educational technology evaluation, which risk overlooking ethical and distributive consequences that matter in practice.

Second, our study highlights the value of deliberative and scenario-based approaches for fairness assessment. Rather than requiring binary judgments, tools that surface tradeoffs and support comparison across cases may better reflect how fairness reasoning operates in educational settings. This aligns with recent calls in AI governance and HCI for participatory, context-aware evaluation methods that foreground stakeholder reasoning [14, 13].

Finally, these findings suggest that teacher-facing governance and professional development efforts should treat fairness as an ongoing reasoning process rather than a checklist. Supporting teachers in articulating and negotiating benefit–burden tensions may be more productive than prescribing fixed fairness criteria, particularly as AI systems are deployed across diverse instructional contexts and student populations.

In conclusion, this paper examined how 29 teachers from two educational contexts, the United States and the Philippines, reason about unfairness and harm when evaluating AI interventions in education. Using a mixed-methods design, we show that fairness judgments are shaped not by learning benefits alone, but by how benefits interact with harms, burdens, risks, and instructional goals. Teachers articulated conditional, context-sensitive reasoning, endorsing unequal benefit in some cases while expressing concern when benefits were paired with harm, risk, or competitive stakes. Methodologically, this work contributes a practical, stakeholder-centered approach to fairness sensemaking by operationalizing

reflective equilibrium through structured scenario comparison, highlighting the importance of bottom-up reasoning for responsible AI design and evaluation in education. Future work will study how demographic, institutional, or cultural factors may shape differences in fairness prioritization, how the deliberation process may introduce social desirability, the long-term stability of shifts in teacher perceptions, and how to elicit and operationalize fairness notions from teachers' deliberation.

Acknowledgments. Support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation. We thank Shona Acquaye and Ajita Raghavendra for their valuable inputs. We also thank the AIED reviewers for their feedback and our teacher participants for sharing their valuable time and insights with us.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Brighouse, H.: Normative case studies and thought experiments: How they differ and why we need both. Educational Theory **74**(3), 329–339 (2024)
2. Castro, C., O'Brien, D., Schwan, B.: Egalitarian machine learning. Res Publica **29**(2), 237–264 (2023)
3. Cheng, H.F., Stapleton, L., Wang, R., Bullock, P., Chouldechova, A., Wu, Z.S.S., Zhu, H.: Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–17 (2021)
4. Crawford, K.: The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press (2021)
5. Fricker, M.: Epistemic injustice: Power and the ethics of knowing. Oxford university press (2007)
6. Guerdan, L., Saxena, D., Chancellor, S., Wu, Z.S., Holstein, K.: Measurement as bricolage: Examining how data scientists construct target variables for predictive modeling tasks. ACM on Human-Computer Interaction **9**(7), 1–37 (2025)
7. Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S.B., Santos, O.C., Rodrigo, M.T., Cukurova, M., Bittencourt, I.I., et al.: Ethics of ai in education: Towards a community-wide framework. International Journal of Artificial Intelligence in Education **32**(3), 504–526 (2022)
8. Holstein, K., McLaren, B.M., Aleven, V.: Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms. In: International conference on artificial intelligence in education. pp. 154–168. Springer (2018)
9. Hu, L.: What is "race" in algorithmic discrimination on the basis of race? Journal of Moral Philosophy **21**(1-2), 1–26 (2023)
10. Jacobs, A.Z., Wallach, H.: Measurement and fairness. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 375–385 (2021)
11. Jakesch, M., Buçinca, Z., Amershi, S., Olteanu, A.: How different groups prioritize ethical values for responsible ai. In: proceedings of the 2022 ACM conference on fairness, accountability, and transparency. pp. 310–323 (2022)

12. Jung, J.Y., Saxena, D., Park, M., Kim, J., Forlizzi, J., Holstein, K., Zimmerman, J.: Making the right thing: Bridging hci and responsible ai in early-stage ai concept selection. In: Proceedings of the 2025 ACM Designing Interactive Systems Conference. pp. 2992–3012 (2025)
13. Karumbaiah, S., Yin, Y., Bharadwaj, A.: Aibat: Ai behavior analysis tool for teacher-driven contextual evaluation of language models in education. In: Artificial Intelligence in Education. Lecture Notes in Computer Science, vol. 15877, pp. 56–71. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-98414-3_5
14. Madaio, M.A., Stark, L., Wortman Vaughan, J., Wallach, H.: Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. p. 1–14. CHI '20, New York, NY, USA (2020)
15. Microsoft: Responsible AI principles and approach. https://www.microsoft.com/en-us/ai/principles-and-approach (2025), accessed on October 9, 2025
16. OECD: Oecd ai principles. organisation for economic co-operation and development. https://oecd.ai/en/ai-principles (2025), accessed on October 9, 2025
17. Puntambekar, S.: Distributed scaffolding: Scaffolding students in classroom environments. Educational Psychology Review **34**(1), 451–472 (2022)
18. Rawls, J.: A Theory of Justice. Harvard University Press (1971)
19. Reay, D.: What would a socially just education system look like? Sociology **46**(4), 587–599 (2012)
20. Selwyn, N.: The future of ai and education: Some cautionary notes. European Journal of Education **57**(4), 620–631 (2022)
21. Shen, H., Deng, W.H., Chattopadhyay, A., Wu, Z.S., Wang, X., Zhu, H.: Value cards: An educational toolkit for teaching social impacts of machine learning through deliberation. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 850–861 (2021)
22. Tabassi, E.: Artificial intelligence risk management framework (ai rmf 1.0) (2023)
23. Yin, Y., Karumbaiah, S., Acquaye, S.: Responsible ai in education: Understanding teachers' priorities and contextual challenges. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. pp. 2705–2727 (2025)
24. Yin, Y., Karumbaiah, S., Lim, J., Acquaye, S., Schwartz, S.: Toward teacher-centered ai design: Exploring the role of pedagogical values and contextual factors in k-12 teachers' perceptions of responsible ai. In: Proceedings of the 18th Computer-Supported Collaborative Learning, pp. 196-204 (2025)
25. Zhang, A., Walker, O., Nguyen, K., Dai, J., Chen, A., Lee, M.K.: Deliberating with ai: improving decision-making for the future through participatory ai design and stakeholder deliberation. Proceedings of the ACM on Human-Computer Interaction **7**(CSCW1), 1–32 (2023)