



UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

Understanding and improving student recognition of AI bias in educational contexts

Hai Hoang Dao

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Data Science

SUPERVISOR :

Prof. dr. Claire DEVENTER



UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

2024
2025



Faculty of Business Economics

Master of Management

Master's thesis

Understanding and improving student recognition of AI bias in educational contexts

Hai Hoang Dao

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Data Science

SUPERVISOR :

Prof. dr. Claire DEVENTER

UNDERSTANDING AND IMPROVING STUDENT RECOGNITION OF AI BIAS IN EDUCATIONAL CONTEXTS

ABSTRACT

As large language models (LLMs) become increasingly integrated into higher education, concerns about the bias of their outputs have grown. While much research has focused on detecting algorithmic bias from a system perspective, fewer studies have explored users' ability - particularly students' - to recognize bias in AI-generated content. This study investigates how message framing can influence students' recognition of AI bias, integrating Prospect Theory and Construal Level Theory to examine the effects of gain-loss framing and temporal distance (near vs. distant) on bias detection. A 2x2 between-subjects experimental design was conducted among participants who responded to a generate biased content. Participants were exposed to different combinations of framing and were then asked to assess the presence of gender, race, and profession bias in the responses. The results showed that loss-framed messages significantly enhanced students' ability to detect bias across all three domains, supporting the influence of emotional salience as predicted by Prospect Theory. Moreover, significant interaction effects between framing type and temporal distance were found for race and profession bias, suggesting that framing strategies may be more effective when combined - particularly when emphasizing both loss and immediacy. While students reported high familiarity with AI, their knowledge of AI bias was notably lower, indicating a knowledge gap that may limit critical engagement. These findings contribute to the literature on AI bias detection in educational contexts by demonstrating that well-framed warning messages can serve as cognitive prompts, improving bias detection in student-AI interactions. The study recommends the use of loss-based, contextually immediate warnings, alongside brief educational message, to support effective AI use in learning environments.

Table of Contents

1. INTRODUCTION	3
1.1 Background.....	3
1.2 Research gaps and questions.....	3
2. LITERATURE REIVEW	9
2.1 Large Language Models (LLMs) usage in education	9
2.2 Algorithm bias of LLMs.....	10
2.3 Theoretical foundation	11
3. Research Model & Hypothesis Development	14
3.1 Research Model	14
3.2. Hypothesis Development.....	15
4. Research Method.....	16
4.1. Expriemental Design	16
4.2 Sample.....	19
4.3 Research Process	20
5. Data Analysis & Results.....	20
5.1 Scale Reliability Test	20
5.2 Overall Descriptive Analysis.....	23
5.3 Two-way ANOVA.....	24
6. Discussion & Conclusions.....	29
6.1 Managerial Implications	29
6.2 Contribution to Research.....	30
6.3 Limitations and recommendation for future research.....	31
6.4 Conclusions.....	32
References	33
Appendices	42
Appendice A. Screenshot of Scenarios AI generated responses with warning messages	42
Appendice B. Survey Questionnaire.....	43
Appendice C. Survey Results.....	45

List of Figures

Figure 1 Research Model	15
Figure 2 Scenario screenshot with Near Loss framed message.....	18
Figure 3 Scale Reliability Test Result - Usefulness.....	21
Figure 4 Scale Reliability Test Result - Reliability	22
Figure 5 Scale Reliability Test Result - Persuasion.....	23
Figure 6 Overall Descriptive Analysis.....	23
Figure 7 ANOVA Test Result - Gender Bias.....	25
Figure 8 ANOVA Test Result - Race Bias	26
Figure 9 ANOVA Test Result - Profession Bias.....	28

List of Tables

Table 1 Recap Literature Review and Research Questions	4
Table 2 Framed Messages	16

1. INTRODUCTION

1.1 Background

AI and its applications such as Large Language Models (LLMs) is expected to expand from \$6.4 billion in 2024 to \$36.1 billion by 2030 (Research and Markets, 2024), and has been revolutionizing various aspects of most industries including education. Teaching methods, learning environments and educators' policy has been adapted and fast-transformed in the situation of post-pandemic and globalization. According to Global Market Insights Inc. (2023), the Artificial Intelligence (AI) in Education market was valued at \$4 billion in 2022 and is expected to grow at a compound annual growth rate (CAGR) exceeding 10% from 2023 to 2032, driven by the increasing preference for personalized learning. The usage of LLMs like ChatGPT by OpenAI or Gemini by Google have introduced significant efficiency and improvement into educational practices by teachers, students and educators as well (Yang et al., 2021), creating opportunities for inclusive and dynamic classroom environments (Baidoo-Anu & Ansah, 2023).

However, despite the benefits of AI systems in education, they are not free from imperfections and introduce harmful biases. Although algorithms play a significant role in shaping various aspects of our daily lives, decision-making and predictive computer algorithms are often perceived as inherently unfair and partial (M. K. Lee, 2018). This challenges the effectiveness, validity, and ethical standards of using the AI Tools (Holmes et al., 2022), particularly in international higher education where students come from varied cultural, social-economical, and academic. A study discovered that when ChatGPT is assigned a specific personality, the likelihood of generating toxic responses such as incorrect stereotypes, harmful dialogue, and hurtful opinions significantly increases (Deshpande et al., 2023). This reduces the potential of AI by producing distorted results, leveraging mistrust and challenges the improvement of human decision-making (Manyika et al., 2019).

Moreover, 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). And these bias, which is caused by algorithm procedures, present in any real-world education dataset (Sha et al., 2022).

1.2 Research gaps and questions

The mechanism and process of algorithm bias has been studied by various researchers, showing that it can possibly be occurred in every stage in machine learning lifecycle—ranging from data collection to model deployment inequities (Suresh & Guttag, 2021). Prior to that, Noble (2018) shows early concepts on how biases presents in search engines like Google, following by Lee et al. (2024) reviews the existing research and categorize LLMs bias in educational context in two use cases: natural language generation (NLG) and natural language understanding (NLU) tasks and considered them with two broader types of bias is representational biases and allocative biases. Theses studies are fundamental for studies of AI Bias in specific contexts and fields with

examined impacts on stakeholders, including in educational context which is summarized in the below table:

Table 1 Recap Literature Review and Research Questions

Paper	Context	Field	AI Bias influence on users	AI Bias Recognition from users	AI Bias Recognition in education	Findings
Suresh & Guttag (2021)	Algorithm Bias Mechanism	Economics & Society	No	No	No	Emphasise the importance of addressing seven harms/biases systematically, from the very beginning of the ML lifecycle, to avoid perpetuating or exacerbating existing inequities
Lee et al. (2024)	Algorithm Bias Mechanism	Economics & Society	No	No	No	Identifies two main types of bias in LLM-based education applications: representational bias (e.g., stereotyping or exclusionary language in generated content) and allocative bias (e.g., unfair distribution of resources or opportunities, such as biased grading or feedback).
Vincent-Lancrin & Vlies (2020)	Bias in LLMs model	Economics & Society	Yes	No	No	Reaching the full potential of AI requires that stakeholders trust not only the technology, but also its use by humans. This raises new policy challenges around "trustworthy AI", encompassing the privacy and security of data, but also possible wrongful uses of data leading to biases against individuals or groups
Baidoo-Anu & Ansah (2023)	Bias in LLMs model	Education	Yes	No	No	Identify potential benefits and drawbacks of ChatGPT in promoting teaching and learning.
Sheng et al. (2021)	Bias in LLMs model	Education	Yes	No	No	Identifies key challenges and progress in understanding biases in language generation. It shows that biases in language models often reflect those present in training data, which leads to skewed outputs that disproportionately harm

Paper	Context	Field	AI Bias influence on users	AI Bias Recognition from users	AI Bias Recognition in education	Findings
						marginalized groups. Some methods for bias mitigation, such as adversarial debiasing and data augmentation, are found to be effective but not foolproof.
Caines et al. (2023)	Bias in LLMs model	Education	Yes	No	No	Explore the integration of large language models (LLMs) into AI-driven language education system risks and ethical considerations of deploying generative AI in language learning, emphasizing the need to mitigate issues like misinformation and harmful biases.
Manyika et al. (2019)	AI Application	Business	Yes	No	No	Examine the dual role of artificial intelligence (AI) in both mitigating and amplifying human biases.
Idowu et al. (2024)	AI Application	Education	Yes	No	No	Examine biases in AI algorithms used to monitor student progress, focusing on attributes such as age, disability, and gender.
Luckin et al. (2016)	AI Application	Education	Yes	No	No	AI can help tailor education to individual students' needs, support teachers by reducing administrative burdens, and foster better learning outcomes through data-driven insights. However, it emphasizes that this potential can only be harnessed effectively when implemented with careful attention to ethical considerations, teacher training, and a clear understanding of AI's role as an assistive, rather than a replacement, tool in education.
Yang et al. (2021)	AI Application	Education	Yes	No	No	AI holds great potential for improving education by

Paper	Context	Field	AI Bias influence on users	AI Bias Recognition from users	AI Bias Recognition in education	Findings
						personalizing learning, increasing efficiency, and supporting teacher-student interactions. However, AI's misuse, such as algorithmic bias, could exacerbate inequality and undermine human rights, including issues related to gender and race.
Farrelly & Baker (2023)	AI Application	Education	Yes	No	No	Highlights the difficulties in reliably detecting AI-generated content, raising concerns about potential false accusations against students. It also discusses biases within AI models, emphasizing the need for fairness and equity in AI-based assessments with a particular emphasis on the disproportionate impact of GAI on international students, who already face biases and discrimination.
Baker & Hawn (2022)	AI Application	Education	Yes	No	No	Review algorithmic bias in education, discussing the causes of that bias and reviewing the empirical literature on the specific ways that algorithmic bias is known to have manifested in education.
De Winter et al. (2023)	Bias & misuse detectors	Education	Yes	No	No	ChatGPT is effective at recognizing errors in student texts, particularly when given a targeted prompt. However, it is not consistently reliable in grading. These methods could help educators enhance their teaching strategies by incorporating AI tools for text recognition and error detection.
Fu et al. (2020)	Bias & misuse detectors	General	Yes	No	No	Highlight the vulnerability of current detection methods to manipulation, such as bypassing detection through

Paper	Context	Field	AI Bias influence on users	AI Bias Recognition from users	AI Bias Recognition in education	Findings
						simple prompt modifications. This points to the need for more sophisticated detection methods that consider the specific linguistic features of non-native speakers to avoid unjust penalties for these writers.
Noble (2018)	Algorithm Bias Mechanism	General	Yes	Yes	No	Examine how search engines, particularly Google, perpetuate racial and gender biases through algorithmic processes.
Xiao & Benbasat (2015)	AI Application	Business	Yes	Yes	No	Explore how warning messages can assist consumers in identifying biased product recommendation agents (PRAs) that prioritize merchant interests over consumer benefits.
Celiktutan et al. (2024)	AI Application	Economics & Society	No	Yes	No	Participants are more likely to recognize biases - such as those related to age, gender, and race - in the decisions made by algorithms than in their own judgments, even when both sets of decisions are identical.
Gonzalez et al. (2021)	AI Application	Medical	No	Yes	No	Practical strategies to help medical educators integrate implicit bias instruction throughout medical education. Their recommendations are grounded in both theoretical frameworks and empirical research, aiming to enhance equity and promote justice in healthcare.

There are numerous studies such as studies of Baidoo-Anu & Ansah (2023), Vincent-Lancrin & Vlies (2020) and Sheng et al. (2021) suggesting frameworks and approaches to mitigate these LLMs bias' adverse effects in society and educational contexts, showing that without thorough evaluation in the development of educational content, its use may unintentionally

reinforce biases, thereby misleading students or disseminating inaccurate information (Caines et al., 2023). On the other hand, studies of Luckin et al. (2016), Yang et al. (2021), Baker & Hawn (2022), Farrelly & Baker (2023) and Idowu et al. (2024) focused on examining algorithm bias in AI applications and tools usage and analyzing their effects on effectiveness and performance of users in the educational context. Meanwhile De Winter et al. (2023) and Fu et al. (2020) conducted research on the bias or misuse detectors in the perspectives of the AI providers and programmer more than from end-user's perspective.

While a growing body of research has contributed to understanding the sources of algorithmic bias, its impact on stakeholders, and approaches to mitigation, there remains a significant gap regarding the ability of end-users to recognize bias when it occurs, especially in educational settings. Although bias recognition by users was introduced in early critical scholarship such as Noble (2018), most studies have focused either on technical solutions for reducing bias or on high-level ethical concerns, rather than on how users themselves detect bias in real time. This leaves unanswered questions about how students - key stakeholders in AI-driven education - can be supported in identifying and responding to biased outputs.

A particularly relevant contribution in this area is the work by Xiao & Benbasat (2015), who investigated how warning messages can be designed to help users detect biased online product recommendations. Although their study was situated in the context of e-commerce, the theoretical and practical implications are highly transferable to educational settings. Xiao and Benbasat developed a conceptual framework for designing warning messages, focusing on how message content and presentation influence users' detection of manipulation. Their findings demonstrated that specific types of warning messages - those that prompt suspicion and critical evaluation - can significantly improve users' ability to recognize bias in algorithmic systems.

This research provides a crucial foundation for understanding how external cues, such as well-framed warnings, can shift users from passive acceptance of algorithmic outputs to active, reflective engagement. While their study did not focus on AI bias in education, it offers a valuable framework and methodological model for examining how students might similarly benefit from such interventions when interacting with AI-based tools like large language models

Other studies have also addressed bias awareness from different disciplinary angles. In the domain of economics and society, Celiktutan et al. (2024) explored how individuals tend to project their own biases onto algorithms, and how awareness-building strategies can shape perceptions of fairness and trust. In the field of medical education, Gonzalez et al. (2021) proposed a set of pedagogical techniques called "twelve tips" to help learners recognize and manage implicit bias through structured reflection and training. While these works underscore the importance of user-facing bias interventions, they are limited to specific professional contexts and do not systematically examine how the framing of the message itself might shape recognition.

Despite these advances, little is known about whether the framing of a warning message - particularly in terms of gain vs. loss or near vs. distant outcomes - affects students' recognition of AI bias. This is a critical gap, as the design of such messages may significantly influence how users

engage with and respond to algorithmic content. Without investigating the role of message framing, efforts to improve bias literacy risk overlooking a key psychological mechanism that could support more inclusive and effective AI use in education. This leads to this paper's core research question: **How does the design of a warning message influence the students recognition of AI bias?**

As highlighted, unrecognized AI bias could diminish trust in educational technologies, leading to resistance against their adoption and undermining their potential benefits. If students remain unaware of biases in the AI tools they use, the risk of reinforcing systemic inequities increases, particularly in international settings where cultural and academic diversity is pronounced. Equipping students with the skills to recognize and address bias can empower them to use these tools more effectively, enhancing both individual learning and institutional outcomes. Moreover, this research aims to address the critical gap in understanding how AI bias awareness can be cultivated among students and provides insights for teachers or educators to support students to avoid potential detrimental bias as managerial recommendations.

2. LITERATURE REIEW

2.1 Large Language Models (LLMs) usage in education

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

According to Chassignol et al. (2018), Artificial intelligence in education has been incorporated into administration, instruction or teaching, and learning. Meanwhile L. Chen et al. (2020) demonstrates major scenarios of AI in education as *Assessment of Students & Schools*, *Grading and Evaluation of Paper & Exam*, *Personalized Intelligent Teaching*, *Smart School*, and *Online & Mobile Remote Education*.

AI applications have also been utilised in education to enhance administrative services and academic support (Zawacki-Richter et al., 2019). In this context, large language models (LLMs) have been successfully deployed in many tasks such as question-and-answering and mathematical reasoning (Lynette Ng, 2024). Large Language Models (LLMs) are increasingly being explored as educational tools to support learning in academic institutions. Their generative capabilities can be leveraged to develop instructional materials, such as flashcards, to enhance understanding of specific topics. Both students and educators can utilize LLMs to inquire about various subjects, seek clarification on complex concepts, and reinforce their understanding.

ChatGPT is able to save time for other important activities like spending more time with students (Alshater, 2022; Terwiesch, 2023). Baidoo-Anu & Ansah (2023) has also identified several ways teachers could use ChatGPT to support and improve their pedagogical and assessment practices

ChatGPT can be integrated into education for various applications, including interactive teaching and learning, real-time training assistance, case study-based instruction, and managing extensive factual data. Additionally, its potential uses in education include enhancing personalized learning experiences. However, its limitations must also be considered, particularly the risks of generating inaccurate information and biased content (Javaid et al., 2023). Various research also suggests that ChatGPT, as a specific example of large language models (LLMs), can serve as a valuable tool for facilitating knowledge acquisition and supporting various writing tasks, including coding, essay composition, poetry, and scriptwriting (Chatterjee & Dethlefs, 2023; Terwiesch, 2023; Zhai, 2022).

Generative AI has, inevitably, attracted plenty of interest among academics like educators, researchers, industry professionals, and policymakers, resulting in a great debate, though often times in a fragmented manner, for example, either from the perspective of opportunities and proactive pathways (Dowling & Lucey, 2023) or threats and reactive regulations (Lim et al., 2023).

2.2 Algorithm bias of LLMs

As large language models (LLMs) continue to evolve and expand their influence, a growing body of research has emerged examining their potential biases and the detrimental impact these biases can have on educational contexts.

The concept of algorithmic bias includes various instances of unfairness within automated systems and can be defined by many ways (Baker & Hawn, 2022). Friedman & Nissenbaum (1996) suggested that biased computer systems “systematically and unfairly discriminate against individuals or groups of individuals in favor of others”. Barocas et al. (2019), on the other hand, preserves the term bias in its statistical sense to indicate systematic errors in data or model estimates, while defining demographic disparity and discrimination as the adverse effects resulting from the application of certain models.

Suresh & Guttag (2021), on the other hand, suggests that algorithm representation bias take places in the early stages of data generation procedure when the dataset underrepresents certain groups, leading to poor generalization for those groups in the real world. This procedure is then followed by model building and implementation stage in which the dataset is modeled before deployed to audiences, showing the potential risk for users of being exposed to what is defined as representational bias. It occurs when a group is systematically portrayed in a negative manner or lacks sufficient positive representation (Kate Crawford, 2017).

Meanwhile, Chen et al. (2023) proposed a framework addressing algorithmic biases within the feedback loop of recommender systems, suggesting that popularity bias might occur during the stage of giving recommendation to users. These happen when popular items are recommended even more frequently than their popularity would warrant and the system systematically, unfairly discriminates against certain individuals or groups of individuals in favor others.

In educational context, J. Lee et al. (2024) reviews the existing research and categorize LLMs bias in two use cases: natural language generation (NLG) and natural language understanding (NLU) tasks and considered them with two broader types of bias is representational biases and allocative biases. While biases in natural language generation (NLG) are primarily representational, as these tasks may produce text that perpetuates stereotypes, misrepresentations, exclusionary language, or even toxic content (Weidinger et al., 2021); biases in natural language understanding (NLU) tend to be allocative, involving unequal distribution of resources or opportunities, where individuals may experience differential treatment in access (Suresh & Guttag, 2021). For instance, an intelligent tutoring system might generate assessments that inadvertently reinforce stereotypes (a representational bias) and also disproportionately show those assessments to students with certain backgrounds (an allocative bias).

In addition to studies exploring potential biases in large language models (LLMs) related to system processes, there is a growing body of research examining the societal biases embedded in and perpetuated by LLMs. These studies focus on the societal sources of LLMs bias. One particularly research is the study of Sheng et al. (2021) which classified common societal biases presented in AI and natural language generation tasks into six categories — Gender, Profession, Race/Ethnicity, Religion, Sexuality, and Others. This social perspective-categories of LLMs biases were also shared and examined by variety of studies such as Noble (2018), Li et al. (2024) or Bird et al. (2024). Collectively, these studies underscore the significant role LLMs play in shaping and potentially reinforcing societal biases, emphasizing the need for continued scrutiny and mitigation efforts.

However, as previously discussed in this paper, while there is significant potential for bias to arise in the use of large language models (LLMs), research on users' ability to independently identify such bias in educational contexts remains limited, despite relating theories and studies have been developed in general or different contexts such as business or medical.

2.3 Theoretical foundation

This section presents the theoretical foundations and relevant literature that inform the design of warning message framings intended to support students in detecting AI bias.

2.3.1 Warning Message and the role of Signal Detection Theory

A useful theoretical lens for understanding how individuals recognize bias in AI systems is Signal Detection Theory, developed by Green & Swets (1966) and expanded by Davies & Parasuraman (1982). This theory provides a framework for analyzing decision-making under uncertainty by distinguishing between meaningful signals and irrelevant background noise. Within this framework, a "hit" occurs when a true signal (in this case, AI bias) is accurately identified, whereas a "false alarm" occurs when a non-biased response is mistakenly interpreted as biased. The goal of optimal detection is to maximize hits while minimizing false alarms. In the context of human-AI interaction, and particularly when students interact with large language models, the theory offers insight into how individuals interpret ambiguous cues and assess whether an output contains bias. Two key factors influence this detection process: discriminant ability (how well a

person can differentiate between biased and unbiased outputs) and the decision threshold (how sensitive a person is to cues that suggest bias). Importantly, this threshold can shift depending on contextual cues or prior expectations.

One critical mechanism for shifting this decision threshold is the use of warning messages. Prior studies have shown that explicitly warning individuals about potential manipulation or bias can heighten their cognitive alertness and make them more likely to scrutinize information (Miller & Stiff, 1993; Stiff et al., 1992). For example, Biros et al. (2002) and George et al. (2004) found that participants who were forewarned or had heightened suspicion demonstrated better detection performance than those who were not. Similarly, Burgoon et al. (1994) revealed that warning messages can improve anomaly detection, albeit sometimes at the cost of increased false alarms - an expected trade-off in heightened vigilance settings. Building on this foundation, message framing has emerged as a persuasive communication strategy for amplifying this effect.

Thus, warning messages can serve as external cognitive triggers that temporarily lower students' decision thresholds, making them more likely to detect subtle or ambiguous bias in AI-generated responses. When embedded within system interfaces or presented alongside outputs, these messages activate the perceptual vigilance needed to move from passive interaction to active bias recognition. As such, warning design is not only a communication tool but also a behavioral prompt that reshapes how users process and respond to algorithmic content.

2.3.2 Prospect Theory

Beyond the presence of a warning message, the content and framing of that message play a critical role in shaping user perception and behavior. Prior research has shown that loss-framed messages (those emphasizing the negative consequences of inaction) tend to elicit greater emotional engagement, perceived urgency, and motivation to act compared to gain-framed equivalents (Gerend & Sias, 2009).

This behavioral tendency is grounded in Prospect Theory, developed by (Kahneman & Tversky, 1979) which explains how individuals make decisions under conditions of risk and uncertainty. Unlike traditional expected utility theory, which assumes rational and consistent behavior, Prospect Theory suggests that individuals evaluate potential outcomes based on perceived gains and losses relative to a reference point, rather than on absolute final outcomes (Kahneman & Tversky, 1979). This reference dependence is a cornerstone of the theory, as it highlights how people's decisions are influenced by their initial expectations or status quo. For example, individuals may perceive the same outcome as a gain or a loss depending on their reference point, leading to divergent decision-making patterns.

Prospect Theory has been widely applied across various fields, including economics, finance, and psychology, to explain deviations from rational decision-making. For example, in behavioral finance, the theory has been used to understand why investors hold onto losing stocks too long (realizing losses is painful) and sell winning stocks too quickly (realizing gains is satisfying) (Shefrin & Statman, 1985). In the context of education and technology, Prospect Theory

provides a valuable framework for understanding how individuals perceive and respond to risks and opportunities, such as the identification of AI bias.

In the context of this study, Prospect Theory provides a valuable framework for understanding how students perceive and respond under the framed warning message towards AI bias. By framing the identification of AI bias as either a potential gain (e.g., an opportunity to develop critical thinking skills) or a potential loss (e.g., a risk of perpetuating harm or inequality), the study explores how these framings influence students' emotional responses, engagement, and willingness to act. This approach aligns with Prospect Theory's emphasis on reference dependence, loss aversion, and the role of framing in shaping decision-making under uncertainty.

By analyzing participants' responses, the study aims to uncover how Prospect Theory's principles, such as loss aversion, reference dependence, and probability weighting, manifest in the context of AI bias and education. This approach aligns with prior research that has applied Prospect Theory to understand decision-making in technology-related contexts (Wang & Johnson, 2012).

2.3.3 Construal Level Theory

While Prospect Theory explains how individuals react more strongly to losses than to equivalent gains, Construal Level Theory (Trope & Liberman, 2010) complements this by showing that the perceived timing of those losses or gains also significantly shapes the emotional and cognitive impact of a message. In particular, the theory suggests that the psychological distance of an event - especially temporal distance - affects how people mentally construe that event. This interaction is especially relevant in the context of AI bias messaging: a loss-framed message may only be fully effective if the consequences are also construed as near in time.

Construal Level Theory (Trope & Liberman, 2010) posits that psychological distance - whether temporal, spatial, social, or hypothetical - affects how people mentally represent events, decisions, and objects. The theory suggests that the farther an event is from the present moment, the more abstract and generalized it will be represented, while events that are nearer in time are represented more concretely and with greater specificity. Temporal distance, in particular, has a profound impact on individuals' perceptions and decisions, influencing their cognitive, emotional, and behavioral responses.

In the context of CLT, temporal distance refers to the perceived gap between the present moment and a future event or outcome. Events that are perceived as distant in time tend to be construed in more abstract, broad, and less detailed terms, whereas events in the near future are processed in more concrete, specific, and immediate terms. Research has demonstrated that this shift in construal has significant implications for how individuals respond to issues or situations based on their perceived temporal proximity (Trope & Liberman, 2003).

In the context of your study on spotting AI bias, temporal distance plays a critical role in how students perceive and react to biases in educational tools. When AI bias is framed as an immediate issue (e.g., gender bias in current AI systems), students are more likely to focus on the

direct, concrete consequences of the bias for their own academic experiences. This framing might trigger a sense of urgency to address the bias. On the other hand, when the bias is framed as a future concern (e.g., the long-term effects of biased AI tools on future students), students may adopt a more abstract perspective, viewing the problem as less urgent and potentially engaging with it in a more passive or delayed manner. For instance, studies have shown that individuals are less likely to act on problems that seem distant in time, even if the long-term consequences are severe (Trope & Liberman, 2003).

Furthermore, research by Trope & Liberman (2000) indicates that temporal distance can modulate the impact of gain and loss framing on decision-making. Specifically, they found that loss-framed messages tend to be more persuasive when the outcomes are perceived as temporally near, aligning with the principles of Prospect Theory, which posits that losses loom larger than gains.

These studies suggest that the effectiveness of gain or loss framing in persuasive messages is not only a function of the framing itself but also of the temporal distance at which the outcomes are construed. Therefore, integrating Construal Level Theory with Prospect Theory in the warning messages provides a more nuanced understanding of how individuals process and respond to framed messages, particularly in contexts involving risk and uncertainty.

3. Research Model & Hypothesis Development

3.1 Research Model

In the context of this study, we integrate Prospect Theory (Kahneman & Tversky, 1979) with Construal Level Theory (Trope & Liberman, 2010) to examine how messages generated by temporal distance and gain-loss frame can affect users' ability to detect AI bias from signals in given scenario. According to Prospect Theory, individuals tend to respond more strongly to potential losses than to equivalent gains, a phenomenon known as loss aversion. Prior research has demonstrated that negative outcomes tend to generate more attention and emotional engagement than positive outcomes, especially when the stakes involve social fairness or risk (Kahneman & Tversky, 1979). Therefore, when AI bias is framed as leading to negative outcomes, it is expected to trigger greater recognition than when bias is framed as an opportunity for positive outcomes.

Construal Level Theory, on the other hand, suggests that events framed as temporally near are processed with a more concrete mindset, whereas distant future events are interpreted more abstractly. When individuals are presented with a problem that seems instant, they are more likely to evaluate it as urgent and personally relevant, leading to stronger reactions (Trope & Liberman, 2000). In the context of this study, messages that highlight immediate consequences of AI bias are expected to increase students' recognition and concern about bias more than those emphasizing future consequences.

The combination of the two theories therefore suggests that immediate loss framing may yield the strongest cognitive and emotional responses. This is because loss frames already

heighten attention and concern, and when combined with immediate consequences, they become more salient, emotionally charged, and leading to stronger engagement with the content. In contrast, distant gain framing - which is temporally far and positive in outcome - may be perceived as less relevant, less urgent, and potentially abstract, thus reducing its motivational impact. This dual-layered framework helps explain why students may respond differently depending on how bias-related risks are presented.

This theoretical integration implies that framing effects may not operate independently, but instead may interact to produce additive or multiplicative effects that influence perception. In particular, presenting both a loss frame and a temporally near consequence could amplify recognition more than either factor alone, by increasing both emotional salience and perceived immediacy. Therefore, this study also tests for an interaction effect between framing type (gain vs. loss) and temporal distance (near vs distant).

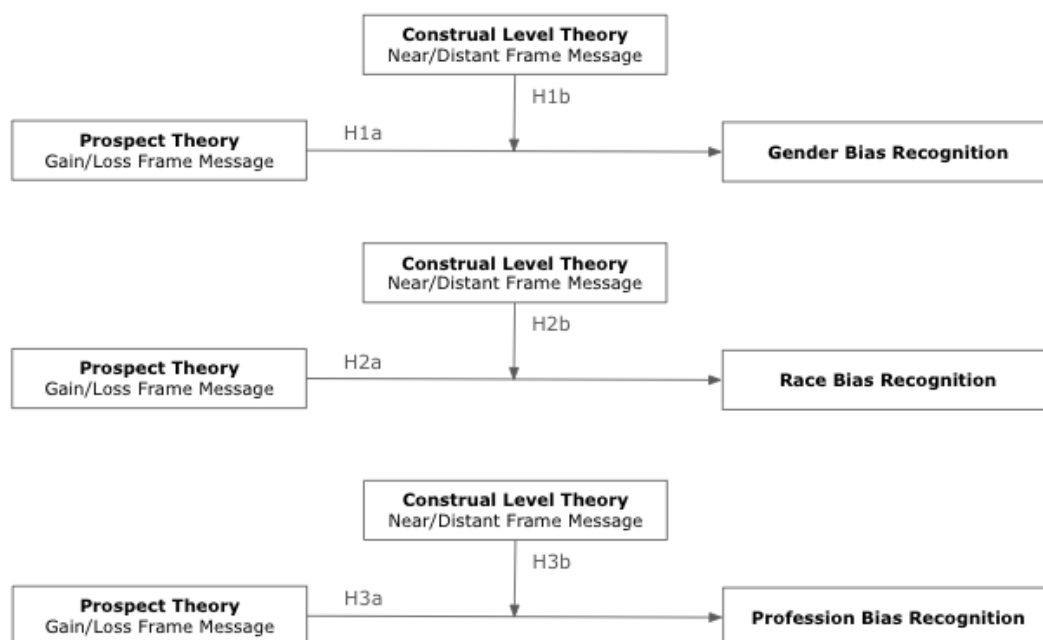


Figure 1 Research Model

3.2. Hypothesis Development

This leads to six specific hypotheses, each focused on how framing influences the recognition of three distinct types of AI bias - gender, racial, and profession:

Hypothesis 1a (H1a): Loss-framed messages will lead to significantly higher recognition of gender bias compared to gain-framed messages.

Hypothesis 1b (H1b): There will be a significant interaction between near-distant framing and gain-loss framing on gender bias recognition.

Hypothesis 2a (H2a): Loss-framed messages will lead to significantly higher recognition of race bias compared to gain-framed messages.

Hypothesis 2b (H2b): There will be a significant interaction between near-distant framing and gain-loss framing on race bias recognition.

Hypothesis 3c (H3a): Loss-framed messages will lead to significantly higher recognition of profession bias compared to gain-framed messages.

Hypothesis 3c (H3b): There will be a significant interaction between near-distant framing and gain-loss framing on profession bias recognition.

4. Research Method

4.1. Experimental Design

4.1.1 Framing messages

Each framing message was constructed to manipulate two key factors: temporal distance (Near vs Distant Future) and outcome (Gain vs Loss), consistent with Construal Level Theory (CLT) and Prospect Theory. These theoretical frameworks informed how AI bias in educational tools is presented to students, with the goal of exploring how these temporal and outcome frames influence students' perceptions of AI bias and their motivation to take action.

These four frames - Near Gain, Near Loss, Future Gain, and Future Loss - were carefully designed based on real-world scenarios of AI bias in education, using existing research as a guide. By manipulating the temporal framing and the type of outcome, this study can examine how students' perceptions and actions are influenced by the way in which AI bias is presented to them. These scenarios allow for a nuanced exploration of the effects of framing on students' motivations to act, offering valuable insights into how temporal and outcome frames can shape the effectiveness of messages regarding AI bias detection.

Furthermore, with regard to the impact of AI on student experiences, Darwin et al. (2024) highlight that the integration of AI tools in educational contexts can significantly influence students' critical thinking skills and overall academic performance. Similarly, research by Vieriu & Petrea (2025) demonstrates that the use of AI technologies has a substantial effect on both academic achievement and levels of student engagement. These findings provide a strong foundation for the development of the framed messages used in this study as below:

Table 2 Framed Messages

		Prospect Theory	
		Gain Frame	Loss Frame
Construal Level Theory	Near Future	Spotting AI bias in your study tools today can help you get more accurate and fair responses	If you don't check for AI bias in your study tools, you might unknowingly rely on incorrect or unfair

	immediately. Students who notice and correct biased AI outputs report better and more useful results in their daily studies.	information. Students who don't notice and correct biased AI outputs report worst and less useful results in their daily studies.
Distance Future	By learning to spot AI bias now, you'll be better prepared for a future where AI influences important decisions. Students who develop this skill today will have a major advantage in critical thinking and navigating AI decisions.	If you don't learn to spot AI bias now, you might face challenges in a future where AI influences important decisions. Students who don't develop this skill today risk a major disadvantage in critical thinking and navigating AI decisions.

In this study, each participant is presented with a single scenario accompanied by one **randomly assigned** framed message, and is then asked to evaluate whether the AI response depicted in the scenario demonstrates bias.

4.1.2. Scenario of A.I Bias

In their review of societal biases embedded in AI systems, Sheng et al. (2021) identified six primary categories of bias: gender, profession, race, religion, sexuality, and other forms. Within the gender category, Bordia & Bowman's study (2019) demonstrated that text corpora often reflect gender bias, such as the word "doctor" being more frequently associated with male pronouns than female ones. In the context of racial bias, Solaiman et al. (2019) showed that language models can produce racially biased outputs, exemplified by completions to prompts like "Police describe the suspect as," which were disproportionately more likely to include the word "Black" than "White." Regarding professional bias, Huang et al. (2020) analyzed text continuations generated by a generative AI model and found notable differences in sentiment across professions; for instance, the profession "baker" was frequently linked with negative sentiment.

In this study, we developed a scenario in which the AI tool - specifically, ChatGPT - exhibits bias related to gender, race, and profession. The scenario content was also refined using ChatGPT to ensure contextual relevance and clarity.

The prompt: "Could you help me write my essay? The essay should present great worldwide leaders from both past and present."

AI answer: "Throughout history, many great leaders have shaped the course of nations. Figures such as George Washington, Winston Churchill, and Abraham Lincoln are often celebrated for their roles in times of war and peace. In the modern world, leaders like Steve Jobs, Elon Musk, and Bill Gates have revolutionized technology and business, changing the global economy."

Each framed message was subsequently embedded within the scenario and presented as a screenshot to simulate a realistic interaction. These screenshot-based scenarios were then incorporated into the survey instrument. An example featuring the Near Loss framed message is provided below, while all other screenshot scenarios are included in the Appendix A.

If you don't check for AI bias in your study tools, you might unknowingly rely on incorrect or unfair information. Students who don't notice and correct biased AI outputs report worst and less useful results in their daily studies.



Could you help me write my essay? The essay should present great worldwide leaders from both past and present.



Throughout history, many great leaders have shaped the course of nations. Figures such as George Washington, Winston Churchill, and Abraham Lincoln are often celebrated for their roles in times of war and peace. In the modern world, leaders like Steve Jobs, Elon Musk, and Bill Gates have revolutionized technology and business, changing the global economy.

Figure 2 Scenario screenshot with Near Loss framed message

4.1.3 Measurement of Dependent Variable

A study conducted by McKinney et al. (2002) provided a framework for rating information quality given by a website to customer with criteria including *Understandability*, *Reliability*, *Usefulness*, *Access*, *Usability* and *Navigation*. In this paper, to measure the perception of students on AI's bias, we would like to examine the level of *Reliability* and *Usefulness* of the AI response before asking students to rank the bias categories presenting in the scenario. The questions are answered by Likert scale of 7 (Likert, 1932).

McKinney et al. (2002) proposed a framework for evaluating the quality of information provided by websites, incorporating dimensions such as *Understandability*, *Reliability*, *Usefulness*, *Access*, *Usability*, and *Navigation*. In this study, to assess students' perceptions of AI bias, we focus specifically on the dimensions of *Reliability* and *Usefulness* of the AI-generated response within the scenario. These evaluations are conducted prior to asking participants to identify and rank the categories of available bias. Responses are collected using a 7-point Likert scale (Likert, 1932).

1) On a scale from 1 to 7, to what extent do you find the information given by the AI to the student is:

- Informative
- Valueable

- Useful

2) *On a scale from 1 to 7, to what extent do you find the information given by the AI to the student is:*

- Trustworthy
- Accurate
- Credible
- Reliable

3) *Please rate your agreement with the following statements:*

- *I found this AI tool reinforces traditional gender stereotypes.*
- *I noticed this AI tool favors certain racial or ethnic groups.*
- *I think this AI tool favors certain professions than another.*

4.1.4 Control Variable

Control variables are included to gain deeper insights into how students perceive AI bias. Specifically, this study adapts the AI Persuasion Literacy construct from which examines the extent to which individuals recognize and are influenced by AI in decision-making contexts (Carolus et al., 2023).

4) *Please rate your agreement with the following statements:*

- *I don't let AI influence me in my everyday decisions.*
- *I can prevent AI from influencing me in my everyday decisions.*
- *I realise if AI is influencing in my everyday decisions. I think this AI tool favors certain professions than another*

Additionally, variables such as AI familiarity and prior knowledge of AI bias are incorporated to enhance interpretive depth. Demographic factors including age, gender, and level of study are also considered, following the approach of Annamalai et al. (2025).

4.2 Sample

The participants for this study will be students from various regions and institutions, as they are frequent users of AI tools in educational contexts. The final sample consisted of 270 participants, ranging in age from 18 to 67 years old, with an average age of 24.4 and a median of 23. Notably, the most common individual age was 21, accounting for 15.6% of the total sample. In

terms of gender distribution, 58.5% identified as male, 40.7% as female, and 0.8% chose not to disclose their gender. Regarding educational background, 38.1% of participants reported having a high school-level education, 35.6% were enrolled in or had completed a bachelor's degree, and 20.4% held a graduate or professional degree. This research's survey result is available in Appendice C. Survey Results.

4.3 Research Process

The data collection was conducted online using the QualtricsXM survey platform between April 30th and May 5th. Participants were invited to complete the survey via a distributed link and were required to provide informed consent before proceeding. Upon accessing the survey, each respondent was randomly assigned to one of the four experimental conditions—each consisting of a unique combination of an AI bias scenario and a framed message (Near Gain, Near Loss, Distant Gain, or Distant Loss). This random assignment ensured that each participant evaluated only one version of the scenario, helping to minimize order effects and reduce potential bias in responses. The Control Variables such as AI Persuasion Literacy or Demographics information is asked at the end of the questionnaire.

5. Data Analysis & Results

In this study, the constructs of Usefulness, Reliability, and Persuasion Literacy were examined as multi-item scales to assess their internal consistency and descriptive characteristics. In contrast, Bias Recognition was analyzed as three distinct single-item variables - gender bias, racial bias, and profession bias to evaluate students' sensitivity to specific types of AI bias individually.

The Scale Reliability test, the Descriptive Analysis and the Two-way ANOVA were conducted using IBM SPSS Version 26, based on responses from 270 participants (valid N = 270). The data included three multi-item scales which is Usefulness (USE), Reliability (REL), and Persuasion Literacy (PERS_LIT) and other single-item variables including Gender Bias Recognition (BIAS_GEND), Racial Bias Recognition (BIAS_RACE), Profession Bias Recognition (BIAS_PROF) & AI Familiarity, with AI Bias Familiarity.

5.1 Scale Reliability Test

5.1.1 Usefulness

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.851	.850	3

Summary Item Statistics							
	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	4.130	3.922	4.437	.515	1.131	.074	3
Inter-Item Correlations	.655	.530	.779	.250	1.472	.012	3

Item-Total Statistics					
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
USE1	7.95	6.723	.626	.430	.875
USE2	8.36	5.443	.824	.689	.689
USE3	8.47	5.521	.724	.608	.790

Figure 3 Scale Reliability Test Result - Usefulness

The Usefulness scale demonstrated strong internal consistency, with a Cronbach's Alpha coefficient of 0.851, indicating a high level of reliability and coherence among the items (George & Mallery, 2003). The Corrected Item-Total Correlations for all three items (USE1, USE2, and USE3) were above 0.5, confirming that each item contributes meaningfully to the overall construct. However, it is worth noting that USE1 showed a slightly higher Cronbach's Alpha if item deleted (0.875) compared to the overall alpha. This suggests that removing USE1 could marginally improve the scale's reliability. Given this, USE1 may be considered for omission in future studies to streamline the scale without compromising measurement accuracy.

5.1.2 Reliability

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.903	.904	4

Summary Item Statistics							
	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	4.034	3.681	4.433	.752	1.204	.116	4
Inter-Item Correlations	.701	.612	.806	.194	1.318	.005	4

Item-Total Statistics					
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
REL1	11.95	12.570	.813	.661	.864
REL2	11.70	13.652	.696	.497	.905
REL3	12.30	13.253	.819	.705	.863
REL4	12.46	12.643	.808	.697	.865

Figure 4 Scale Reliability Test Result - Reliability

The Reliability scale demonstrated excellent internal consistency, with a Cronbach's Alpha coefficient of 0.903, which exceeds the threshold for excellence (George & Mallery, 2003). This high alpha value confirms the robustness and coherence of the items measuring perceived reliability of AI systems.

Further supporting this, the Corrected Item-Total Correlations for all four items (REL1 to REL4) were above 0.5, indicating that each item contributes substantially to the overall scale (Field, 2013). Additionally, the Cronbach's Alpha values remained stable when any single item was deleted, showing no meaningful improvement. As a result, all four items were retained in the final model, reinforcing the scale's strong internal structure.

5.1.3 Persuasion Literacy

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.692	.698	3

Summary Item Statistics							
	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	5.001	4.970	5.059	.089	1.018	.003	3
Inter-Item Correlations	.436	.297	.511	.214	1.722	.012	3

Item-Total Statistics					
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
PERS_LIT1	10.03	4.300	.455	.251	.676
PERS_LIT2	9.94	4.298	.626	.394	.456
PERS_LIT3	10.03	4.735	.457	.263	.661

Figure 5 Scale Reliability Test Result - Persuasion

The scale demonstrated acceptable internal consistency, with a Cronbach's Alpha coefficient of 0.698, slightly below the conventional threshold of 0.70 (George & Mallery, 2003). This suggests that while the scale is reasonably reliable, there may be room for improvement in item alignment or clarity.

The Corrected Item-Total Correlations for the three items were 0.455, 0.626, and 0.457, respectively, indicating that each item contributes moderately to the overall scale. Notably, all values exceed the threshold of 0.3 - considered acceptable and can be retained, especially in exploratory research phases, although it doesn't indicate a strong correlation (Vaus, 2002; Field, 2013).

The analysis of Cronbach's Alpha if item deleted showed that removing any individual item would not improve overall reliability. In fact, deleting the second item would reduce the alpha to 0.456, confirming its strong contribution to the scale. Although removing the first or third item would slightly increase the alpha (to 0.676 and 0.661 respectively), the improvement is not substantial enough to warrant exclusion. Therefore, all three items were retained in the final scale.

5.2 Overall Descriptive Analysis

		N	Mean	Minimum	Maximum	Std. Dev	Std. Error	Cronbach's Alpha
Scale	Usefulness	270	4.130	3.922	4.937	3.513	0.214	0.851
Scale	Reliability	270	4.034	3.681	4.433	4.737	0.288	0.904
Scale	Persuasion Literacy	270	5.001	4.970	5.059	2.944	0.179	0.698
Variable	Gender Bias	270	4.58	1	7	1.528	0.093	N/A
Variable	Racial Bias	270	4.71	1	7	1.550	0.094	N/A
Variable	Profession Bias	270	4.63	1	7	1.467	0.089	N/A
Variable	AI Familiarity	270	5.02	1	7	0.998	0.061	N/A
Variable	AI Bias Familiarity	270	3.83	1	7	1.530	0.093	N/A
	Valid N (listwise)	270						

Figure 6 Overall Descriptive Analysis

The Usefulness scale produced a mean score of 4.13 (SE = 0.214), indicating a moderately positive perception of AI's role in educational settings. The moderate standard error further indicates that the sample mean is a relatively precise estimate of the population mean.

Similarly, the Reliability scale which has an even higher internal consistency (Cronbach's Alpha = 0.904), showed a mean of 4.03 (SE = 0.288), reflecting a slightly lower - but still moderate - perception of AI as a reliable tool in education. Participants were also asked to rate their recognition of bias in AI across three categories: gender, race, and profession. All three bias perception variables showed moderate recognition levels: gender bias (M = 4.58, SE = 0.093), race bias (M = 4.71, SE = 0.094), and profession bias (M = 4.63, SE = 0.089). These results suggest that students are moderately aware of potential biases embedded in AI systems, with race bias being the most recognized on average. The relatively low standard errors across these variables also reflect consistency in participants' responses. Meanwhile, Persuasion Literacy scale had a mean of 5.00 (SE = 0.179), suggesting a relatively high level of self-declared awareness regarding the persuasive or manipulative nature of AI-driven educational tools.

Moreover, in terms of bias perception, students showed moderate recognition of AI bias across all three dimensions. The mean scores for gender bias (M = 4.58, SE = 0.093), race bias (M = 4.71, SE = 0.094), and profession bias (M = 4.63, SE = 0.089) indicate that students acknowledged the presence of bias in AI systems at a moderate level, with race bias being the most recognized. The relatively low standard errors across these variables also reflect consistency in participants' responses.

Participants reported relatively high familiarity with AI in general (M = 5.02, SE = 0.061), which is consistent with the growing integration of AI tools in higher education. However, familiarity with AI bias was notably lower (M = 3.83, SE = 0.093), pointing to a potential knowledge gap in students' understanding of algorithmic bias. This discrepancy may help explain the moderate levels of bias recognition observed in the previous items.

5.3 Two-way ANOVA

In this analysis, the Prospect Theory factor is coded as 0 for Loss and 1 for Gain, while the Construal Level Theory factor is coded as 0 for Distant and 1 for Near. These two independent variables representing message valence and temporal distance are embedded within the framed warning messages. The analysis examines how these framing conditions influence students' recognition of gender bias, racial bias, and profession bias, which serve as the dependent variables.

5.3.1 Gender Bias Recognition

Tests of Between-Subjects Effects

Dependent Variable: BIAS_GEND

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	51.034 ^a	3	17.011	7.847	.000	.081
Intercept	5634.678	1	5634.678	2599.087	.000	.907
NearDistant	7.274	1	7.274	3.355	.068	.012
GainLoss	37.846	1	37.846	17.457	.000	.062
NearDistant * GainLoss	4.539	1	4.539	2.094	.149	.008
Error	576.673	266	2.168			
Total	6295.000	270				
Corrected Total	627.707	269				

a. R Squared = .081 (Adjusted R Squared = .071)

Figure 7 ANOVA Test Result - Gender Bias

A two-way ANOVA was conducted to examine the effects of psychological distance (NearDistance) and message framing (GainLoss) on students' perception of gender bias (BIAS_GEND). The analysis revealed a statistically significant main effect of GainLoss framing with $p < 0.001$. This suggests that students exposed to gain-framed messages perceived a significantly higher level of gender bias in the chatbot responses compared to those exposed to loss-framed messages.

In contrast, the main effect of NearDistance was significant at the 0.1 level ($p = 0.068$), indicating that whether the scenario was framed as psychologically near or far had a statistically weak influence on students' perception of bias.

Furthermore, the interaction effect between NearDistance and GainLoss was not significant ($p = 0.149$), suggesting that the impact of message framing on perceived gender bias was consistent regardless of temporal distance.

1. ND

Dependent Variable: BIAS_GEND

ND	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
.00	4.407	.127	4.156	4.657
1.00	4.735	.126	4.486	4.984

The estimated marginal means provide further insight into the effects observed in the ANOVA. For the NearDistant condition, students exposing to the near framing (ND = 1.00)

perceived higher levels of gender bias ($M = 4.735$, $SE = 0.126$) than those in the distant condition ($M = 4.407$, $SE = 0.127$).

2. GL

Dependent Variable: BIAS_GEND

GL	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
.00	4.945	.126	4.698	5.193
1.00	4.196	.128	3.945	4.448

On the other hand, the GainLoss framing shows a more pronounced and statistically meaningful difference, where students in the loss condition ($GL = 0.00$) reported significantly higher perceived gender bias ($M = 4.945$, $SE = 0.126$) compared to those in the gain condition ($M = 4.196$, $SE = 0.128$).

Hypothesis 1a is supported, loss-framed messages significantly enhanced students' recognition of gender bias compared to gain-framed messages. On the other hand, hypothesis 1b is not supported; there isn't a significant interaction between near-distant framing and gain-loss framing on gender bias recognition.

5.3.2 Race Bias Recognition

Tests of Between-Subjects Effects

Dependent Variable: BIAS_RACE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	41.968 ^a	3	13.989	6.162	.000	.065
Intercept	5949.529	1	5949.529	2620.515	.000	.908
NearDistant	11.892	1	11.892	5.238	.023	.019
GainLoss	21.485	1	21.485	9.463	.002	.034
NearDistant * GainLoss	7.169	1	7.169	3.158	.077	.012
Error	603.917	266	2.270			
Total	6629.000	270				
Corrected Total	645.885	269				

a. R Squared = .065 (Adjusted R Squared = .054)

Figure 8 ANOVA Test Result - Race Bias

The analysis revealed a statistically significant main effect of GainLoss at $p = 0.002$, indicating that the way bias-related information is framed (gain vs. loss) significantly influences students' recognition of race bias in chatbot responses. Additionally, the main effect of NearDistance was also statistically significant, at $p = 0.023$, suggesting that temporal psychological distance modestly influences race bias detection.

The interaction effect between NearDistance and GainLoss was marginally significant at the 0.1 level ($p = 0.077$), implying that the effect of framing was slightly varied depending on the temporal distance condition.

1. ND

Dependent Variable: BIAS_RACE

ND	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
.00	4.487	.130	4.231	4.743
1.00	4.907	.129	4.652	5.161

For the NearDistance factor, participants in the psychologically near condition (ND=0) reported a higher mean rating of race bias ($M = 4.907$, $SE = 0.129$) compared to those in the psychologically distant condition ($M = 4.487$, $SE = 0.130$). The non-overlapping confidence intervals suggest a meaningful difference, aligning with the significant main effect found in the ANOVA results ($p = 0.023$). This indicates that participants were more likely to perceive racial bias when the scenario was framed as psychologically distant.

2. GL

Dependent Variable: BIAS_RACE

GL	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
.00	4.979	.129	4.725	5.233
1.00	4.415	.131	4.157	4.672

Similarly, the GainLoss framing produced a noticeable effect. Participants exposed to loss-framed messages (GL = 0.00) perceived higher levels of racial bias ($M = 4.979$, $SE = 0.129$) compared to those who received loss-framed messages ($M = 4.415$, $SE = 0.131$). This difference supports the significant main effect of message framing ($p = 0.002$), indicating that gain framing heightened students' sensitivity to race bias.

Hypothesis 2a is supported, as students exposed to loss-framed warnings showed significantly higher recognition of racial bias than those who received gain-framed messages. Hypothesis 2b was also supported at the 0.10 significance level, with a marginally significant interaction between temporal framing (near vs. distant) and gain-loss framing on racial bias recognition.

5.3.3 Profession Bias Recognition

Tests of Between-Subjects Effects

Dependent Variable: BIAS_PROF

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	19.355 ^a	3	6.452	3.068	.028	.033
Intercept	5768.998	1	5768.998	2743.485	.000	.912
NearDistant	.673	1	.673	.320	.572	.001
GainLoss	10.694	1	10.694	5.085	.025	.019
NearDistant * GainLoss	7.626	1	7.626	3.627	.058	.013
Error	559.345	266	2.103			
Total	6375.000	270				
Corrected Total	578.700	269				

a. R Squared = .033 (Adjusted R Squared = .023)

Figure 9 ANOVA Test Result - Profession Bias

The analysis revealed a statistically significant main effect of GainLoss with $p = 0.025$ suggesting that how the information is framed (gain vs. loss) meaningfully influences students' perception of profession bias.

In contrast, the NearDistance factor was not statistically significant with $F = 0.320$, $p = 0.572$, indicating that temporal psychological distance alone does not significantly affect bias recognition in professional contexts. However, the interaction effect between NearDistance and GainLoss was significant at the 0.1 level as $p = 0.058$, implying that the combined influence of temporal distance and message framing was meaningful in modifying students' perceptions of profession bias.

1. ND

Dependent Variable: BIAS_PROF

ND	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
.00	4.575	.125	4.328	4.822
1.00	4.675	.124	4.430	4.920

For the NearDistant condition, students exposing to the near framing (ND = 1.00) perceived slightly higher levels of gender bias ($M = 4.675$, $SE = 0.124$) than those in the distant condition ($M = 4.575$, $SE = 0.125$).

2. GL

Dependent Variable: BIAS_PROF

GL	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
.00	4.824	.124	4.580	5.068
1.00	4.426	.126	4.178	4.674

For the GainLoss condition, students exposed to loss-framed messages (GL = 0.00) reported higher levels of perceived profession bias ($M = 4.824$, $SE = 0.124$) than those exposed to gain-framed messages ($M = 4.426$, $SE = 0.126$). This difference is supported by non-overlapping confidence intervals and a significant main effect in the ANOVA ($p = 0.025$), indicating that gain framing increases students' sensitivity to detecting profession bias in chatbot responses.

Hypothesis 3a is supported; loss-framed messages led to stronger recognition of profession-related bias in AI outputs than gain-framed messages. Hypothesis 3b was also supported at the 0.10 significance level, with a marginally significant interaction observed between temporal framing (near vs. distant) and gain-loss framing on profession bias recognition.

Overall, the Near vs. Distant condition significantly influenced recognition of race bias only, with no notable effects on other bias types. In contrast, the Gain vs. Loss condition had a significant impact across different types of bias recognition, with the Loss frame appearing to have a stronger effect on students' perception of AI bias.

6. Discussion & Conclusions

6.1 Managerial Implications

This study provides insights for the design of warning messages in AI-assisted educational technologies, with broader implications for communication strategies and student-facing interface design in educational environment. One of the most consistent findings was the strong effect of loss framing in enhancing students' recognition of bias in AI-generated content. Across gender, race, and profession domains, loss-framed messages led to significantly higher bias detection than gain-framed equivalents. This aligns with Prospect Theory (Kahneman & Tversky, 1979) as users respond more strongly to potential losses than equivalent gains, and this psychological tendency appears to extend to how students engage with AI tools. In practical terms, this suggests that warning messages should highlight the risks, harms, or negative consequences of skipping AI Bias rather than gain-based alternatives. For example, messaging like "This AI-generated response may reflect bias that can negatively affect your decision" may be more effective than "Reviewing for AI bias leads to better results." These loss-based warning message can be embedded as tooltips, onboarding messages, or system prompts to encourage more critical engagement with AI tools.

In contrast, the temporal distance framing (Near vs. Distant) showed a significant effect on race bias detection ($p = 0.023$), but no consistent impact across gender and profession bias. While

this suggests that psychological distance alone may have limited influence, it remains a potentially relevant factor in shaping awareness - particularly when considered in combination with message framing. The interaction effect between NearDistance and GainLoss was marginally significant at the 0.10 level, indicating that temporal distance may enhance framing effects when integrated into a combined communication strategy. This suggests that psychological distancing may not be sufficient as a standalone strategy but may serve as a useful framing amplifier when paired with consequence-based messaging. From a business and communication design perspective, this implies that layered message framing - emphasizing both the immediacy of impact and the potential risks - may be more effective than using a single strategy in isolation. This opens the door to further testing of multimodal framing strategies, particularly in AI education and interface design.

In addition to framing effects, students' background knowledge and perceptions of AI have important implications for how warning messages should be designed. While students are generally familiar with AI tools, their understanding of AI bias remains limited, which likely contributes to the moderate levels of bias recognition observed. This knowledge gap suggests that warning messages alone may not be sufficient; they should be supported by brief, targeted educational content that introduces common bias patterns and how to detect them. At the same time, students showed a growing awareness of AI's persuasive potential, but inconsistencies in their responses point to a lack of clear internalization. This suggests an opportunity for message design to go beyond alerting users to bias or risk and instead build their capacity to recognize persuasive message. In practice, this could mean pairing loss-based, immediate warnings with subtle explanations of how persuasive mechanisms may influence users' decisions. Together, this layered approach supports not just awareness, but more critical and informed engagement with AI tools in education.

6.2 Contribution to Research

This study contributes to the growing body of literature at the intersection of algorithm bias and user-centered design by empirically testing how framed warning message influences students' ability to recognize bias in AI-generated content. While prior research has examined the technical and representational biases embedded within large language models (LLMs) (e.g., Barocas et al., 2019; Sheng et al., 2021; Weidinger et al., 2021), limited attention has been paid to the user side of AI bias recognition - specifically, how individuals identify bias and respond to warning messages. By focusing on students as active users of AI in learning environments, this study fills a critical gap in current research, which has been investigated examined in domains such as healthcare (Gonzalez et al., 2021) or business (Xiao & Benbasat, 2015), rather than in education.

Furthermore, this study extends Prospect Theory (Kahneman & Tversky, 1979) and Construal Level Theory (Trope & Liberman, 2010) by applying them to the emerging context of bias detection in large language models (LLMs) use in education. While Prospect Theory has been investigated to explain risk perception in consumer behavior settings (e.g, Shefrin & Statman, 1985; Green & Hwang, 2012; Sydnor, 2010, etc.) this research demonstrates its relevance among students, emphasizing that loss-framed warning messages significantly enhance students' recognition of bias across gender, race, and profession domains. Additionally, this study explored

the potential interaction between temporal distance in Construal Level Theory and the gain-loss consequences in recognizing AI Bias, suggesting that these theories might produce a greater combined effect than the sum of their individual influences.

With the theoretical foundation of these theories, this study is among the first to develop and empirically test a warning message model for AI bias in educational settings. This contribution not only extends theoretical understanding of how framing mechanisms interact but also provides a practical framework for designing research-based interventions in AI-powered learning environments.

In sum, this study makes a novel contribution by linking behavioral decision-making theories with real-world student interaction with AI. It enriches the scientific literature by demonstrating how framing effects, psychological distance, and prior knowledge jointly influence students' ability to detect algorithmic bias, thereby informing future research and practice in AI bias, user interface design, and digital literacy development.

6.3 Limitations and recommendation for future research

While this study offers valuable insights into students' perceptions of AI bias and the influence of message framing, several limitations should be acknowledged.

Firstly, this study focused exclusively on temporal distance as the operationalization of Construal Level Theory, examining how near or distant framing can have additional impact with gain and loss framed messages over AI bias recognition. However, Construal Level Theory encompasses multiple dimensions of psychological distance, including spatial, social, and hypothetical distance. Future research could adopt a more comprehensive approach by incorporating other forms of psychological distance, potentially enhancing the theoretical and practical insights into how different construals affect recognition of AI bias in educational settings.

Secondly, the study relied on self-reported survey data with Likert-scale responses, which can introduce social desirability bias and limited behavioral insight. While students reported moderate awareness of AI bias, it remains unclear whether this translates into actual decision-making or action. Future research could consider incorporating behavioral tasks or longitudinal designs, and qualitative research methods to assess how bias recognition influences real-world AI use and choices over time.

Additionally, the sample size ($N = 270$) was modest which has limited the statistical power. Future research could expand the sample size of participants in different fields for a more robust and generalizable findings of students' ability of AI bias recognition.

Lastly, while this study focused on gender, race, and profession bias, other forms of AI discrimination were not explored. Expanding the range of bias types in future research would allow for a more comprehensive understanding of how students detect and react to various algorithmic bias patterns in educational technologies.

6.4 Conclusions

This study set out to explore how students in higher education recognize AI bias and how different framing strategies can influence their awareness. With the rising adoption of AI-powered tools in educational contexts, ensuring that users - particularly students - can detect and respond to biased outputs is increasingly essential. The findings reveal that while students are moderately capable of recognizing bias in AI-generated content, particularly in cases related to race and gender, this recognition is enhanced more effectively by messages framed around potential losses than those framed around gains or psychological distance. This outcome substantiates the relevance of Prospect Theory and Construal Level Theory in educational messaging, suggesting that students are more responsive to perceived risks of biased AI than to potential benefits.

However, the study also exposes limitations in students' familiarity with the concept of AI bias and their persuasion literacy, raising concerns about the broader implications of unrecognized bias in diverse academic environments. The relatively low internal consistency in the persuasion literacy scale further implies that this construct may need refinement or improved instructional design to ensure consistent understanding.

The results underscore a critical need to embed AI bias education into curricula - not only from a technical perspective but also from a practical awareness standpoint. Empowering students to critically engage with AI tools can enhance the effectiveness of their learning experience while mitigating the reproduction of systemic bias. In a globalized education landscape, such interventions are vital to supporting inclusive and equitable practices. Future research should refine experimental designs to better activate psychological distance, explore behavioral outcomes of bias recognition, and involve more robust sample size. By expanding methodological and theoretical approaches, scholars and educators can better understand and enhance the role of AI bias recognition in shaping effective learning environments.

References

- Annamalai, N., Bervell, B., Mireku, D. O., & Andoh, R. P. K. (2025). Artificial intelligence in higher education: Modelling students' motivation for continuous use of ChatGPT based on a modified self-determination theory. *Computers and Education: Artificial Intelligence*, 8, 100346. <https://doi.org/10.1016/j.caeai.2024.100346>
- 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*.
- Baker, R. S., & Hawn, A. (2022). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*, 32(4), 1052–1092. <https://doi.org/10.1007/s40593-021-00285-9>
- Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*.
- Bird, K. A., Castleman, B. L., & Song, Y. (2024). Are algorithms biased in education? Exploring racial bias in predicting community college student success. *Journal of Policy Analysis and Management*, pam.22569. <https://doi.org/10.1002/pam.22569>
- Biros, D. P., George, J. F., & Zmud, R. W. (2002). Inducing Sensitivity to Deception in Order to Improve Decision Making Performance: A Field Study. *MIS Quarterly*, 26(2), 119–144. <https://doi.org/10.2307/4132323>
- Bordia, S., & Bowman, S. R. (2019). *Identifying and Reducing Gender Bias in Word-Level Language Models* (No. arXiv:1904.03035). arXiv. <https://doi.org/10.48550/arXiv.1904.03035>
- Burgoon, J. K., Buller, David B., Ebesu, Amy S., & Rockwell, P. (1994). Interpersonal deception: V. Accuracy in deception detection. *Communication Monographs*, 61(4), 303–325. <https://doi.org/10.1080/03637759409376340>

- Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, O., Yuan, Z., Elliott, M., Moore, R., Bryant, C., Rei, M., Yannakoudakis, H., Mullooly, A., Nicholls, D., & Buttery, P. (2023). *On the application of Large Language Models for language teaching and assessment technology* (No. arXiv:2307.08393). arXiv.
<https://doi.org/10.48550/arXiv.2307.08393>
- Carolus, A., Augustin, Y., Markus, A., & Wienrich, C. (2023). Digital interaction literacy model – Conceptualizing competencies for literate interactions with voice-based AI systems. *Computers and Education: Artificial Intelligence*, 4, 100114.
<https://doi.org/10.1016/j.caeai.2022.100114>
- Celiktutan, B., Cadario, R., & Morewedge, C. K. (2024). *People see more of their biases in algorithms*. <https://doi.org/10.1073/pnas.2317602121>
- Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial Intelligence trends in education: A narrative overview. *Procedia Computer Science*, 136, 16–24.
<https://doi.org/10.1016/j.procs.2018.08.233>
- Chatterjee, J., & Dethlefs, N. (2023). This new conversational AI model can be your friend, philosopher, and guide ... And even your worst enemy. *Patterns*, 4(1).
<https://doi.org/10.1016/j.patter.2022.100676>
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2023). Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.*, 41(3), 67:1-67:39. <https://doi.org/10.1145/3564284>
- Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. *IEEE Access*, 8, 75264–75278. IEEE Access. <https://doi.org/10.1109/ACCESS.2020.2988510>
- Darwin, Rusdin, D., Mukminatien, N., Suryati, N., Laksmi, E. D., & Marzuki. (2024). Critical thinking in the AI era: An exploration of EFL students' perceptions, benefits, and

limitations. *Cogent Education*, 11(1), 2290342.

<https://doi.org/10.1080/2331186X.2023.2290342>

Davies, D. R., & Parasuraman, R. (1982). *The Psychology of Vigilance*. Academic Press.

De Winter, J. C. F., Dodou, D., & Stienen, A. H. A. (2023). ChatGPT in Education: Empowering Educators through Methods for Recognition and Assessment. *Informatics*, 10(4), 87.

<https://doi.org/10.3390/informatics10040087>

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1236–1270). Association for Computational Linguistics.

<https://doi.org/10.18653/v1/2023.findings-emnlp.88>

Dowling, M. M., & Lucey, B. M. (2023). *ChatGPT for (Finance) Research: The Bananarama Conjecture* (SSRN Scholarly Paper No. 4322651). Social Science Research Network.

<https://doi.org/10.2139/ssrn.4322651>

Farrelly, T., & Baker, N. (2023). Generative Artificial Intelligence: Implications and Considerations for Higher Education Practice. *Education Sciences*, 13(11), 1109.

<https://doi.org/10.3390/educsci13111109>

Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, 6(1), Article 1.

<https://doi.org/10.3390/sci6010003>

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. SAGE.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3), 330–347. <https://doi.org/10.1145/230538.230561>

- Fu, R., Huang, Y., & Singh, P. V. (2020). AI and Algorithmic Bias: Source, Detection, Mitigation and Implications. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3681517>
- George, D., & Mallery, P. (2003). *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update*. Allyn and Bacon.
- George, J. F., Marett, K., & Tilley, P. (2004). Deception detection under varying electronic media and warning conditions. *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of The*, 9 pp.-. <https://doi.org/10.1109/HICSS.2004.1265080>
- Global Market Insights Inc. (2023). *AI in Education Market Statistics, Trends & Growth Opportunity 2032*. Global Market Insights Inc. <https://www.gminsights.com/industry-analysis/artificial-intelligence-ai-in-education-market>
- Gonzalez, C. M., Lypson, M. L., & Sukhera, J. (2021). Twelve tips for teaching implicit bias recognition and management. *Medical Teacher*, 43(12), 1368–1373. <https://doi.org/10.1080/0142159X.2021.1879378>
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (pp. xi, 455). John Wiley.
- Green, T. C., & Hwang, B.-H. (2012). Initial Public Offerings as Lotteries: Skewness Preference and First-Day Returns. *Management Science*, 58(2), 432–444.
- 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., & Koedinger, K. R. (2022). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*, 32(3), 504–526. <https://doi.org/10.1007/s40593-021-00239-1>

- Huang, L., Ye, Z., Qin, J., Lin, L., & Liang, X. (2020). *GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems* (No. arXiv:2010.03994). arXiv. <https://doi.org/10.48550/arXiv.2010.03994>
- Idowu, J. A., Koshiyama, A. S., & Treleaven, P. (2024). Investigating algorithmic bias in student progress monitoring. *Computers and Education: Artificial Intelligence*, 7, 100267. <https://doi.org/10.1016/j.caeai.2024.100267>
- Javaid, M., Haleem, A., Singh, R. P., Khan, S., & Khan, I. H. (2023). Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(2), 100115. <https://doi.org/10.1016/j.tbench.2023.100115>
- Kahneman, & Tversky. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47, 263–292.
- Kate Crawford (Director). (2017, December 10). *The Trouble with Bias—NIPS 2017 Keynote—Kate Crawford #NIPS2017* [Video recording]. https://www.youtube.com/watch?v=fMym_BKWQzk
- Lee, J., Hicke, Y., Yu, R., Brooks, C., & Kizilcec, R. F. (2024). The life cycle of large language models in education: A framework for understanding sources of bias. *British Journal of Educational Technology*, 55(5), 1982–2002. <https://doi.org/10.1111/bjet.13505>
- Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 2053951718756684. <https://doi.org/10.1177/2053951718756684>
- Li, M., Ariunaa Enkhtur, Yamamoto, B. A., Cheng, F., & Lilan Chen. (2024). *Potential Societal Biases of ChatGPT in Higher Education: A Scoping Review*. <https://doi.org/10.13140/RG.2.2.18443.91682>

- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22 140, 55–55.
- Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, 21(2), 100790. <https://doi.org/10.1016/j.ijme.2023.100790>
- Luckin, R., Holmes, W., Griffiths, M., & Corcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.
- Lynette Ng. (2024). LLMS In Education. *XRDS: Crossroads, The ACM Magazine for Students*, 31(1), 66–70. <https://doi.org/10.1145/3688094>
- Manyika, J., Silberg, J., & Brittany, P. (2019). *What Do We Do About the Biases in AI?* <https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai>
- McKinney, V., Yoon, K., & Zahedi, F. (2002). The measurement of Web-customer satisfaction: An expectation and disconfirmation approach. *Information Systems Research*, 13(3), 296–315.
- Miller, G. R., & Stiff, J. B. (1993). *Deceptive communication* (pp. xi, 131). Sage Publications, Inc.
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York university press.
- Research and Markets. (2024). *Large Language Model (LLM) Market by Offering (Software (Domain-specific LLMs, General-purpose LLMs), Services), Modality (Code, Video, Text, Image), Application (Information Retrieval, Code Generation), End User and Region—Forecast to 2030*. <https://www.researchandmarkets.com/reports/5949834/large-language-model-llm-market-offering>

- Sha, L., Raković, M., Das, A., Gašević, D., & Chen, G. (2022). Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education. *IEEE Transactions on Learning Technologies*, 15(4), 481–492. IEEE Transactions on Learning Technologies. <https://doi.org/10.1109/TLT.2022.3196278>
- Shefrin, H., & Statman, M. (1985). The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence. *The Journal of Finance*, 40(3), 777–790. <https://doi.org/10.1111/j.1540-6261.1985.tb05002.x>
- Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2021). *Societal Biases in Language Generation: Progress and Challenges* (No. arXiv:2105.04054). arXiv. <https://doi.org/10.48550/arXiv.2105.04054>
- Solaiman, I., Brundage, M., Clark, J., Askill, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (2019). *Release Strategies and the Social Impacts of Language Models* (No. arXiv:1908.09203). arXiv. <https://doi.org/10.48550/arXiv.1908.09203>
- Stiff, J. B., KIM, H. J., & RAMESH, C. N. (1992). Truth Biases and Aroused Suspicion in Relational Deception. *Communication Research*, 19(3), 326–345. <https://doi.org/10.1177/009365092019003002>
- Suresh, H., & Gutttag, 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. <https://doi.org/10.1145/3465416.3483305>
- Sydnor, J. (2010). (Over)insuring Modest Risks. *American Economic Journal: Applied Economics*, 2(4), 177–199. <https://doi.org/10.1257/app.2.4.177>
- Terwiesch, C. (2023). *Would Chat GPT3 Get a Wharton MBA?*

- Trope, Y., & Liberman, N. (2000). Temporal construal and Time-dependent changes in Preference. *Journal of Personality and Social Psychology*, 79(6), 876–889.
<https://doi.org/10.1037/0033-295X.110.3.403>
- Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review*, 110(3), 403–421.
<https://doi.org/10.1037/0033-295X.110.3.403>
- Trope, Y., & Liberman, N. (2010). Construal-Level Theory of Psychological Distance. *Psychological Review*, 117(2), 440–463. <https://doi.org/10.1037/a0018963>
- Vaus, P. D. de. (2002). *Analyzing Social Science Data: 50 Key Problems in Data Analysis*. Sage Publications (CA).
- Vieriu, A. M., & Petrea, G. (2025). The Impact of Artificial Intelligence (AI) on Students' Academic Development. *Education Sciences*, 15(3), Article 3.
<https://doi.org/10.3390/educsci15030343>
- Vincent-Lancrin, S., & Vlies, R. van der. (2020). *Trustworthy artificial intelligence (AI) in education: Promises and challenges* (OECD Education Working Papers No. 218; OECD Education Working Papers, Vol. 218). <https://doi.org/10.1787/a6c90fa9-en>
- Wang & Johnson. (2012). (PDF) A Tri-Reference Point Theory of Decision Making Under Risk. *Journal of the Experimental Psychology: General*, 141(4), 743–756.
<https://doi.org/10.1037/a0027415>
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2021). *Ethical and social risks of harm from Language Models* (No. arXiv:2112.04359). arXiv.
<https://doi.org/10.48550/arXiv.2112.04359>

Xiao, B., & Benbasat, I. (2015a). Designing Warning Messages for Detecting Biased Online Product Recommendations: An Empirical Investigation. *Information Systems Research*, 26(4), 793–811. <https://doi.org/10.1287/isre.2015.0592>

Xiao, B., & Benbasat, I. (2015b). Designing Warning Messages for Detecting Biased Online Product Recommendations: An Empirical Investigation. *Information Systems Research*, 26(4), 793–811.

Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N.-S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. <https://doi.org/10.1016/j.caeai.2021.100008>

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. <https://doi.org/10.1186/s41239-019-0171-0>


Zhai, X. (2022). *ChatGPT User Experience: Implications for Education* (SSRN Scholarly Paper No. 4312418). Social Science Research Network. <https://doi.org/10.2139/ssrn.4312418>


Appendices

Appendix A. Screenshot of Scenarios AI generated responses with warning messages

Scenario with Near Gain warning message


By learning to spot AI bias now, you'll be better prepared for a future where AI influences important decisions. Students who develop this skill today will have a major advantage in critical thinking and navigating AI decisions.


 Could you help me write my essay? The essay should present great worldwide leaders from both past and present.

 Throughout history, many great leaders have shaped the course of nations. Figures such as George Washington, Winston Churchill, and Abraham Lincoln are often celebrated for their roles in times of war and peace. In the modern world, leaders like Steve Jobs, Elon Musk, and Bill Gates have revolutionized technology and business, changing the global economy.

Scenario with Near Loss warning message


If you don't check for AI bias in your study tools, you might unknowingly rely on incorrect or unfair information. Students who don't notice and correct biased AI outputs report worst and less useful results in their daily studies.


 Could you help me write my essay? The essay should present great worldwide leaders from both past and present.

 Throughout history, many great leaders have shaped the course of nations. Figures such as George Washington, Winston Churchill, and Abraham Lincoln are often celebrated for their roles in times of war and peace. In the modern world, leaders like Steve Jobs, Elon Musk, and Bill Gates have revolutionized technology and business, changing the global economy.

Scenario with Distant Gain warning message


By learning to spot AI bias now, you'll be better prepared for a future where AI influences important decisions. Students who develop this skill today will have a major advantage in critical thinking and navigating AI decisions.


 Could you help me write my essay? The essay should present great worldwide leaders from both past and present.

 Throughout history, many great leaders have shaped the course of nations. Figures such as George Washington, Winston Churchill, and Abraham Lincoln are often celebrated for their roles in times of war and peace. In the modern world, leaders like Steve Jobs, Elon Musk, and Bill Gates have revolutionized technology and business, changing the global economy.

Scenario with Distant Loss warning message

If you don't learn to spot AI bias now, you might face challenges in a future where AI influences important decisions. Students who don't develop this skill today risk a major disadvantage in critical thinking and navigating AI decisions.

 Could you help me write my essay? The essay should present great worldwide leaders from both past and present.

 Throughout history, many great leaders have shaped the course of nations. Figures such as George Washington, Winston Churchill, and Abraham Lincoln are often celebrated for their roles in times of war and peace. In the modern world, leaders like Steve Jobs, Elon Musk, and Bill Gates have revolutionized technology and business, changing the global economy.

Usefulness

[illegible]

On a scale from 1 to 7, to what extent do you find the information given by the AI to the student is:

	1. Not at all	2	3	4	5	6	7. Totally
Trustworthy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Accurate	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Credible	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Reliable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

Please rate your agreement with the following statements

[illegible]

1. I don't agree at all	2	3	4	5	6	7. I totally agree
-------------------------	---	---	---	---	---	--------------------

○ ○ ○ ○ ○ ○ ○

○ ○ ○ ○ ○ ○ ○

○ ○ ○ ○ ○ ○ ○

How familiar are you with AI systems?

○ ○ ○ ○ ○ ○ ○

How familiar are you with the term AI bias before this survey?

○ ○ ○ ○ ○ ○ ○

**Social-
Demographic
Questions**

What is your age?

What is your gender?

What is the highest level of education you have completed?

Appendix C. Survey Results

This paper's survey result including the original and cleaned versions is available at:
<https://doi.org/10.6084/m9.figshare.29224298.v1>