Understanding And Improving Student Recognition of A.I Bias in Educational Contexts

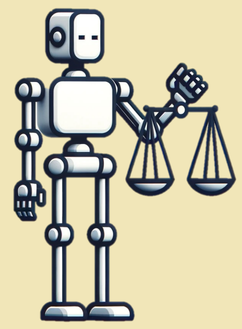
Author: Dao Hai Hoang
Master of Management in Data Science
Supervisor: Prof. Dr. Claire Deventer

What is *A.I Bias* in the education?

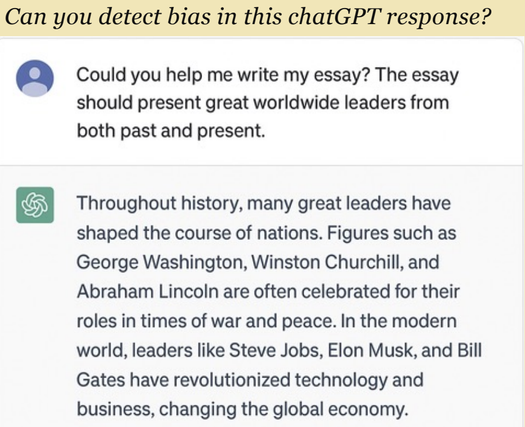
The application of artificial intelligence to education (AIEd) has been studied for decades, examining learning in various settings to support both formal education and lifelong learning (Luckin et al., 2016).

However, despite the benefits of AI systems in education, they are not free from imperfections and introduce harmful biases, *as decision-making and predictive computer algorithms are often perceived as inherently unfair and partial* (M. K. Lee, 2018).

These biases potentially mislead students and can amplify existing inequalities in educational contexts (Baidoo-Anu & Ansah, 2023), leading to discrimination against marginalized groups and new forms of discrimination based on skin color, ethnicity, or physical appearance (Ferrara, 2024).



Can you detect bias in this chatGPT response?



The Problem

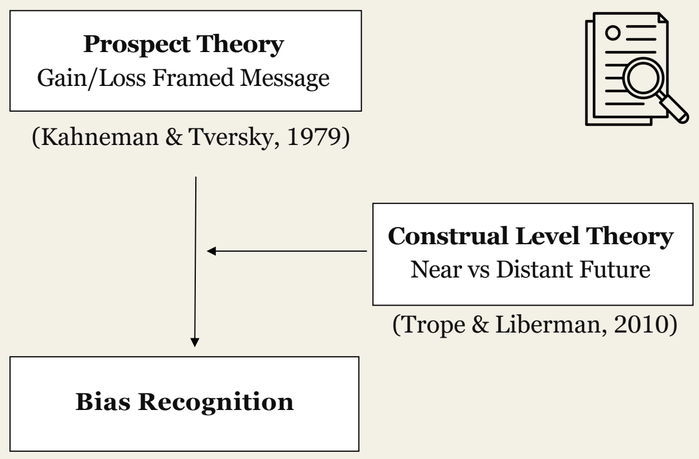
- *Research has examined algorithm bias from the **system-design perspectives** (Suresh & Guttag, 2021; Lee et al., 2024) and in education (Baker & Hawn, 2022).*
- *Some has explored how end users perceive algorithm bias (Noble, 2018) in different fields.*
- ⁉ But **few** have explored how end users like **students - recognize algorithm biases** in educational contexts.
- ⁉ This gap is critical in educational settings where students increasingly rely on AI tools such as ChatGPT.
- ⁉ Unrecognized AI bias could diminish trust in educational technologies, leading to resistance against their adoption and undermining their potential benefits.

The Study

Recognizing the need to better understand user-side detection of bias, this study investigates how students' recognition of AI bias is influenced by the framing of warning messages with the core research question:

How does the design of a warning message influence the students recognition of AI bias?

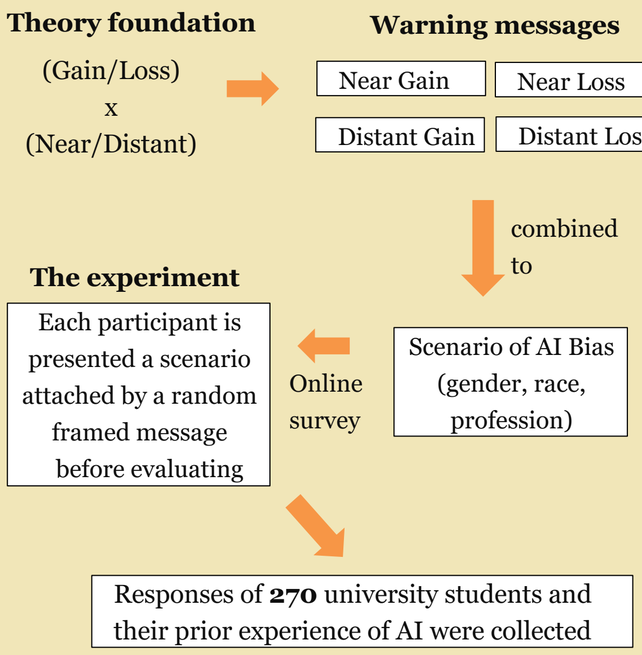
The research model is drawn on the literature of Prospect Theory (Kahneman & Tversky, 1979), Construal Level Theory (Trope & Liberman, 2010) and warning messages.



Hypotheses:

- Loss-framed messages will lead to significantly higher bias recognition of compared to gain-framed messages.
- There will be a significant interaction between near-distant and gain–loss framing on bias recognition.

Research Method



Findings

- **Loss-framed messages significantly enhanced** students' ability to recognize AI bias across all three domains: gender, race, and profession.
- While the main effect of temporal distance was generally weaker, it showed a **significant impact** at 0.1 level for race and profession bias detection **when combining with loss/gain frame**.
- Although most participants reported high familiarity and persuasion literacy with AI tools, **their knowledge of AI bias was notably lower.**
- Students consider AI responses are moderately useful and reliable.

ANOVA test	NearDistant	GainLoss	NearDistant * GainLoss
Gender bias	p=0.068	p=0.000 M(L)=4.945 M(G)=4.196	p=0.149
Race bias	p=0.023	p=0.002 M(L)=4.979 M(G)=4.415	p=0.077
Profession bias	p=0.572	p=0.025 M(L)=4.824 M(G)=4.426	p=0.058

Descriptive Analysis	Mean	Min	Max	S.Error	Cronbach's Alpha
Usefulness	4.130	3.922	4.937	0.214	0.851
Reliability	4.034	3.681	4.433	0.288	0.904
Persuasion Literacy	5.001	4.970	5.059	0.179	0.698
AI familiarity	5.02	1	7	0.061	N/A
AI bias familiarity	3.83	1	7	0.093	N/A

Managerial Implications

1. Warning messages should **highlight the risks, harms, or negative consequences** of skipping AI Bias rather than gain-based alternatives.
2. **Psychological distancing** may not be sufficient as a standalone strategy but may serve as **a useful framing amplifier when paired with consequence-based messaging**.
3. **Build students' capacity to recognize persuasive message,** not solely alerting them to bias. This layered approach supports not just awareness, but more critical engagement with AI tools in education.

Limitations & Future Research

For a more comprehensive understanding of how students detect algorithmic bias, future research could:

- Explore psychological distance beyond temporal distance such as including spatial, social, and hypothetical distance.
- Conduct deeper behavioral tasks or qualitative research methods & Expand the sample size.
- Explore more bias forms beyond gender, race, and profession bias (e.g, religious, etc.).

References
Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI*.
Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci, 6(1), Article 1.
Kahneman, & Tversky. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*, 263–292.
Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684.
Luckin, R., Holmes, W., Griffiths, M., & Corcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.
Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York university press.
Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9.
Trope, Y., & Liberman, N. (2010). Construal-Level Theory of Psychological Distance. Psychological Review, 117(2), 440–463.