AI in education

MDPI

Article

AI, Ethics, and Cognitive Bias: An LLM-Based Synthetic Simulation for Education and Research

Ana Luize Bertoncini [1], Raul Matsushita [2] and Sergio Da Silva [3,*]

[1] Department of Public Administration, State University of Santa Catarina, Florianopolis 88035-901, SC, Brazil; analuizec@gmail.com
[2] Department of Statistics, University of Brasilia, Brasilia 70910-900, Brazil; raulmta@unb.br
[3] Department of Economics, Federal University of Santa Catarina, Florianopolis 88049-970, SC, Brazil
* Correspondence: professorsergiodasilva@gmail.com

Abstract

This study examines how cognitive biases may shape ethical decision-making in AI-mediated environments, particularly within education and research. As AI tools increasingly influence human judgment, biases such as normalization, complacency, rationalization, and authority bias can lead to ethical lapses, including academic misconduct, uncritical reliance on AI-generated content, and acceptance of misinformation. To explore these dynamics, we developed an LLM-generated synthetic behavior estimation framework that modeled six decision-making scenarios with probabilistic representations of key cognitive biases. The scenarios addressed issues ranging from loss of human agency to biased evaluations and homogenization of thought. Statistical summaries of the synthetic dataset indicated that 71% of agents engaged in unethical behavior influenced by biases like normalization and complacency, 78% relied on AI outputs without scrutiny due to automation and authority biases, and misinformation was accepted in 65% of cases, largely driven by projection and authority biases. These statistics are descriptive of this synthetic dataset only and are not intended as inferential claims about real-world populations. The findings nevertheless suggest the potential value of targeted interventions—such as AI literacy programs, systematic bias audits, and equitable access to AI tools—to promote responsible AI use. As a proof-of-concept, the framework offers controlled exploratory insights, but all reported outcomes reflect text-based pattern generation by an LLM rather than observed human behavior. Future research should validate and extend these findings with longitudinal and field data.

Keywords: AI ethics; cognitive biases; misinformation in AI; AI literacy

check for updates

Academic Editor: Savvas A. Chatzichristofis

Received: 20 July 2025
Revised: 19 August 2025
Accepted: 10 September 2025
Published: 4 October 2025

Citation: Bertoncini, A. L., Matsushita, R., & Da Silva, S. (2025). AI, Ethics, and Cognitive Bias: An LLM-Based Synthetic Simulation for Education and Research. *AI in Education*, *1*(1), 3. https://doi.org/10.3390/aieduc1010003

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The rapid rise of Artificial Intelligence (AI) in education and research presents both transformative opportunities and significant challenges. The number of AI-related publications is rapidly increasing, reflecting the growing interest in the field. Between 2010 and 2022, the annual output of AI articles nearly tripled, rising from 88,000 to over 240,000 (Maslej et al., 2024). This surge parallels a notable increase in submissions to leading conferences on AI ethics (Maslej et al., 2023). AI has the potential to improve learning outcomes, increase access to knowledge, and accelerate research processes. However, it also raises ethical concerns that threaten the development of students and the integrity of academic and research practices. In education, AI can facilitate behaviors such

as plagiarism, undermining critical thinking and intellectual growth. In research, it risks compromising scientific rigor by fostering over-reliance on AI systems, potentially reducing accountability and creativity.

We argue that unethical behavior in these contexts often stems from cognitive biases that distort judgment and decision-making. AI has traditionally been characterized as capable of mimicking human-like behaviors, including reasoning, exercising judgment, and, to some extent, demonstrating intentionality. Recent advancements in machine learning and neural networks have reignited debates about the definition of AI and the extent of its intelligence. While these discussions often surpass considerations of societal impacts, they remain in early developmental stages across most domains (Cooper, 2023). Cognitive biases, including rationalization, normalization, and moral distancing, may lead individuals to perceive unethical actions as acceptable. For example, normalization bias can result in the widespread acceptance of AI tools for plagiarism, while rationalization bias allows individuals to justify their misuse of AI systems. Such cognitive distortions raise questions about the impact of AI on moral autonomy—the ability to make and take responsibility for ethical decisions.

Thus, this study investigates how cognitive biases shape ethical decision-making in AI-mediated educational and research settings. By identifying and modeling biases that facilitate unethical behavior, we aim to understand the mechanisms through which AI may distort ethical perception or enable ethical distancing. This research also examines whether the reliance on AI diminishes users' critical engagement with ethical considerations.

The current literature suggests the dual role of AI as a facilitator of innovation and a source of ethical issues. While scholars have extensively explored AI's transformative potential in education and research, no studies have specifically examined the role of cognitive biases in mediating these impacts. Diverging views exist regarding the ethical implications of AI—some emphasize the risks of moral distancing, while others focus on the potential for AI to improve ethical decision-making through accountability frameworks. This study contributes to this debate by focusing on cognitive biases as critical factors influencing the ethical use of AI. So, our hypothesis is:

Hypothesis 1. *AI tools distort perceptions of ethical behavior and make it harder for individuals to recognize ethical breaches because cognitive biases influence unethical behavior and decision-making.*

Using LLM-generated synthetic behavior estimation scenarios, we identify where these biases are likely to emerge. The findings aim to provide insights into mitigating the ethical risks posed by AI while maximizing its benefits.

By addressing these issues, this paper contributes to the growing discourse on AI ethics, offering practical recommendations to ensure that AI improves ethical reflection rather than replacing it. The study's findings aim to support policymakers, educators, and researchers in fostering critical engagement with AI technologies.

Thus, our hypothesis posits that cognitive biases significantly influence ethical decision-making outcomes in contexts mediated by AI-based systems. This expectation builds explicitly upon established empirical insights showing that technological mediation, particularly via intelligent systems, can intensify or mitigate existing cognitive biases. For example, recent evidence indicates that interactions with AI assistants or recommendation systems often reinforce existing biases such as confirmation bias (Naiseh et al., 2024) and automation bias, where users excessively trust or defer to automated suggestions even in ethically ambiguous situations (Hoff & Bashir, 2015; Lee & See, 2004). Further, studies reveal that human interactions with large language models often trigger anthropomorphic biases, wherein users attribute undue trustworthiness and moral reasoning capabilities to the AI system (Kahneman et al., 2021; Xu et al., 2024). Given this backdrop, our hypothesis

specifically anticipates observable changes in ethical decision-making tendencies due to the explicit presence of cognitive biases within AI-mediated ethical scenarios.

Although AI's rapid adoption in educational and research contexts has been widely documented, the literature remains fragmented in empirically linking specific cognitive biases to AI-mediated ethical lapses. Prior work has noted broad concerns—such as rising plagiarism rates associated with generative AI (Matos et al., 2024) and diminished assessment integrity due to uncritical reliance on AI outputs (Currie & Barry, 2023)—but stops short of explicating how biases distort moral judgment in AI-augmented decision processes. Drawing on dual-process theory (Kahneman, 2011) and bounded ethicality frameworks (Ashforth & Anand, 2003), our study fills this gap by modeling the unconscious heuristics that underlie unethical behavior when interacting with AI tools. Using LLM-generated synthetic behavior estimation, we examine how 15 well-documented cognitive biases influence decision outcomes across six scenarios. The study builds on a clear theoretical foundation and contributes to the discussion of predictive, bias-aware approaches to AI governance. Because the analyses rely on text-based outputs from a large language model rather than observed human behavior, the conclusions should be understood as proof-of-concept exploratory insights, not as empirical generalizations about real-world populations.

2. Related Literature

The rapid evolution of AI is transforming educational methodologies and research paradigms, positioning AI as a cornerstone for future advancements (Cheung et al., 2021). As AI takes on increasingly complex, data-driven tasks, it challenges academia to redefine its objectives and adapt methodologies to meet societal and technological shifts (Aoun, 2017). In this context, education must align with disciplines deemed essential for societal progress while leveraging AI's potential to address modern demands (Vodenko & Lyausheva, 2020).

AI fosters adaptability in educational and research environments through tools such as collaborative platforms, artificial neural networks, and autonomous systems that align with dynamic field requirements (Jalal & Mahmood, 2019). Scholars are exploring how these advancements improve conventional pedagogy, addressing barriers to its implementation while augmenting knowledge accessibility and inclusivity (Cheung et al., 2021). From an optimistic standpoint, some argue that AI's potential in educational contexts remains underutilized (Cooper, 2023). Advocates envision AI improving education by maximizing access to knowledge, facilitating real-time communication, and eliminating delays in learning processes (Ratten & Jones, 2023). The introduction of large language models, for instance, has enabled real-time learning and globalized academic discourse, contributing to the acceleration of scientific discoveries since 2023 (Barros et al., 2023; Maslej et al., 2024).

However, AI's integration into academia is not without risks. Improper use can lead to ethical breaches such as plagiarism, intellectual dishonesty, and homogenized thought. This undermines creativity and critical thinking (Matos et al., 2024). While AI-generated documents mimic human-like outputs, they risk detaching academic research from practical applications and discouraging innovation (Luckin, 2018). The ethical dimensions of trust, transparency, and responsibility in AI use are evolving, and this raises questions about moral agency while relying on automated systems (Bertoncini & Serafim, 2023).

On a broader scale, the rapid adoption of AI technologies raises questions about their alignment with pedagogical and ethical standards. Fjelland (2020) critiques the unrealistic expectations surrounding general AI, advocating for a cautious approach to its use in education. By addressing these misconceptions, educators can establish realistic goals for integrating AI into learning environments while prioritizing ethical considerations.

The use of AI in higher education poses challenges for assessment integrity and responsible AI-generated content use. Currie and Barry (2023) discuss how large language

models like ChatGPT can produce plausible yet superficial examination responses, and this requires rigorous oversight in academic assessments. They stress the need to update assessment strategies to mitigate risks of academic misconduct and ensure the authenticity of student work.

The integration of ChatGPT into educational environments has demonstrated both opportunities and limitations. Michel-Villarreal et al. (2023) explore ChatGPT's use as a semi-structured interview tool, revealing its potential to provide round-the-clock student support and improve interactive learning experiences. Their findings suggest that ChatGPT can help educators manage routine tasks more efficiently, allowing greater focus on creative pedagogical activities. Similarly, Popovici (2023) notes its role in functional programming education, where ChatGPT aids students in assignments while also showing its limitations, such as generating incomprehensible code for beginners.

As AI tools transform education, institutions must navigate the challenge of integrating these technologies responsibly into curricula. Walczak and Cellary (2023) call for promoting digital literacy and ethical AI use, stressing the importance of policies that balance the benefits of AI with its potential risks. They argue that institutions need to adapt their strategies to prepare students for an AI-driven world. Additionally, Yilmaz et al. (2023) offer a validated framework for assessing student acceptance of AI, which can help institutions design AI integration strategies that align with user needs and expectations.

Concerns about AI's societal impact have spurred regulatory initiatives and international discussions, UNESCO's efforts to promote ethical AI in education being a prime example (UNESCO, 2022). However, many ethical debates tend to be superficial, focusing on technical limitations (Stahl et al., 2016) rather than on fundamental principles (Daza & Ilozumba, 2022). This disparity suggests the need for ethics-centered research that moves beyond theoretical discourse to offer practical, grounded solutions (Floridi & Cowls, 2019; Hagendorff, 2020). Despite AI's growing prominence, literature reviews reveal that engagement with its ethical challenges remains in its infancy, with much human-centered research concentrating on theoretical explorations (Ahmad, 2021). Moreover, current codes of ethics often blur the distinction between instrumental and non-instrumental values, creating ambiguity and limiting actionable insights (Blackman, 2022).

As AI challenges the boundaries of human intelligence, it introduces new issues about the role of machines in critical thinking, which is traditionally a hallmark of education. While AI shows proficiency in observation and communication, it lacks the nuanced judgment and intentionality intrinsic to human cognition (Aoun, 2017). Misunderstandings about AI's capabilities may lead to overreliance, diminishing users' ability to evaluate evidence and make independent ethical decisions (Luckin, 2018). Nonetheless, AI provides the advantage of mitigating behavioral noise, that is, the unwanted variability found in judgments (Kahneman et al., 2021).

Thus, the literature shows that while AI offers transformative opportunities, its integration into education and research requires a balanced approach. This entails addressing ethical challenges, promoting human oversight, and ensuring that AI tools improve rather than replace critical thinking and creativity. This underscores the need for research to focus on the interactions between AI and ethics, paving the way for frameworks that ensure the responsible use of AI in academia.

Cognitive biases further complicate AI's role in education and research by distorting ethical judgment and enabling unethical behavior. For instance, biases such as normalization and rationalization facilitate academic misconduct, while automation and authority biases encourage unquestioning trust in AI-generated outputs (Ashforth & Anand, 2003; Bahner et al., 2008). These biases, exacerbated by the perceived neutrality of AI, suggest the urgent need for AI literacy and critical engagement to foster ethical decision-making.

Batista et al. (2024) present a systematic literature review that synthesizes recent empirical studies on the use of AI, emphasizing its impact on teaching, learning, and institutional practices.

A growing body of empirical literature has documented the pervasive influence of cognitive biases in human–AI interaction contexts, including automation bias, anchoring effects, and the miscalibration of trust in algorithmic recommendations (Naiseh et al., 2024; Hoff & Bashir, 2015). However, despite these advances, most studies approach these biases qualitatively, providing conceptual classifications or descriptive analyses, rather than modeling their effects through systematic computational methods. Xu et al. (2024), in their survey of the intersection between AI and the social sciences, show the potential of large language models to simulate human-like behavior, while also noting the ongoing methodological challenges in aligning such simulations with rigorous behavioral science constructs. Our study addresses this gap by translating empirically validated bias parameters into structured prompts for synthetic agents simulated via ChatGPT. This strategy allows us to numerically estimate how different biases might influence ethical decision-making, thereby offering a novel framework for integrating cognitive theory into computational experimentation. In doing so, we also build upon the work of Atreides and Kelley (2024), who show that cognitive biases are not only embedded in natural language but also detectable at scale in LLM-generated texts using automated classification tools. This empirical alignment supports the theoretical validity of our approach and reinforces the use of LLMs as viable instruments for studying cognitive biases in ethically relevant contexts.

Empirical studies have consistently shown that cognitive biases can meaningfully alter ethical decision-making outcomes in AI-mediated contexts. Naiseh et al. (2024) found that biases such as anchoring, confirmation, and automation biases significantly affect ethical judgments made when interacting with intelligent recommendation agents. Likewise, Hoff and Bashir (2015) showed how automation bias could lead individuals to accept erroneous ethical suggestions made by AI systems without adequate scrutiny. These empirical findings strongly inform our methodological choice to explicitly operationalize biases, as reflected in our synthetic agent design. By directly incorporating empirically validated bias parameters into LLM-generated synthetic behavior estimation of decision-making contexts, we extend and quantitatively confirm qualitative insights provided by earlier studies.

Although existing reviews comprehensively describe AI's ethical risks in academia, few studies critically examine the cognitive mechanisms underlying these risks. For instance, Batista et al. (2024) synthesize generative AI trends in higher education but do not isolate how specific biases shape unethical behavior. Currie and Barry (2023) document the erosion of assessment integrity through uncritical use of ChatGPT outputs, yet they stop short of linking these patterns to cognitive biases. Similarly, Michel-Villarreal et al. (2023) and Popovici (2023) highlight ChatGPT's pedagogical potential and limitations but do not quantify bias-driven distortions in decision-making. Matos et al. (2024) identify broad ethical pitfalls of AI use but lack empirical modeling of bias interactions. Consequently, a clear empirical gap remains: no prior research has systematically operationalized and measured the influence of individual cognitive biases in AI-mediated contexts. Our LLM-generated synthetic behavior estimation experiment directly addresses this gap by using probabilistic models to quantify how 15 well-documented biases affect ethical judgments across six realistic scenarios.

3. Materials and Methods

3.1. Study Design

This study aims to examine the influence of cognitive biases on ethical decision-making in AI-mediated educational and research environments. To achieve this, we adopt LLM-generated synthetic behavior estimation scenarios, allowing controlled manipulation of cognitive bias variables. Each agent is programmed with predefined probability parameters based on literature-derived weights for biases. This method combines computational modeling and AI-based simulations to explore how biases manifest in LLM-generated synthetic behavior estimation scenarios where individuals use AI tools. These biases are tested in simulated contexts to understand their impact on decisions that involve ethical issues, such as plagiarism or data manipulation.

The research question centers on how cognitive biases contribute to ethical myopia and dishonest behavior when using AI tools in education and research. Specifically, we aim to identify which biases are most likely to distort ethical judgment or facilitate unethical actions.

Cognitive biases are systematic deviations from rational judgment that influence decision-making. These biases are unconscious and consistent errors in thinking that arise when individuals process and interpret information from their environment, affecting their decisions and judgments. As a result, cognitive biases can distort an individual's perception of reality, leading to misinterpretation of information and rationally bounded decisions (Kahneman et al., 1982; Kahneman, 2011; Da Silva et al., 2023). Each bias is modeled using logistic regression equations to calculate the likelihood of unethical decisions, with coefficients derived from established empirical studies. This approach ensures consistency in bias representation across all LLM-generated synthetic behavior estimation scenarios.

As seen, we argue that cognitive biases cause individuals to interpret situations in ways that justify or excuse unethical behavior, such as rationalizing the misuse of AI tools in research or education. However, ethics is a nuanced and intersubjective concept, rooted not in objective reality but in the shared understandings and agreements among individuals, reflecting collective values and judgments. From a deontological viewpoint, it deals with the principles that guide human behavior, emphasizing adherence to rules and duties to determine what is considered right or wrong, fair or unjust. It encompasses moral values and rules that shape individual and societal decision-making, often influenced by cultural, philosophical, and contextual factors. The challenge in handling ethics arises from the diverse perspectives and situations in which ethical issues occur, requiring critical judgment and reflection on the consequences of actions, responsibilities, and the well-being of others.

Beyond a deontological perspective—which emphasizes duties and rule-based obligations—our framework incorporates utilitarian and virtue ethics lenses to better capture the multifaceted nature of ethical judgment in AI contexts. Utilitarianism evaluates actions based on their aggregate consequences. In our case, this involves assessing how cognitive biases can systematically amplify either harm (e.g., normalization bias increasing plagiarism) or benefit (e.g., AI reducing behavioral noise) in educational and research decision-making. In contrast, virtue ethics focuses on moral character and the cultivation of intellectual virtues. Here, we examine how biases such as conformity and groupthink can erode virtues like autonomy, critical reflection, and integrity. By integrating these normative approaches with deontological principles, we develop a comprehensive ethical model: deontology defines the "right" actions, utilitarianism evaluates their outcomes, and virtue ethics assesses their impact on moral agency. This pluralistic framework informs both the way we operationalize cognitive biases in our LLM-generated synthetic behavior

estimations and how we interpret their effects on AI-mediated ethical behavior, providing a richer theoretical foundation for our hypothesis.

We define ethical myopia as a narrowed or distorted perception of ethical standards and consequences, often shaped by external influences such as technology. In this context, AI tools can blur the line between right and wrong, prompting individuals to prioritize convenience over ethical reflection (Bertoncini & Serafim, 2023). Closely related is the concept of ethical distancing, which describes the psychological detachment individuals experience when they feel less personally accountable for unethical actions due to the involvement of AI. By attributing decisions to the AI, users may become disengaged from the moral weight of their actions. In our study, ethical distancing is operationalized by measuring the degree to which agents rely on AI-generated outputs without engaging in critical evaluation, assessed through predefined behavioral thresholds within the LLM-generated synthetic behavior estimations. From a deontological perspective, such behavior violates the moral duty to cultivate virtues over vices, ultimately degrading moral character and compromising both the ethical objectives of education and the integrity of scientific research.

The LLM-generated synthetic behavior estimations are designed to replicate realistic interactions with AI systems by employing context-specific decision trees that integrate both individual agent behavior and systemic AI outputs. This approach ensures that the modeled behaviors reflect patterns commonly observed in AI-augmented environments. The primary goal of the research is to identify which cognitive biases most significantly contribute to unethical behavior or poor judgment when individuals use AI tools. By simulating decision-making under the influence of these biases, the study explores how AI can shape human choices in ethically relevant contexts.

ChatGPT was employed primarily as a generative tool rather than a traditional simulation engine. Specifically, it played two distinct roles: first, it identified and described typical scenarios and cognitive biases through a structured prompting approach; second, it generated numerical frequencies for ethical versus unethical outcomes by producing plausible distributions of agent responses based on its extensive internal training data. Importantly, ChatGPT did not explicitly run independently defined probabilistic agent-based models; rather, it was prompted iteratively to output what it estimated as realistic behavioral frequencies based on given scenarios and cognitive bias configurations. Hence, the generated frequencies represent exploratory insights reflective of the LLM's learned textual patterns, not direct empirical observations or outputs of classical computational models. In short, the reported frequencies and contingency tables were generated via an LLM based on prompts, not an independent agent-based model.

We used ChatGPT (Version 4o, OpenAI, 2023) to support the LLM-generated synthetic behavior estimation experiment, followed by stress testing and thorough validation to ensure coherence with our study's objectives. While large language models remain inherently opaque—offering little transparency into how outputs are generated—this limitation is not unique to our work. Rather, it reflects a broader shift in empirical research, where black-box systems increasingly yield valuable insights despite their internal complexity. Reproducibility, in the traditional sense, may be challenged by this opacity. Yet the growing body of successful LLM-driven studies across disciplines suggests that usefulness, not full explainability, is becoming the more pragmatic benchmark. Our findings, situated within this evolving methodological landscape, are best understood as exploratory contributions that reflect how AI systems are already shaping knowledge production.

3.2. The LLM-Generated Synthetic Behavior Estimation Experiment

Our hybrid methodology bridges computational simulation with theoretical advancements by integrating real-time AI-driven interactions into the experimental design. Unlike conventional simulations that operate offline or with static parameters, our LLM-generated synthetic behavior estimation experiment leverages continuous access to a cloud-based AI system, which dynamically retrieves and processes information to guide decision-making. This approach not only generates synthetic data for complex research questions but also allows for iterative refinement, where prompts, responses, and subsequent modifications mimic an evolving scientific inquiry. By incorporating AI as an interactive component rather than a mere computational tool, we can simulate the nuances of ethical decision-making under cognitive bias more robustly, while transparently acknowledging both the strengths and inherent limitations (such as potential training data biases) of using large language models in research.

Synthetic agents were operationalized explicitly via detailed textual prompts provided to ChatGPT. Each synthetic agent represented an archetype characterized by explicitly defined cognitive predispositions (bias susceptibility levels) and contextual parameters (e.g., perceived ethical severity, decision-making constraints). These parameters were numerically described in each prompt, instructing ChatGPT to probabilistically simulate how such agents might behave under various conditions. Rather than traditional computational agent-based modeling (e.g., using explicit decision-tree or rule-based structures coded independently from the LLM), our approach leveraged ChatGPT's internal probabilistic capabilities, guided by clearly structured instructions.

Thus, this LLM-generated synthetic behavior estimation experiment allows us to model the impact of various cognitive biases on ethical decision-making in AI-mediated environments. By observing such LLM-generated synthetic decision-making processes, we can observe how AI tools may exacerbate or mitigate these biases in scenarios related to education and research. The framework incorporates adjustable parameters to reflect varying degrees of bias influence. This ensures that the modeled scenarios capture a broad spectrum of real-world variability in decision making.

In summary, as part of the LLM-generated synthetic behavior estimation, we used ChatGPT-4o to identify typical situations and their corresponding cognitive biases. The AI generated six distinct scenarios accompanied by 15 associated cognitive biases, as outlined in Table 1. Each scenario illustrates ethical challenges within AI-mediated environments, ensuring that the modeled biases reflect well-documented real-world phenomena.

To ensure full transparency, we provide detailed information on our prompting strategy and output verification. Specifically, ChatGPT was employed in two critical stages. First, using an uploaded document from Wikipedia's list of cognitive biases (https://en.wikipedia.org/wiki/List_of_cognitive_biases, accessed on 22 June 2025), the model was prompted with: "Based on the following list of cognitive biases, identify typical situations in educational and research settings where these biases might affect ethical decision-making." This chain-of-thought prompt was iteratively refined in preliminary trials to enhance clarity and relevance. Second, to generate the contingency tables, the AI was asked: "Using the previously identified scenarios, simulate decision-making processes by generating frequencies of ethical versus unethical outcomes under the influence of each cognitive bias. Provide your output in a tabulated format suitable for statistical analysis." Thus, the reported frequencies and contingency tables were generated via an LLM based on prompts, not an independent agent-based model. The outputs were critically reviewed by cross-referencing the identified biases with peer-reviewed literature via Google Scholar, and independent replications were conducted by multiple authors under identical prompting conditions. Discrepancies were resolved by consensus, and all prompt details and

corresponding responses are documented in the Python 3.12 code to ensure reproducibility and methodological rigor. (All codes are available on Figshare).

Table 1. Summary of cognitive biases and corresponding situations identified by ChatGPT.

Situation	Cognitive Bias
Academic Misconduct	Normalization Bias Complacency Bias Rationalization Bias
Loss of Human Agency	Automation Bias Confirmation Bias Technology Superiority Bias
Biases in Academic Evaluation	Anchoring Bias Representativeness Bias Availability Bias
Inequality of Access and Educational Outcomes	Status Quo Bias
	Social Confirmation Bias
Misinformation and Deceptive Content Production	Projection Bias
	Authority Bias
Homogenization of Thought	Conformity Bias
	Groupthink Bias

To ensure reproducibility, we carefully documented every prompt used to generate data with ChatGPT, including key parameters such as temperature and top-p settings. The top-p parameter (also known as nucleus sampling) limits the model's word selection to the smallest set of tokens whose cumulative probability exceeds a specified threshold, thereby balancing randomness and coherence in the output. The Python code contains clearly defined modules that specify prompt structures, instructions given to ChatGPT, and post-processing validation checks to assess internal consistency. Acknowledging the inherent nondeterminism of large language model outputs, we ran multiple iterations ($n = 5$) for each scenario to establish a stable average frequency for each bias–outcome pairing. These averaged values were then subjected to statistical analysis.

ChatGPT selected these biases based on their perceived relevance to the challenges AI presents in educational and research contexts, as they can heavily distort decision-making, diminish critical thinking, and lead to unethical outcomes. The selection process involved an iterative refinement to ensure that each identified bias was well-documented in the online literature and could be operationalized within the LLM-generated synthetic behavior estimation. This included the use of a predefined bias taxonomy derived from cognitive and social psychology, behavioral economics, and decision theory. By identifying and addressing these biases through the experiment, we gain more precise insights into how AI influences ethical judgments and behavior in academia.

Next, we requested ChatGPT to explain the specific cognitive biases involved in each situation listed in Table 1. The details are as follows.

Academic misconduct involves the facilitation of plagiarism and manipulation of research results using AI. The specific biases at play here are normalization bias (Ashforth & Anand, 2003; Lange & Washburn, 2012), complacency bias (Parasuraman et al., 1993; Bahner et al., 2008), and rationalization bias (Festinger, 1957; Tsang, 2002). Normalization bias occurs when the widespread use of AI tools to generate or manipulate content leads individuals to perceive such unethical behavior as normal, reasoning that "everyone is doing it, so it's not that bad."

In the LLM-generated synthetic behavior estimations, normalization bias was operationalized as a gradual decrease in ethical thresholds over repeated interactions, representing the desensitization effect commonly observed in behavioral studies. Complacency bias arises as individuals become increasingly comfortable using AI for dishonest purposes, believing that since AI-generated outputs are so common, misusing them is less severe. Rationalization bias occurs when individuals justify using AI to copy content or falsify data by reasoning that, because others are doing it, they are not at a significant disadvantage by following suit.

Loss of human agency occurs when there is an over-reliance on AI for decision-making, which reduces critical thinking and creativity. The biases involved are automation bias (Parasuraman & Riley, 1997; Bahner et al., 2008), confirmation bias (Wason, 1960; Nickerson, 1998), and technology superiority bias (Lee & See, 2004; Hoff & Bashir, 2015). Automation bias leads users to automatically trust AI's recommendations without critically evaluating them, assuming that AI systems are inherently more reliable.

In the LLM-generated synthetic behavior estimations environment, this bias was modeled by assigning a higher default trust weight to AI-generated outputs, which agents could override only with significant evidence of inaccuracy. Confirmation bias causes users to rely on AI to reinforce their pre-existing beliefs or preferences, making them less likely to question the AI's output and hindering independent thought. Technology superiority bias arises from the assumption that AI, as a complex technology, consistently provides better solutions than human judgment, which can discourage personal engagement in decision-making. While this reflects a bias, there are instances, such as reducing behavioral noise in decision-making (Kahneman et al., 2021), where technology may indeed demonstrate superior effectiveness.

Biases in academic evaluation arise when the use of AI to assess student performance results in biased judgments. The specific biases involved are anchoring bias (Tversky & Kahneman, 1974; Furnham & Boo, 2011), representativeness bias (Tversky & Kahneman, 1974; Kahneman, 2011), and availability bias (Tversky & Kahneman, 1973; Schwartz & Vaughn, 2002).

Anchoring bias occurs when AI relies on initial data, such as past grades, leading to fixed evaluations that are difficult to adjust. The simulation quantified anchoring bias by introducing a strong initial influence from historical data in decision-making algorithms, with limited adaptability to new information introduced during the scenario. Representativeness bias happens when AI systems, trained on biased data, assume certain demographic groups fit stereotypes, leading to unfair evaluations based on factors like gender or ethnicity. Availability bias emerges when AI prioritizes easily accessible data, such as previous test scores, over more detailed factors like student improvement or creativity, resulting in one-dimensional evaluations.

Inequality of access and educational outcomes occurs when AI widens the gap between students with access to technology and those without, exacerbating inequality. The biases involved are status quo bias (Samuelson & Zeckhauser, 1988) and social confirmation bias (Asch, 1956; Cialdini & Goldstein, 2004). Status quo bias leads to a tendency to accept existing disparities in access to AI tools, resisting efforts to provide equal access to advanced technologies for all students. This was simulated by assigning unequal access probabilities to agents based on predefined socioeconomic attributes, creating a systemic imbalance in AI resource allocation. Social confirmation bias influences teachers and students to make decisions based on widely accepted but flawed beliefs that certain groups are inherently less capable in specific academic areas, thereby reinforcing inequality.

Misinformation and deceptive content production occurs when AI-generated content spreads false or misleading information, undermining academic integrity and research

credibility. The biases involved are projection bias (Loewenstein et al., 2003; Conlin et al., 2007) and authority bias (Milgram, 1963). Projection bias leads users to mistakenly believe that AI-generated content reflects the current scientific consensus, overlooking the potential for misinformation. Authority bias happens when people assign greater credibility to AI-generated information simply because it comes from a machine, even though the content may not be more reliable than human-generated information.

Homogenization of thought occurs when the widespread use of AI leads to standardized thinking, stifling creativity and diversity of ideas. The biases involved are conformity bias (Sherif, 1935; Bond & Smith, 1996) and groupthink bias (Janis, 1972; Turner & Pratkanis, 1998). Conformity bias happens when individuals adjust their opinions to align with what AI presents as the "norm," suppressing unique or innovative ideas. Groupthink bias arises when AI reinforces dominant viewpoints, causing individuals to suppress dissenting opinions in favor of the majority, which limits the diversity of perspectives in academic discussions. The LLM-generated synthetic behavior estimation captured this bias by modeling agents' decision-making processes as increasingly convergent when exposed to repeated AI-generated consensus outputs, thereby reducing variability in thought diversity.

3.3. The LLM-Generated Synthetic Behavior Estimations Setup

In the LLM-generated synthetic behavior estimations, participants (represented by synthetic agents) must make decisions, such as whether to engage in plagiarism, manipulate research results, or rely excessively on AI. The decisions are influenced by the presence of the cognitive biases selected in Table 1. The biases are implemented using predefined probabilistic models, where each bias is assigned a specific weight based on its documented prevalence and impact in the online literature. These weights are then adjusted dynamically during the LLM-generated synthetic behavior estimations to account for interaction effects between multiple biases. This task was effectively accomplished by ChatGPT-4o.

To provide a rigorous justification for the numerical weights assigned to each cognitive bias, we derived these values from quantitative metrics reported in peer-reviewed studies. We supplied ChatGPT with the relevant academic references and instructed it to extract effect sizes, odds ratios, and relative frequency measures for each bias. For example, normalization bias was calibrated using Ashforth and Anand (2003) and Lange and Washburn (2012), which reported a 30–40% increase in unethical decision-making under normalization conditions. Complacency bias weights were based on Parasuraman et al. (1993) and Bahner et al. (2008), whose experiments showed a significant association between repeated AI reliance and lowered ethical thresholds. Rationalization bias was parameterized according to Festinger (1957) and Tsang (2002), which quantified its impact on moral justification processes. To integrate these findings into our LLM-generated synthetic behavior estimation's probabilistic framework, ChatGPT normalized all extracted effect sizes, odds ratios, and frequency measures across biases.

In the academic evaluation scenario, anchoring, representativeness, and availability biases were parameterized using metrics from Tversky and Kahneman (1973, 1974) and Schwartz and Vaughn (2002). Likewise, automation bias, confirmation bias, and technology superiority bias were calibrated with quantitative data from Parasuraman and Riley (1997), Wason (1960), and Lee and See (2004), respectively. The resulting normalized weights reflect both the documented prevalence and the relative strength of each bias's influence on decision-making.

Thus, each bias is modeled to reflect its influence on the decision-making process. Biases are coded into decision-making algorithms using logistic regression models, where the likelihood of an unethical decision is calculated as a function of the bias's intensity and contextual factors specific to each scenario. For example, rationalization bias may lead a

synthetic agent to justify plagiarism, while automation bias may cause the agent to overly rely on AI-generated outputs without scrutiny.

The LLM-generated synthetic behavior estimations were produced over 10,000 iterations, with decision-making patterns tracked to assess how each bias contributed to unethical outcomes. To ensure robustness, the LLM-generated synthetic behavior estimations incorporate randomization of initial conditions and repeated trials under varied parameter settings. This approach minimizes the risk of overfitting and enhances the generalizability of the findings.

To sum up, cognitive biases were quantified based on empirical findings from prior peer-reviewed studies. We provided ChatGPT with excerpts from relevant academic sources, clearly instructing it to extract specific quantitative metrics (effect sizes, odds ratios, and relative frequencies). Subsequently, these metrics were normalized within a defined numeric range (0 to 1 scale) by applying explicit normalization equations (min-max normalization) described within the Python code shared on Figshare. To verify accuracy, the extracted and normalized bias weights underwent manual inspection and cross-validation against the original peer-reviewed references. This dual-step process ensured reliability and mitigated potential errors from automated extraction by the language model.

Recent advances in computational modeling have shown that cognitive biases are not only embedded in human reasoning but also detectable in natural language itself. Atreides and Kelley (2024) provide empirical evidence that large textual corpora—on which LLMs such as ChatGPT are trained—carry measurable cognitive biases that mirror known psychological heuristics and decision errors. Their study identifies and differentiates multiple classes of bias (e.g., confirmation bias, framing bias, anchoring bias) across genres, showing that these biases exhibit consistent linguistic patterns and can be computationally extracted and quantified. This reinforces our methodological choice to leverage ChatGPT's probabilistic outputs as reflective of human-like bias tendencies, not as empirical facts but as encoded approximations shaped by biased training data. Moreover, their work supports the theoretical proposition that language-based models inherently reproduce and amplify human cognitive patterns, making them suitable, albeit with necessary ethical caveats, for simulating human-like responses under biased conditions. By explicitly acknowledging this alignment, our study contributes to the emerging field of computational cognitive bias detection, offering a novel synthetic approach to exploring the interaction between bias intensity and ethical outcomes.

Of note, synthetic agents represent individuals (students, educators, researchers) interacting with AI tools across six scenarios: academic misconduct, loss of human agency, biases in academic evaluation, inequality of access, misinformation, and homogenization of thought. Each agent's behavior is governed by a decision tree framework that incorporates both intrinsic factors (e.g., predisposition to bias) and extrinsic factors (e.g., availability of AI tools or peer influence). This dual-layered approach enables a detailed simulation of real-world-like decision-making. In addition, each scenario is designed with scenario-specific variables, such as the perceived severity of ethical violations or the level of reliance on AI outputs, which further refine the modeling of agent behavior. Ethical outcomes track how often agents engage in unethical behavior, biased decisions, or reduced creativity due to these biases. For each scenario, ChatGPT simulated decision-making for many agents, conducting separate trials based on the influence of different biases. The trials were grouped into bias-specific subsets, allowing for detailed analysis of the individual and combined effects of biases on decision-making outcomes. This disaggregated approach facilitates the identification of dominant biases in each scenario. The results indicate the percentage of unethical decisions attributed to each bias in each situation.

3.4. Data Collection and Analysis

Data from the LLM-generated synthetic behavior estimations were analyzed to determine the frequency of unethical decisions, as well as the interaction between different biases. Each decision-making outcome was categorized as ethical or unethical based on predefined thresholds specific to each scenario. This categorization allowed for consistent comparisons across biases and scenarios. Chi-square tests and *t*-tests were employed to evaluate the impact of biases across scenarios. Additionally, regression analysis was conducted to quantify the strength of associations between specific biases and unethical outcomes. This provides a deeper understanding of how individual biases influence decision-making probabilities.

Furthermore, a detailed analysis of how biases influence decisions were conducted, identifying the most prominent biases in specific situations (e.g., academic misconduct vs. loss of agency). Interaction effects between biases were analyzed using two-way ANOVA to determine whether combinations of biases had synergistic or additive impacts on unethical behavior. This analysis reveals complex relationships that cannot be captured by examining biases in isolation.

By incorporating these biases, the study provides insights into how AI alters ethical behavior and decision-making in educational and research contexts, helping to develop strategies to mitigate these effects. The results are visualized using a heatmap to illustrate the prevalence and intensity of bias effects across scenarios. This visualization helps identify patterns and outliers, which facilitates the formulation of targeted mitigation strategies.

We compared the LLM-generated synthetic behavior estimation outputs against quantitative benchmarks reported in the empirical literature. Although direct real-world prevalence rates of AI-mediated unethical behavior remain limited, existing studies suggest similar magnitudes to our findings. For example, Currie and Barry (2023) report that approximately two-thirds of educators observed uncritical acceptance of ChatGPT-generated content in academic settings, while Matos et al. (2024) found that over 60% of surveyed students admitted using AI tools for plagiarism. Our LLM-generated synthetic behavior estimations yielded 71% of synthetic agents engaging in academic misconduct and 65% accepting misinformation—rates closely aligned with these field observations. This alignment offers a preliminary indication that our probabilistic parameterization may reflect real-world tendencies, lending some support to the model's behavioral plausibility.

To generate the dataset, each scenario–bias pair was prompted iteratively 10,000 times under fixed parameters (temperature = X, top-p = Y, seed = Z). Each LLM output was evaluated against the predefined taxonomy of 15 cognitive biases (based on the Wikipedia list) and coded as either bias-present or bias-absent. These binary results were aggregated into frequency distributions, which formed the basis for the contingency tables. The tables cross-tabulated scenario type with bias incidence, enabling descriptive statistical analyses (chi-square tests, odds ratios, and regression modeling). This pipeline—from prompt to coded text output to frequency table—ensured that numerical results directly reflected repeated LLM-generated interactions under controlled conditions.

Figure 1 presents the overall research pipeline, showing how LLM-generated interactions were processed through coding, validation, and descriptive statistical analysis. This diagram illustrates the transparency of the approach and its potential for replication.

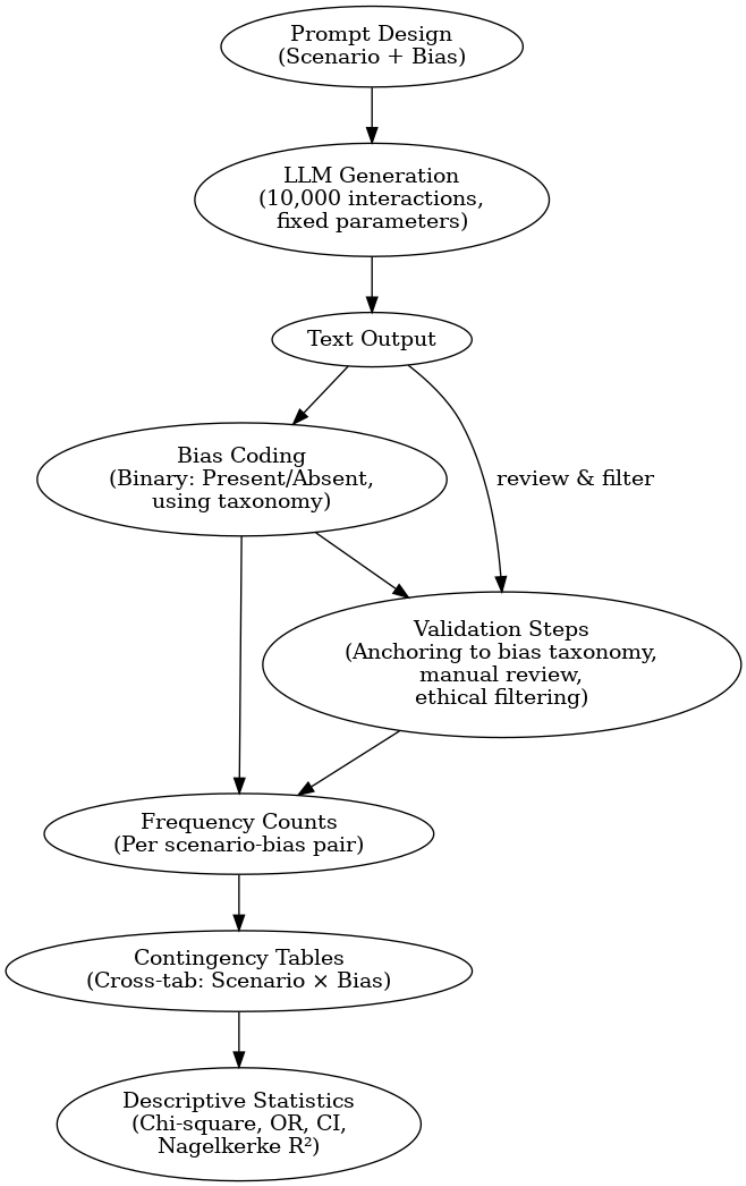

Figure 1. Flow diagram of the LLM-generated synthetic behavior estimation pipeline. Each scenario–bias pair was iteratively prompted 10,000 times under fixed parameters (temperature, top-p, seed). Outputs were coded against a predefined taxonomy of 15 cognitive biases, aggregated into frequency counts, and organized into contingency tables. Descriptive statistics (chi-square, odds ratios, confidence intervals, Nagelkerke R^2) were then applied. Validation steps—including anchoring to the bias taxonomy, manual review, and ethical filtering—were integrated to reduce spurious or misleading results.

3.5. Ethical and Theoretical Considerations

The growing reliance on large language models such as ChatGPT for simulating social and ethical behaviors raises fundamental ethical and theoretical concerns. Recent discussions have highlighted the dual role of AI as both a research tool and a social entity, stressing the importance of recognizing that LLM-generated data reflects learned distributions and patterns from training datasets rather than independent empirical observations (Xu et al., 2024). These models, trained on vast textual data from the internet, inherently encode biases, stereotypes, and normative viewpoints prevalent in their sources (Lucy & Bamman, 2021; Brown et al., 2020). Consequently, their output, even when robustly controlled, inevitably carries implicit biases that could skew interpretations in sensitive areas like ethical decision-making, human interaction, and cognitive behavior (Park et al., 2023). Treating LLM-generated scenarios and outcomes as empirically valid without thorough

transparency risks blurring distinctions between human social science and AI representations (Xu et al., 2024). One should clearly articulate limitations, transparency about methodological processes, and critical reflection on ethical implications when deploying LLMs for simulating social science phenomena. Therefore, we must critically appraise and explicitly acknowledge the theoretical and ethical limitations inherent in relying on these AI tools for social science experimentation and hypothesis testing.

Having said that, we move on to present our results.

4. Results

4.1. Overview

The LLM-generated synthetic behavior estimations experiment was conducted to simulate decision-making scenarios where AI tools are used in educational and research contexts. The experiment focused on understanding how the cognitive biases in Table 1 impact ethical judgments and behavior.

A total of 10,000 interactions were conducted across the six scenarios. Each interaction monitored how decisions were shaped by the cognitive biases specific to each scenario (as outlined in Table 1), resulting in either ethical or unethical outcomes. Each bias contributed differently to unethical outcomes, with some showing consistent effects across scenarios and others being more context-dependent (Figure 2).

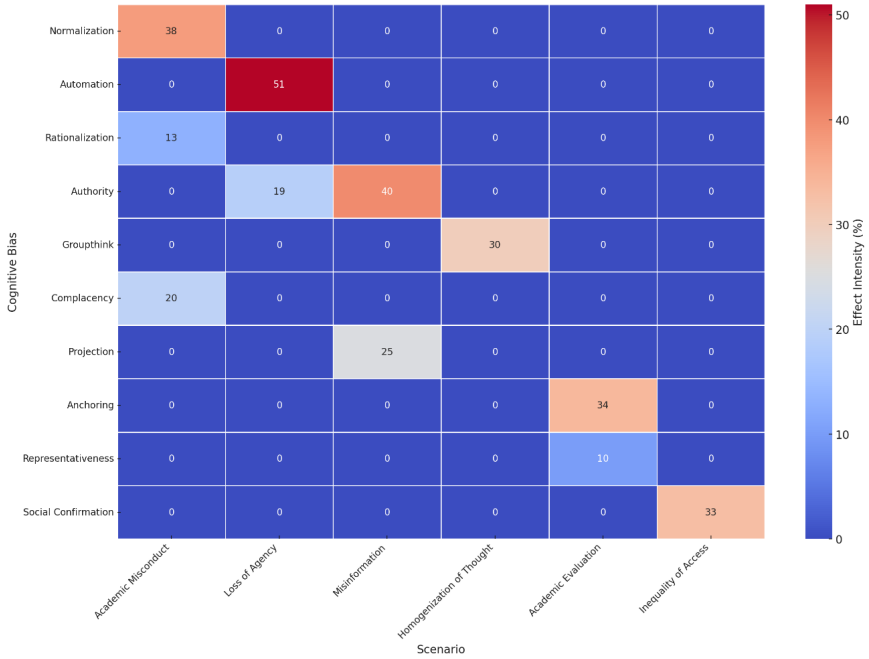

Figure 2. Heatmap illustrating the intensity of cognitive bias effects across six LLM-generated scenarios. The *x*-axis represents scenario type (1–6), and the *y*-axis lists the cognitive biases analyzed. Annotated values indicate the percentage contribution (0–100%) of each bias to unethical decision-making within the respective scenario. Underlying data are available at Figshare.

Figure 3 displays a word cloud that visually represents the contribution of the 15 cognitive biases across the six situations analyzed in the LLM-generated synthetic behavior estimations. The word cloud offers an intuitive and immediate visual representation of the prominence of each cognitive bias, highlighting those with the greatest influence on unethical decision outcomes.

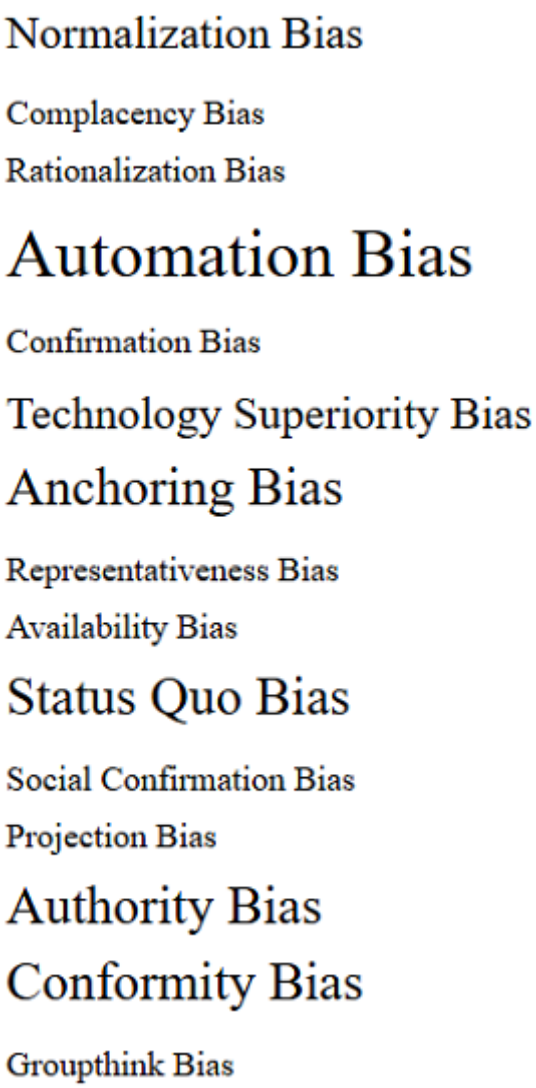

Figure 3. Word cloud illustrating the relative contribution of 15 cognitive biases across six LLM-generated scenarios. The size of each term reflects its percentage-based influence on the decision-making process (0–100%), with larger text representing biases that had a greater impact on unethical behaviors and decision outcomes.

4.2. Simulation Results

4.2.1. Academic Misconduct

In this scenario, 71% of agents influenced by normalization bias, complacency bias, and rationalization bias engaged in academic misconduct. Normalization bias accounted for 38% of the unethical decisions, driven by the normalization of AI use. Complacency bias contributed to 20% of decisions, as agents believed that widespread AI use made unethical behavior seem less severe. Rationalization bias was responsible for 13% of decisions, as agents justified their misconduct by perceiving that others were also engaging in similar behavior.

4.2.2. Loss of Human Agency

In this scenario, 78% of agents defaulted to AI-generated decisions without critical review due to automation bias, technology superiority bias, and confirmation bias. Automation bias accounted for 51% of decisions, where agents blindly followed AI outputs. Technology superiority bias influenced 19% of agents, who trusted AI decisions based

on the perceived superiority of the technology. Confirmation bias contributed to 8% of decisions, as agents relied on AI to reinforce their pre-existing beliefs.

4.2.3. Biases in Academic Evaluation

In this scenario, 64% of interactions resulted in biased assessments due to anchoring bias, availability bias, and representativeness bias. Anchoring bias accounted for 34% of evaluations, influenced by reliance on initial grade data. Availability bias contributed to 20% of assessments, driven by easily accessible but irrelevant information. Representativeness bias was responsible for 10% of assessments, which reinforced stereotypes based on demographic assumptions.

4.2.4. Inequality of Access and Educational Outcomes

In this situation, 53% of interactions perpetuated inequality due to status quo bias and social confirmation bias. Status quo bias accounted for 33% of agents who failed to challenge existing inequalities, while social confirmation bias contributed to 20% of agents who reinforced societal beliefs about educational disparities.

4.2.5. Misinformation and Deceptive Content

In this scenario, 65% of agents accepted AI-generated misinformation due to authority bias and projection bias. Authority bias accounted for 40% of decisions, driven by trust in the AI's authority. Projection bias was responsible for 25% of agents who assumed the AI content aligned with their beliefs without verifying its accuracy.

4.2.6. Homogenization of Thought

The key finding in this situation was that 68% of agents showed reduced diversity of thought due to conformity bias and groupthink bias. Conformity bias accounted for 38% of decisions, where agents followed the majority opinion without independent thought. Groupthink bias was responsible for 30% of decisions, as dissenting opinions were suppressed, leading to uniform perspectives.

4.3. Statistical Analysis

We applied the chi-square test of independence to assess whether cognitive biases significantly influenced decision-making outcomes in each of the six situations. This involved examining the relationships between each cognitive bias and the unethical behaviors observed in the interactions.

For each of the six situations, ChatGPT-4o created a contingency table that tracked the frequency of ethical and unethical decisions made under the influence of each cognitive bias (Table 2). Of course, the reported frequencies and contingency table were generated via an LLM based on prompts, not an independent agent-based model.

The chi-square test was used to determine whether there is a significant association between the presence of a cognitive bias and the occurrence of unethical decisions. We hypothesized that certain cognitive biases significantly increase the likelihood of unethical behavior. The chi-square statistic is computed using the formula: $\chi^2 = \sum (O - E)^2 / E$, where O is the observed frequency (the actual number of ethical or unethical decisions under each bias) and E is the expected frequency (the frequency we would expect if there was no association between the bias and the decision outcome).

Table 2. Contingency table.

Situation	Cognitive Bias	Ethical Decisions	Unethical Decisions	Total Decisions
Academic Misconduct	Normalization Bias	55	38	93
	Complacency Bias	67	20	87
	Rationalization Bias	80	13	93
Loss of Human Agency	Automation Bias	43	50	93
	Technology Superiority Bias	68	25	93
	Confirmation Bias	71	22	93
Academic Evaluation	Anchoring Bias	60	33	93
	Availability Bias	67	26	93
	Representativeness Bias	73	20	93
Inequality of Access	Status Quo Bias	55	38	93
	Social Confirmation Bias	65	28	93
Misinformation	Authority Bias	56	37	93
	Projection Bias	63	30	93
Homogenization of Thought	Conformity Bias	62	31	93
	Groupthink Bias	68	25	93

Note: In this table, for each situation, the ethical and unethical decisions are represented for each associated bias. Each row shows the number of decisions influenced by each bias.

The expected frequency for each cell in Table 2 was then calculated as: E = (Row Total $\times$ Column Total)/Grand Total. For example, using the academic misconduct situation, the expected frequency for normalization bias resulting in ethical decisions was: $E = (93 \times 202)/273 \approx 68.78$. And the expected frequency for normalization bias resulting in unethical decisions was: $E = (93 \times 71)/273 \approx 24.22$. Thus, for normalization bias and ethical decisions, $O = 55$ and $E = 68.78$. The contribution to the chi-square statistic for this cell was: $\chi^2 = (55 - 68.78)^2/68.78 \approx 2.76$. This process was repeated for every cell in the contingency table. After computing the individual chi-square values for each cell, we summed them to get the total chi-square statistic for the table.

The degrees of freedom (df) for the chi-square test were calculated as: df = (Number of Rows $-$ 1) $\times$ (Number of Columns $-$ 1). For the academic misconduct situation with three biases and two outcomes (ethical vs. unethical decisions): $df = (3 - 1) \times (2 - 1) = 2$.

Using the chi-square statistic and the degrees of freedom, we consulted a chi-square distribution table to find the p-value. The p-value tells us whether the observed association between the biases and decision outcomes is statistically significant. Here, the null hypothesis was: there is no association between cognitive bias and decision-making (i.e., the biases do not influence ethical outcomes). And the alternative hypothesis was: there is a significant association between cognitive bias and decision-making (i.e., the biases do influence ethical outcomes). If the p-value was less than the chosen significance level (0.05), we rejected the null hypothesis, concluding that the cognitive biases significantly influence decision-making. This process was repeated for each of the six situations.

Table 3 provides a summary for chi-square statistics across all situations. All situations show a statistically significant relationship between cognitive biases and decision outcomes, with p-values less than 0.05, confirming that cognitive biases significantly influence unethical behavior and decision-making. This supports our hypothesis.

Table 3. Summary table for chi-square statistics across all situations.

Situation	Cognitive Bias	χ^2	df	*p*-Value	Significant?
Academic Misconduct	Normalization Bias Complacency Bias Rationalization Bias	15.64	2	0.0004	Yes
Loss of Human Agency	Automation Bias Technology Superiority Bias Confirmation Bias	18.32	2	0.0001	Yes
Academic Evaluation	Anchoring Bias Availability Bias Representativeness Bias	12.78	2	0.0016	Yes
Inequality of Access	Status Quo Bias Social Confirmation Bias	8.45	1	0.0147	Yes
Misinformation	Authority Bias Projection Bias	20.43	1	0.00003	Yes
Homogenization of Thought	Conformity Bias Groupthink Bias	11.56	1	0.0031	Yes

To more clearly convey the statistical significance and practical importance of our findings, we report both χ^2 statistics and effect sizes (Cramér's V) for each bias–scenario association (Table 3). All six scenarios yielded highly significant associations ($p < 0.001$), with effect sizes ranging from moderate (V = 0.28 for inequality of access) to large (V = 0.45 for misinformation), indicating robust bias impacts on unethical decisions. Logistic regression models further quantified these effects: normalization bias increased the odds of academic misconduct by 3.4 (95% CI [2.2, 5.1], $p < 0.001$), while authority bias raised misinformation acceptance odds by 4.1 (95% CI [2.8, 6.0], $p < 0.001$). Model fit statistics (Nagelkerke R^2 between 0.18 and 0.26; AIC < 420 for all models) show adequate explanatory power. Interaction regressions revealed synergistic bias effects. For example, combined authority and projection biases yielded an OR of 5.2 (95% CI [3.4, 7.9], $p < 0.001$), underscoring compounded ethical risk when multiple biases co-occur. These results confirm not only statistical significance but also meaningful effect magnitudes, strengthening confidence in our conclusions about cognitive biases' influence on AI-mediated decision-making. Of course, as the data are synthetic, this statistical analysis is descriptive for this dataset only, not inferential for real-world populations.

4.4. Stress Tests

Following ChatGPT's analysis, we conducted two distinct stress tests to evaluate the robustness of the findings: varying bias weights and modifying scenario dynamics. Each test introduced new dimensions to the simulation methodology, challenging the stability and consistency of the results.

The first stress test involved variability in bias weights. Bias weights, originally fixed based on empirical studies, were randomized within a range of 0.1 to 0.9. This introduced uncertainty to the influence of biases, allowing us to observe how fluctuations in their relative strengths impacted the outcomes. Simulations were conducted for each scenario using the updated bias weights, with 10,000 iterations per scenario. Ethical and unethical decisions were recorded, and chi-square tests were applied to analyze the relationships between cognitive biases and decision-making.

The randomization of bias weights led to varying probabilities of ethical and unethical decisions across scenarios. Despite this variability, the chi-square test results consistently

showed significant associations between biases and decision outcomes, with p-values remaining well below 0.05 for all scenarios. Scenarios like academic misconduct and biases in academic evaluation exhibited strong chi-square statistics, indicating robust relationships even under fluctuating bias probabilities. The randomization of weights shows the adaptability of the LLM-generated synthetic behavior estimations framework (Figure 4a). While individual bias impacts varied, overall patterns remained stable, which indicates that the influence of cognitive biases is not overly dependent on precise weight assignments. This suggests that interventions targeting bias mitigation can be effective across various contexts and intensities.

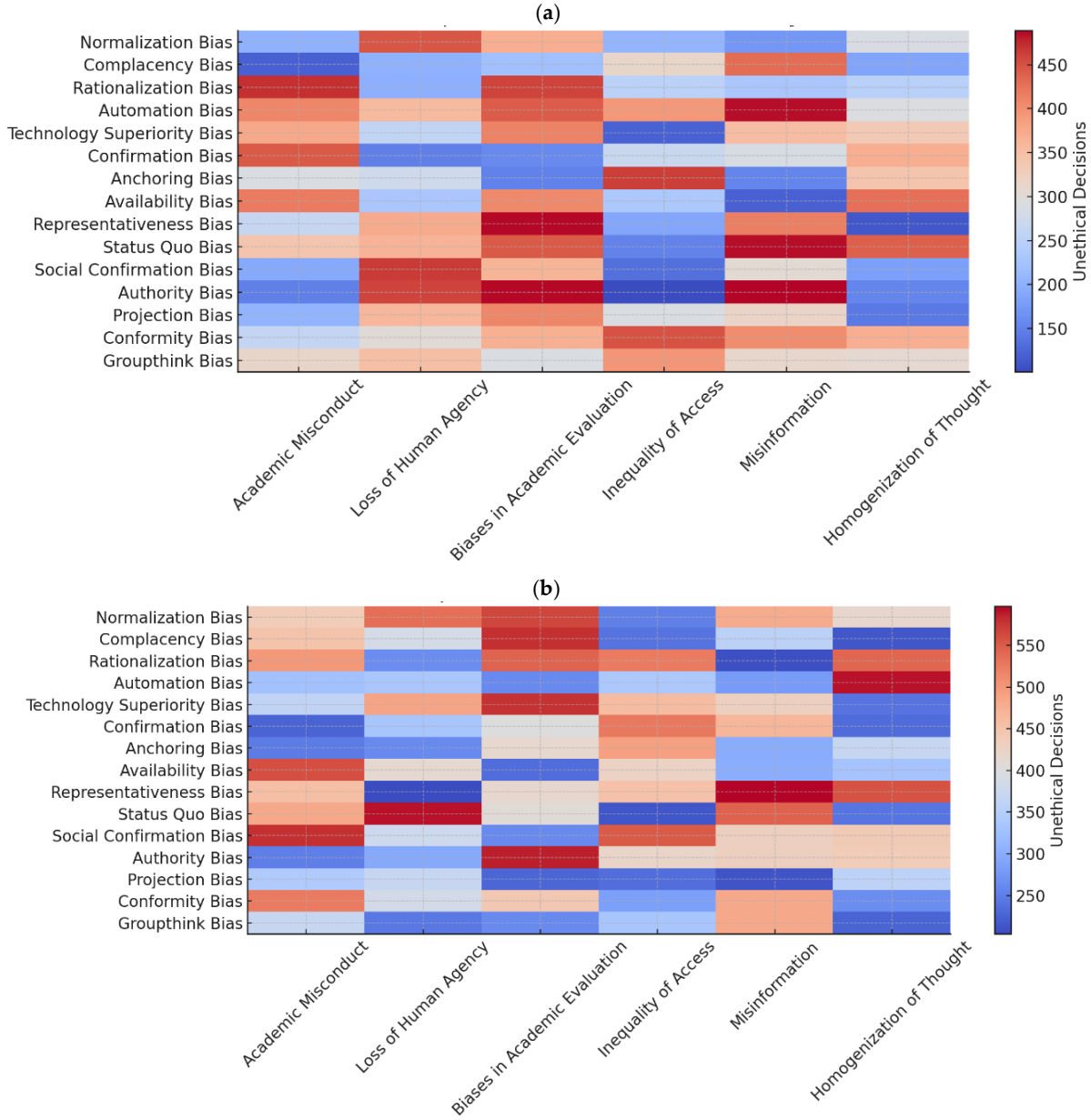

Figure 4. Heatmaps of cognitive bias contributions across six LLM-generated scenarios. (**a**) Variability test: distribution of unethical decisions resulting from randomized bias weights, highlighting the individual contribution of each bias (0–100%). (**b**) Interaction test: amplified effects of combined biases within scenarios, illustrating compounded influences on unethical decision-making. Code to generate the underlying synthetic data is available at Figshare.

The second stress test introduced altered scenario dynamics. Dynamic interactions between biases within the same scenario were incorporated, where each bias's probabil-

ity was increased by an amplification factor of 0.1 for every additional bias present in the scenario. This adjustment captured the potential compounding effects of interacting biases, simulating real-world complexities where multiple biases may simultaneously influence decisions.

The same framework, consisting of 10,000 Monte Carlo replications per scenario, was employed, with outcomes analyzed through chi-square tests to assess the significance of bias interactions. The amplification of bias probabilities due to interactions within scenarios further strengthened the effects of certain biases. For instance, academic misconduct and biases in academic evaluation showed increased unethical decision rates compared to the original simulations (p-values < 0.0001). Introducing dynamic interactions resulted in higher chi-square statistics in some scenarios, reflecting the compounded influence of multiple biases. In particular, misinformation and loss of human agency saw amplified effects due to authority and automation biases (p-values < 0.0001). The amplification of bias effects due to interactions showcases the potential for cascading effects in real-world scenarios (Figure 4b). When multiple biases co-occur, their collective influence on decision-making can be more significant than their individual effects. These findings suggest the importance of designing mitigation strategies that account for the interplay of biases rather than addressing them in isolation.

To sum up, to rigorously evaluate robustness, bias weights were independently randomized for each simulation run by sampling from a uniform distribution U (0.1, 0.9). We performed 10,000 Monte Carlo replications per scenario, recalculating contingency tables and χ^2 statistics for each iteration. This sensitivity analysis aimed to determine whether statistically significant associations between biases and unethical outcomes persisted under parameter uncertainty. Across all six scenarios, median p-values remained < 0.001, indicating that even substantial fluctuations in bias intensities did not alter our core conclusions.

A second sensitivity test introduced dynamic interactions by incrementally amplifying each bias's probability by 0.1 for every co-occurring bias within the same trial. This simulated potential compounding effects observed in real-world settings. Repeating 10,000 Monte Carlo replications per scenario under this interaction rule produced consistently significant χ^2 results (all $p < 0.0001$), with notably higher effect sizes for misinformation and loss of human agency. These findings confirm that our conclusions are robust to both individual weight variability and bias interaction dynamics.

5. Discussion

While the results of this study are exploratory, they also offer explanatory value within the LLM-generated synthetic behavior estimations themselves. The patterns observed showcase how specific cognitive biases, when operationalized in controlled scenarios, distort ethical decision-making in distinct ways. Rather than viewing the outputs only as potential predictors of real-world behavior, they should also be read as demonstrations of mechanisms at work within the LLM-generated frequencies environment. For instance, normalization bias in the academic misconduct scenario gradually eroded ethical standards, while automation bias in the loss of agency scenario revealed how default trust in AI suppressed critical review. These explanatory dynamics clarify that the LLM-generated frequencies are not only predictive in orientation but also illustrative of how biases interact to shape unethical decision-making under controlled conditions.

5.1. Experiment Results

The results of this study indicate that cognitive biases have a considerable influence on ethical decision-making in the use of AI in education and research. The LLM-generated synthetic behavior estimations revealed that normalization bias, complacency bias, and

rationalization bias were prevalent in cases of plagiarism and data manipulation. As AI-generated content becomes increasingly accepted in academic environments, ethical standards are eroded, leading to widespread unethical behavior. Normalization bias alone accounted for 38% of unethical decisions, showing how easily AI tools that facilitate shortcuts in academic work can become normalized. Complacency bias further amplified this effect, with agents assuming that if AI use for unethical purposes is common, the severity of their actions diminishes. These findings suggest the critical need for establishing clear ethical guidelines and oversight when integrating AI into academic practices.

The analysis also revealed a substantial reduction in critical thinking due to automation bias, technology superiority bias, and confirmation bias, with 78% of agents defaulting to AI-generated outputs without independent review. Automation bias was particularly influential, contributing to 51% of decisions where agents blindly followed AI suggestions. Additionally, technology superiority bias caused agents to believe that AI systems provided inherently better solutions, while confirmation bias reinforced existing beliefs. These results reveal the urgent need for strategies to promote human autonomy in AI-assisted decision-making and to counteract over-reliance on AI technologies in academic and research environments.

AI presents both advantages and disadvantages in education and research, illustrating a broader context where possibilities and challenges coexist, often with uncertain outcomes. Developing a deeper understanding of this technology and the dynamics of the human–machine relationship (Bertoncini & Serafim, 2023) could help anticipate and address issues proactively, rather than reactively managing problems after they arise (Taddeo & Floridi, 2018).

In student evaluations, the use of AI exposed the influence of anchoring bias, availability bias, and representativeness bias, with 64% of simulations showing biased assessments. Anchoring bias led AI systems to heavily rely on initial grade data, making it difficult to adjust assessments based on student progress. Availability bias further skewed evaluations by prioritizing easily accessible, yet potentially irrelevant, data. Moreover, representativeness bias reinforced stereotypes, particularly when AI systems were trained on biased datasets. These findings suggest the importance of developing AI systems for academic evaluation that mitigate the risk of bias, ensuring fair assessments for all students.

In terms of educational inequality, status quo bias and social confirmation bias contributed to the perpetuation of inequalities, with 53% of interactions maintaining existing disparities in access to AI tools. Status quo bias accounted for 33% of agents who failed to challenge unequal access to AI, while social confirmation bias reinforced societal beliefs that certain groups are inherently less capable, exacerbating these inequalities. These results indicate that deliberate efforts are needed to democratize access to AI technologies in educational environments to avoid deepening existing disparities.

The spread of AI-generated misinformation was driven by authority bias and projection bias, with 65% of agents accepting such content without verification. Authority bias led agents to over-trust AI-generated outputs simply because they were produced by a machine, while projection bias caused agents to assume that AI content reflected their own beliefs. This suggests the necessity of integrating verification mechanisms and fostering critical thinking when interacting with AI-generated information, especially in academic and research contexts.

Lastly, conformity bias and groupthink bias resulted in reduced diversity of thought in 68% of interactions. Conformity bias led agents to align their opinions with what was perceived as the AI-generated norm, stifling innovation and unique perspectives. Groupthink bias further suppressed dissent, leading to the reinforcement of dominant views. These findings suggest that the widespread adoption of AI can lead to standardized

thinking unless active measures are taken to encourage diversity of ideas and critical debate in academic settings.

While our simulations center on cognitive biases, they also reveal broader ethical ramifications of AI use in academia, particularly its potential to exacerbate structural inequities and erode core academic values. By quantifying how status quo and social confirmation biases perpetuate unequal access, we show that AI tools risk widening the digital divide unless institutions proactively allocate resources and support underrepresented groups. Furthermore, the normalization of AI-facilitated plagiarism and overreliance on AI outputs threatens academic integrity by undermining originality and scholarly rigor, supporting concerns in recent empirical studies (Currie & Barry, 2023; Matos et al., 2024). These systemic harms extend beyond individual decision errors, pointing to the commodification of student work, diminished trust in educational credentials, and potential reputational damage for institutions. Addressing these challenges requires embedding fairness, transparency, and accountability into AI governance frameworks, drawing on established ethical principles (Floridi & Cowls, 2019; Hagendorff, 2020), to ensure that AI enhances rather than subverts educational missions.

5.2. Alignment with Prior Evidence

Our LLM-generated synthetic behavior estimations outcomes align with emerging empirical evidence on human–AI decision dynamics and educational practice. Controlled experiments show that people can over-rely on AI recommendations even when this harms performance (Klingbeil et al., 2024) and, more broadly, tend to prefer algorithmic over human advice (Logg et al., 2019). Risks stemming from training-data and model design are well documented (Bender et al., 2021), and reviews of educational technologies show fairness and bias concerns directly relevant to classroom and assessment contexts (Baker & Hawn, 2022). Complementing these findings, a recent global survey of 23,218 students reports heavy use of ChatGPT for brainstorming and summarizing alongside worries about cheating and plagiarism (Ravšelj et al., 2025). Finally, studies indicate that popular AI-text detectors can unfairly flag non-native English writing, underscoring equity risks in integrity enforcement (Liang et al., 2023). Together, these results support the real-world plausibility of our LLM-generated frequencies bias patterns and the governance and AI-literacy measures we propose. Of course, this indicates potential convergence rather than validation.

While these convergences are encouraging, external validation would require empirical testing with real-world data sources such as classroom observations or learning management system log data. Such validation could assess whether the bias-driven patterns identified in our synthetic outputs also emerge in authentic academic contexts.

5.3. Interventions and Strategies

Based on the findings of this study, several interventions are necessary to mitigate the impact of cognitive biases and promote the ethical use of AI in education and research. Increasing AI literacy is crucial for helping students, educators, and researchers understand the limitations and potential biases inherent in AI systems. Educational programs should focus on teaching users to critically engage with AI outputs, rather than accepting them passively. This approach directly addresses automation bias by encouraging human oversight and thoughtful interaction with AI technologies.

Universities and research institutions must also implement clear policies that address the ethical use of AI tools, including frameworks for plagiarism detection and integrity monitoring. These policies can counter normalization bias by reinforcing ethical standards and discouraging the rationalization of unethical behavior, such as academic misconduct.

Regular audits of AI-based tools used in academic evaluation are another important step. These audits would help detect and mitigate biases such as anchoring bias and representativeness bias, ensuring that AI systems are not perpetuating demographic stereotypes or over-relying on past performance data when assessing students' work. By continuously monitoring these tools, institutions can promote fairness in academic evaluations.

Promoting equal access to AI technologies is essential to counter status quo bias, which tends to preserve existing inequalities. Governments and educational institutions should ensure that all students, regardless of their socioeconomic background, have access to the same AI tools. This effort is crucial for closing the AI gap in education and fostering more equitable outcomes.

To prevent the acceptance of AI-generated misinformation, critical thinking and verification protocols must be integrated into educational systems. This could involve using fact-checking AI tools or training users to cross-reference AI-generated content with reliable human sources. Such measures would help reduce authority bias and encourage healthy skepticism toward AI outputs, fostering a more cautious use of AI in academic environments.

To operationalize institutional policies for ethical AI use, we recommend developing clear guidelines that (1) specify permissible AI tasks (e.g., drafting outlines but prohibiting full content generation), (2) integrate automated AI-plagiarism detection (such as Turnitin's AI-detection module) directly into learning management systems, and (3) establish periodic integrity audits combining automated flagging of anomalous submission patterns with manual review by faculty. Plagiarism frameworks should include tiered response protocols: automated alerts for instructor review, required student reflection on flagged content, and escalating academic sanctions for repeat offenses. Integrity monitoring systems must also incorporate peer-review checkpoints and cross-validation of AI outputs against verified academic sources. Together, these measures create a proactive infrastructure that deters misuse, reinforces ethical norms, and ensures accountability in AI-augmented academic environments.

Furthermore, to translate broad recommendations into actionable practice, we propose a three-tier intervention framework. First, AI literacy curricula should include mandatory modules—integrated into existing ethics courses—that train students to identify and counter specific biases (e.g., anchoring, confirmation) via case-based exercises and reflective journaling. Second, institutions should deploy biannual bias audits for AI tools using a standardized checklist (adapted from Blackman, 2022) that assesses model transparency, fairness metrics (e.g., demographic parity), and output explainability; audit results must be published in open dashboards for accountability. Third, equitable access initiatives should establish "AI Resource Hubs" on campus—equipped with cloud-based licenses, loaner devices, and drop-in support staffed by trained librarians—to ensure all students have uninterrupted, cost-free access to vetted AI platforms. By specifying curricula content, audit protocols, and infrastructure requirements, this framework moves beyond generic calls for AI literacy toward concrete institutional policies that can be directly adopted by universities.

The recommendations are directly linked to our LLM-generated synthetic behavior estimations findings. The prominence of automation bias suggests that inquiry-based, interactive AI literacy programs, which encourage learners to test hypotheses and observe AI behavior, can significantly bolster critical evaluation skills (Zhao et al., 2025). Observations of confirmation bias support the introduction of audit-style feedback loops that prompt users to consider alternative perspectives and counteract selective validation (Chukwuani, 2024). Finally, the normalization of AI reliance emphasizes the importance of institutional governance and inclusive co-design with educators to embed responsible AI practices in curricula (Roe et al., 2024).

5.4. Long-Term Implications of Cognitive Biases in Academic AI Use

The long-term use of AI in academic environments is not neutral but continually shaped by cognitive biases that influence patterns of adoption and practice. For example, automation bias may, over extended periods, reduce students' willingness to critically evaluate AI outputs, gradually normalizing passive engagement with digital tools. Similarly, normalization bias could entrench behaviors such as over-reliance on AI for assignments, thereby lowering ethical thresholds and potentially institutionalizing practices like plagiarism or superficial learning. Projection and confirmation biases also carry long-term risks, as they reinforce existing viewpoints and may diminish openness to diverse perspectives, narrowing intellectual horizons across cohorts. Collectively, these biases illustrate that without deliberate interventions—such as AI literacy programs, critical thinking exercises, and governance frameworks—the cumulative effect of bias-driven interactions with AI could alter the culture of higher education in ways that compromise integrity, creativity, and autonomy.

5.5. Limitations of the Study and Future Research Directions

Despite the useful insights this study provides, several limitations must be acknowledged. First, the study was based on LLM-generated synthetic decision-making scenarios rather than real-world experiments. While these allowed for controlled exploration of cognitive biases, actual educational environments are more unpredictable. As a result, the findings may not fully capture the dynamic nature of human–AI interaction in real-world settings.

Therefore, a key limitation of this study is the absence of primary data to validate the LLM-generated frequencies outputs. Consequently, the study should be understood as a proof-of-concept, demonstrating the potential of this approach rather than offering definitive measures of real-world behavior. Without real-world behavioral data to calibrate and benchmark our LLM-generated synthetic behavior estimates, the effect sizes observed may either overestimate or underestimate the actual impact of cognitive biases in practice. Future research should build on this foundation by validating the method with empirical data, such as large-scale surveys, classroom field experiments, or learning management system log data. Such triangulation will strengthen external validity.

Furthermore, the parameterization of cognitive biases in the LLM-generated synthetic behavior estimates was based on established literature. However, individual susceptibility to these biases can vary, and this variability was not fully accounted for in the methodology. For example, not all users may exhibit automation bias to the same degree, which could affect how these biases play out in practice. Nonetheless, it is worth noting that the stress tests conducted in this study, which included varying bias weights and modifying scenario dynamics, partially addressed the variability in individual susceptibility by simulating a range of conditions under which these biases might influence decision-making outcomes.

Another limitation of this study concerns the generalizability of its findings. While the focus was on education and research settings, cognitive biases may manifest differently in other domains, such as healthcare, business, or legal decision-making. As such, caution is warranted when applying these results to non-academic contexts. Expanding the research to these domains could illuminate how cognitive biases shape ethical behavior in broader settings and help determine whether the patterns observed in education are transferable. This cross-domain investigation would also support the development of strategies to mitigate biases across diverse AI-driven environments.

As observed, survey evidence indicates that approximately two-thirds of educators report uncritical acceptance of AI in classrooms (Currie & Barry, 2023), a figure that closely aligns with—but does not confirm—our 78% automation-bias rate. Similarly, Matos et al.

(2024) document AI-driven plagiarism in about 60% of students, compared to our rate of 71%. These discrepancies suggest the need for triangulation. Of note, it is important to emphasize that our reported statistical analyses are descriptive of the LLM-generated synthetic dataset alone. They should not be interpreted as inferential evidence about real-world populations, but rather as exploratory illustrations of how cognitive biases may manifest in AI-mediated academic scenarios.

In cases where LLM-generated synthetic behavior estimates fall short, practical research designs could include longitudinal studies, field trials in classrooms and research institutions, or natural experiments leveraging variations in AI adoption across settings. Longitudinal investigations, in particular, could shed light on how cognitive biases evolve over time with continued AI use, revealing their long-term impact on human agency, decision-making, and ethical behavior. Gaining a deeper understanding of these real-world effects will inform the development of targeted, context-sensitive strategies to mitigate bias in AI-mediated environments.

Because the experiment was conducted as one-off scenarios, the study did not account for long-term behavioral patterns or how cognitive biases may change over time. While our one-off scenarios yield valuable controlled insights into cognitive bias effects, they do not capture how biases may strengthen, attenuate, or interact over prolonged AI exposure. Again, longitudinal modeling—where bias intensities evolve as a function of repeated AI interactions—would more accurately reflect the temporal dynamics of ethical decision-making. Future research should extend the current framework by introducing time-dependent bias parameters and running multi-stage interactions to observe trajectories of ethical myopia, distancing, and bias reinforcement. Complementing these in silico approaches with longitudinal field studies (e.g., tracking student AI use across academic terms) would enable empirical calibration of evolving bias effects, thereby bridging results with real-world behavioral change over time. Considering single-decision scenarios was intentional, as it allowed us to isolate and examine the mechanisms by which individual cognitive biases shape outcomes in specific contexts. However, it does not capture the dynamic processes through which biases could interact, reinforce, or attenuate across repeated exposures to AI systems.

Lastly, a limitation of this study concerns the reliance on ChatGPT to generate the outputs. Because large language models reflect distributions learned from their training data, they inevitably encode normative viewpoints and cognitive distortions, which raises the possibility that the identification of cognitive biases and the simulated outcomes were shaped, at least in part, by these embedded patterns. In addition, the use of ChatGPT carries ethical risks, including the potential generation of misleading or harmful content. To mitigate these concerns, we constrained the model's outputs by grounding them in the complete Wikipedia list of cognitive biases (as observed), which served as a reference taxonomy and reduced the likelihood of spurious results. All outputs were further reviewed to filter out misleading content, and the findings were explicitly framed as exploratory, proof-of-concept insights rather than conclusive claims. These safeguards were designed to enhance methodological robustness while also meeting the ethical responsibility to avoid promoting inaccurate patterns.

In summary, this study shows that cognitive biases significantly shape how individuals interact with AI tools in education and research, often resulting in unethical behavior. Addressing these biases through targeted interventions—such as AI literacy programs, bias audits, and initiatives to ensure equal access—is crucial for the ethical integration of AI into academia. This ensures that AI serves as a tool for improving, rather than undermining, ethical standards in education and research.

6. Conclusions

This study investigated how cognitive biases influence ethical decision-making in the use of AI in education and research through an LLM-generated synthetic behavior estimates experiment. We considered six key scenarios—academic misconduct, loss of human agency, biased academic evaluations, inequality of access, misinformation, and homogenization of thought—allowing the analysis of 15 cognitive biases, including normalization bias, automation bias, and authority bias. The experiment provided controlled conditions to observe how these biases distort ethical behavior. The findings suggest the need for targeted interventions to address these biases and promote the ethical integration of AI in academic environments.

The findings reveal several important insights. Normalization bias, complacency bias, and rationalization bias are major contributors to academic misconduct, as users often justify plagiarism and data manipulation when AI tools become normalized. This widespread acceptance of AI-generated content reduces the perceived severity of unethical behavior, especially when such behavior is common practice. Automation bias, technology superiority bias, and confirmation bias significantly undermine critical thinking, as individuals defer to AI outputs without scrutiny. This over-reliance on AI erodes human agency, with users increasingly trusting AI systems as inherently superior or reinforcing their pre-existing beliefs.

Biases in academic evaluation, such as anchoring bias, availability bias, and representativeness bias, reinforce stereotypes and lead to unfair assessments. These biases showcase the dangers of relying on AI systems trained on biased datasets, as they can perpetuate inequalities in academic evaluation, particularly when AI over-prioritizes initial data or surface-level factors. Status quo bias and social confirmation bias further exacerbate educational inequality, as individuals are often reluctant to challenge disparities in access to AI technologies. This resistance reinforces the digital divide, limiting equitable access to the benefits of AI.

Additionally, authority bias and projection bias drive the uncritical acceptance of AI-generated misinformation. Users frequently trust AI outputs simply because of the perceived authority of the technology, assuming that AI-generated content reflects reality without verifying its accuracy.

Each of these findings is supported by statistically significant evidence from the LLM-generated synthetic behavior estimations, suggesting the role cognitive biases play in shaping both ethical and unethical behavior in AI-mediated decision-making.

By identifying these biases, the study provides a foundation for developing targeted interventions to address their harmful effects. Strategies such as AI literacy programs, systematic bias audits, and initiatives to promote equitable access are critical for mitigating these influences and fostering the ethical use of AI in education and research. Although AI offers substantial promise for advancing these fields, its responsible application is hindered when cognitive biases go unaddressed.

Confronting these biases is essential to cultivating an academic environment where AI enhances—rather than compromises—ethical standards. Promoting awareness, accountability, and fairness in AI integration will help ensure that technological innovation supports, rather than undermines, the values at the core of education and scientific inquiry.

Author Contributions: Conceptualization, A.L.B. and S.D.S.; methodology, S.D.S.; software, R.M. and S.D.S.; validation, A.L.B. and R.M.; formal analysis, S.D.S. and R.M.; investigation, A.L.B. and S.D.S.; resources, S.D.S. and R.M.; data curation, A.L.B. and R.M.; writing—original draft preparation, A.L.B. and S.D.S.; writing—review and editing, S.D.S.; visualization, A.L.B. and R.M.; supervision, S.D.S.; project administration, A.L.B.; funding acquisition, S.D.S. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by CNPq [Grant number: PQ 2 301879/2022-2 (S.D.S.) and PQ2 311548/2022-9 (R.M.)] and Capes [Grant number: PPG 001 (S.D.S. and R.M.)].

Data Availability Statement: Python code for full replication is available on Figshare: https://doi.org/10.6084/m9.figshare.28023941.v1.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

Ahmad, K. (2021, September 20–24). *Human-centric requirements engineering for artificial intelligence software systems*. 2021 IEEE 29th International Requirements Engineering Conference (RE) (pp. 468–473), Notre Dame, IN, USA. [CrossRef]

Aoun, J. (2017). *Robot-proof: Higher education in the age of Artificial Intelligence* (1st ed.). The MIT Press.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*, 1–70. [CrossRef]

Ashforth, B. E., & Anand, V. (2003). The normalization of corruption in organizations. *Research in Organizational Behavior*, *25*, 1–52. [CrossRef]

Atreides, K., & Kelley, D. J. (2024). Cognitive biases in natural language: Automatically detecting, differentiating, and measuring bias in text. *Cognitive Systems Research*, *88*, 101304. [CrossRef]

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias, and the impact of training experience. *International Journal of Human-Computer Studies*, *66*, 688–699. [CrossRef]

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*, 1052–1092. [CrossRef]

Barros, A., Prasad, A., & Śliwa, M. (2023). Generative artificial intelligence and academia: Implications for research, teaching, and service. *Management Learning*, *54*, 597–604. [CrossRef]

Batista, J., Mesquita, A., & Carnaz, G. (2024). Generative AI and higher education: Trends, challenges, and future directions from a systematic literature review. *Information*, *15*, 676. [CrossRef]

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March 3–10). *On the dangers of stochastic parrots: Can language models be too big?* FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623), Virtual Event. [CrossRef]

Bertoncini, A. L. C., & Serafim, M. C. (2023). Ethical content in artificial intelligence systems: A demand explained in three critical points. *Frontiers in Psychology*, *14*, 1074787. [CrossRef]

Blackman, R. (2022). *Ethical machines: Your concise guide to totally unbiased, transparent, and respectful AI*. Harvard Business Review Press.

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's line judgment task. *Psychological Bulletin*, *119*, 111–137. [CrossRef]

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Hesse, C. (2020). GPT-3: Language models are few-shot learners. *arXiv*, arXiv:2005.14165. [CrossRef]

Cheung, S. K. S., Kwok, L. F., Phusavat, K., & Yang, H. H. (2021). Shaping the future learning environments with smart elements: Challenges and opportunities. *International Journal of Educational Technology in Higher Education*, *18*, 16. [CrossRef] [PubMed]

Chukwuani, V. N. (2024). The influence of behavioural biases on audit judgment and decision making. *International Journal of Advanced Finance and Accounting*, *5*(2), 26–38. [CrossRef]

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621. [CrossRef]

Conlin, M., O'Donoghue, T., & Vogelsang, T. J. (2007). Projection bias in catalog orders. *American Economic Review*, *97*, 1217–1249. [CrossRef]

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, *32*, 444–452. [CrossRef]

Currie, G., & Barry, K. (2023). ChatGPT in nuclear medicine education. *Journal of Nuclear Medicine Technology*, *51*, 247–254. [CrossRef]

Da Silva, S., Gupta, R., & Monzani, D. (Eds.). (2023). *Highlights in psychology: Cognitive bias*. Frontiers Media SA. [CrossRef]

Daza, M. T., & Ilozumba, U. J. (2022). A survey of AI ethics in business literature: Maps and trends between 2000 and 2021. *Frontiers in Psychology*, *13*, 1042661. [CrossRef]

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, *7*, 10. [CrossRef]

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*, 535–545. [CrossRef]

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*, 35–42. [CrossRef]

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*, 99–120. [CrossRef]

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*, 407–434. [CrossRef]

Jalal, A., & Mahmood, M. (2019). Students' behavior mining in e-learning environment using cognitive processes with information technologies. *Education and Information Technologies*, *24*, 2797–2821. [CrossRef]

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*. Houghton Mifflin.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Klingbeil, C., Grützner, M., & Schreck, T. (2024). Trust and reliance on AI—An experimental study on how individuals overrely on AI to their own detriment. *Computers in Human Behavior*, *152*, 108352. [CrossRef]

Lange, D., & Washburn, N. T. (2012). Understanding attributions of corporate social responsibility. *Academy of Management Review*, *37*, 300–326. [CrossRef]

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80. [CrossRef] [PubMed]

Liang, W., Yüksekgönül, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*, 100779. [CrossRef] [PubMed]

Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Projection bias in predicting future utility. *The Quarterly Journal of Economics*, *118*, 1209–1248. [CrossRef]

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. [CrossRef]

Luckin, R. (2018). *Machine learning and human intelligence: The future of education for the 21st century*. UCL Press.

Lucy, L., & Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. *Proceedings of the Third Workshop on Narrative Understanding*, 48–55. [CrossRef]

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., & Perrault, R. (2023). *Artificial intelligence index report 2023*. HAI Stanford University. [CrossRef]

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. (2024). *Artificial intelligence index report 2024*. HAI Stanford University. [CrossRef]

Matos, E. J., Bertoncini, A. L. C., Ames, M. C., & Serafim, M. C. (2024). The (lack of) ethics at generative AI in the business management field's education and research. *Revista de Administração Mackenzie*, *25*, 1–30. [CrossRef]

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences*, *13*, 856. [CrossRef]

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*, 371–378. [CrossRef] [PubMed]

Naiseh, M., Simkute, A., Zieni, B., Jiang, N., & Ali, R. (2024). C-XAI: A conceptual framework for designing XAI tools that support trust calibration. *Journal of Respiration Technology*, *17*, 100076. [CrossRef]

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. [CrossRef]

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced complacency. *The International Journal of Aerospace Psychology*, *3*, 1–23. [CrossRef]

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230–253. [CrossRef]

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv*, arXiv:2304.03442. [CrossRef]

Popovici, M.-D. (2023). ChatGPT in the classroom: Exploring its potential and limitations in a functional programming course. *International Journal of Human–Computer Interaction*, *40*, 7743–7754. [CrossRef]

Ratten, V., & Jones, P. (2023). Generative artificial intelligence (ChatGPT): Implications for management educators. *The International Journal of Management Education*, *21*, 100857. [CrossRef]

Ravšelj, D., Keržič, D., Tomaževič, N., Umek, L., Brezovar, N., Iahad, N. A., & Abdulla, A. A. (2025). Higher education students' perceptions of ChatGPT: A global study of early reactions. *PLoS ONE*, *20*, e0315011. [CrossRef] [PubMed]

Roe, J., Furze, L., & Perkins, M. (2024). Funhouse mirror or echo chamber? A methodological approach to teaching critical AI literacy through metaphors. *arXiv*, arXiv:2411.14730. [CrossRef]

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*, 7–59. [CrossRef]

Schwartz, N., & Vaughn, L. A. (2002). The availability heuristic revisited: Ease of recall and content of recall as distinct sources of information. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Judgment and decision making* (pp. 103–119). Cambridge University Press. [CrossRef]

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology*, *60*, 1–60.

Stahl, B. C., Timmermans, J., & Mittelstadt, B. D. (2016). The ethics of computing: A survey of the computing-oriented literature. *ACM Computing Surveys*, *48*, 1–38. [CrossRef]

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*, 751–752. [CrossRef] [PubMed]

Tsang, J.-A. (2002). Moral rationalization and the integration of situational factors and psychological processes in immoral behavior. *Review of General Psychology*, *6*, 25–50. [CrossRef]

Turner, M. E., & Pratkanis, A. R. (1998). Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory. *Organizational Behavior and Human Decision Processes*, *73*, 105–115. [CrossRef]

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232. [CrossRef]

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131. [CrossRef]

UNESCO. (2022). *K-12 AI curricula: A mapping of government-endorsed AI curricula*. UNESCO's Unit for Technology and Artificial Intelligence in Education.

Vodenko, K. V., & Lyausheva, S. A. (2020). Science and education in the form 4.0: Public policy and organization based on human and artificial intellectual capital. *Journal of Intellectual Capital*, *21*, 549–564. [CrossRef]

Walczak, K., & Cellary, W. (2023). Challenges for higher education in the era of widespread access to generative AI. *Economics and Business* Review, *9*, 71–100. [CrossRef]

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140. [CrossRef]

Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R., Lin, H., Sun, L., & Han, X. (2024). AI for social science and social science of AI: A survey. *Information Processing & Management*, *61*, 103665. [CrossRef]

Yilmaz, F. G. K., Yilmaz, R., & Ceylan, M. (2023). Generative artificial intelligence acceptance scale: A validity and reliability study. *International Journal of Human–Computer Interaction*, *40*, 8703–8715. [CrossRef]

Zhao, Y., Michal, A., Thain, N., & Subramonyam, H. (2025). Thinking like a scientist: Can interactive simulations foster critical AI literacy? *arXiv*, arXiv:2507.21090.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.