DeBiasMe: De-biasing Human-AI Interactions with Metacognitive AIED (AI in Education) Interventions

CHAEYEON LIM, UCL Interaction Centre, UK

While generative artificial intelligence (Gen AI) increasingly transforms academic environments, a critical gap exists in understanding and mitigating human biases in AI interactions, such as anchoring and confirmation bias. This position paper advocates for metacognitive AI literacy interventions to help university students critically engage with AI and address biases across the Human-AI interaction workflows. The paper presents the importance of considering (1) metacognitive support with deliberate friction focusing on human bias; (2) bi-directional Human-AI interaction intervention addressing both input formulation and output interpretation; and (3) adaptive scaffolding that responds to diverse user engagement patterns. These frameworks are illustrated through ongoing work on "DeBiasMe," AIED (AI in Education) interventions designed to enhance awareness of cognitive biases while empowering user agency in AI interactions. The paper invites multiple stakeholders to engage in discussions on design and evaluation methods for scaffolding mechanisms, bias visualization, and analysis frameworks. This position contributes to the emerging field of AI-augmented learning by emphasizing the critical role of metacognition in helping students navigate the complex interaction between human, statistical, and systemic biases in AI use while highlighting how cognitive adaptation to AI systems must be explicitly integrated into comprehensive AI literacy frameworks.

1 Introduction

Generative artificial intelligence (Gen AI) tools, particularly text generators with LLMs, are rapidly transforming academic environments not just as productivity aids but as potential catalysts or barriers to critical thinking [25]. The adoption of these tools for automation in various cognitive tasks (e.g., summarization, argument development, and research synthesis) rather than for augmentation poses cognitive risks in education by promoting over-reliance on AI [58], reinforcing existing beliefs [18], and offloading higher-order thinking skills to AI systems [3, 31]. Despite

Author's Contact Information: Chaeyeon Lim, chaeyeon.lim.24@ucl.ac.uk, UCL Interaction Centre, London, UK.

increasing AI literacy initiatives extending data and digital literacy frameworks to develop the skill set required for effective human-AI collaboration [34], what is less discussed are cognitive biases or human errors in information processing and decision making in this new form of human-computer interactions [2]. For example, automation bias can make students overly reliant on AI outputs [32], confirmation bias reinforces their existing beliefs [44], and anchoring bias locks them into initial AI-generated suggestions [41]. Critically, these human biases interact with algorithmic and systemic biases embedded in AI systems and datasets [45]. This interaction creates a harmful feedback loop that can amplify both types of biases: human biases influencing how users interact with AI systems, and AI biases potentially reinforcing human cognitive biases, negatively impacting the quality of learning and critical thinking [24] (See Fig. 4).
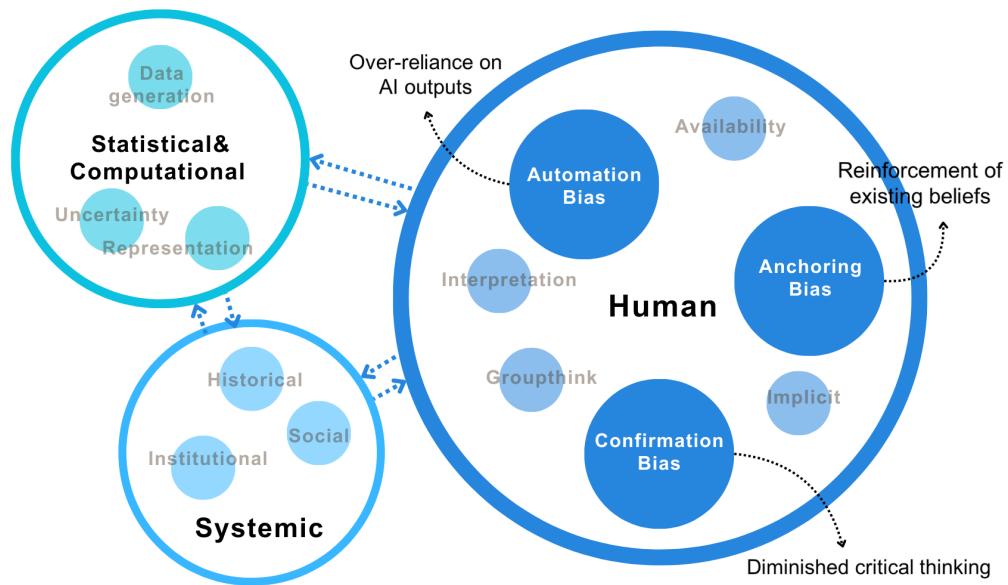
Fig. 1. Interacting categories of bias in human-AI Interaction

To bring this discussion into focus, this position paper argues that AI literacy approaches should prioritize developing human bias awareness and metacognitive skills, the ability to monitor, evaluate, and regulate one's own learning and decision-making processes when interacting with AI tools [51]. By fostering critical engagement and reflection, such tools can transform students from passive AI consumers to active collaborators with AI systems [30]. To illustrate these arguments, this paper presents work-in-progress project *DeBiasMe*, a metacognitive AIED (AI in Education) literacy intervention for human and AI bias mitigation. This will be extended by discussing ongoing research and development directions and their educational and social implications. This position paper invites educators, researchers, designers, AI developers, and students to join in *DeBiasMe*'s exploration of the following questions (See Table 1) to advance AI literacy approaches that empower end-users in AI-augmented learning environments.

Table 1. Research questions

Focused Area	Question
Deliberate Friction for Metacognitive Support	How can strategic integration of friction points in the human-AI workflow support critical thinking and metacognitive skills?
Bias Interaction Patterns	How to address human bias at both the input (prompt formulation) and output stages (interpretation of AI responses) to help students navigate the complex interaction of human, systemic, and statistical biases?
Individual factors	How to design scaffolding mechanism that adapts to individual factors involved in human-AI interaction?

2 Related Works

2.1 Technical gap: human bias in AI literacy tools

Conventional Human-in-the-Loop interventions prioritize refining AI models with less attention to human involvement in both input and output stages of the interaction [21]. As a result, existing AI literacy tools predominantly focus on algorithmic and systemic bias, with a lack of AI literacy interventions that address human biases (See Table 2). Further, existing tools are target domain-specific use cases or users with technical expertise, creating accessibility barriers that widen the digital divide and leave broader populations without effective means to develop critical AI competencies [1].

2.2 Theoretical gap: the role of metacognition in human-AI interaction

Further, there is a gap in competency dimensions that current AI literacy is advocating for: the current emphasis on technical competencies in human intervention, such as prompt engineering, and ethical and responsible AI use, which is essential for human-AI collaboration beyond automation [14, 28]. Bridging this gap in AI literacy frameworks requires AI literacy to move beyond first-level (operational understanding) to incorporate metacognitive skills by providing enhanced explainability and customizability in tool design and educational contents to support metacognitive demands in human-AI interaction [51, 53]. However, their effectiveness depends on understanding the individual factors that shape how students engage with AI. While UNESCO competency frameworks for AI literacy capture various learning dimensions [55], mitigation of biases requires not only knowledge-based analytics but also attributional and affective dimensions involving individual factors, including the confidence [48], trust [29], agency [11], and anthropomorphism tendency [48], which needs to be further addressed to better support diverse learners [16].

3 Conceptual Framework: towards learner-centered AIED

To address identified technical and theoretical gaps, design solutions with a deliberate friction approach and a new theoretical framework for bi-directional intervention will be discussed as key foundations for the development of *DeBiasMe*.

3.1 Deliberate friction in human-AI feedback loops

While the design of AI as an automatic decision-making system tends to reduce cognitive load for users and minimize interruptions in user workflows [35], design interventions known as "deliberate friction" can introduce intentional pauses and reflection opportunities [7]. This reflective and value-sensitive approach facilitates active control and independent

Table 2. A survey of existing interventions

Toolkit / Framework	Educational Focus & Features	Bias Focus Area	Learning Interaction	Target Users
Google What-If Tool [57]	Interactive visual interface to simulate changes in ML models	Fairness, AI bias	Jupyter Notebook / GUI	Undergraduate students, Data Science beginners
AI Blindspot [36]	Card-based prompts for uncovering cognitive and design biases	Cognitive bias	Workshop / Cards	High school and early college students
D-BIAS Tool [23]	Web-based visual tool showing dataset influence on AI decisions	Dataset-level bias	Web GUI	University students, HCI/UX learners
Fairlearn [43]	Fairness metrics and disparity analysis in ML outcomes	Statistical bias	Coding / Dashboard	Technical undergrads, grad students
IBM AI Fairness 360 [42]	Metrics & bias mitigation strategies toolkit	Algorithmic bias	Jupyter notebooks	Grad students, ML-focused courses
Perspective API [27]	Detects toxic or biased language	Linguistic bias	API / Front-end integration	Communication/media studies
OpenAI Moderation [37]	Filters harmful outputs from LLMs	Content moderation	API / Embedded	AI system designers, CS learners

thinking over automation, helping users adjust their agency and trust levels [12] through strategic interventions [20, 47]. These friction points serve as metacognitive triggers and knowledge scaffolding, reducing susceptibility to and amplification between cognitive and AI biases.

3.2   Bi-Directional human-AI collaboration

Such design intervention can be effectively intergrated when users are perceived as active participants in AI interaction and decision-making processes [50]. Adopting this framework in AI literacy allows reimagining the Human-in-the-Loop as a bi-directional process where humans can intervene in both input and output stages for close human-AI collaboration [15]. From this view, algorithmic and systemic biases are not imposed on users in a top-down manner, but end-users can address them with bias recognition and mitigation efforts. By empowering end-users with enhanced interpretability and transparency [22], this approach also contributes to both human-centered adoption of AI to human as well as human adaptation to AI systems [49].

4   Illustrative Case: *DeBiasMe*

To demonstrate the application of the above frameworks, tthis paper presents DeBiasMe, a metacognitive AI literacy tool designed to enhance awareness of human and AI biases in both the input and output stages of human-AI collaboration (See Fig. 2).

4.1   Identifying metacognitive support needs

Initial mixed-methods user research with university students using surveys and interviews identified three metacognitive support needs, informing the design to meet the following learner requirements: (1) bias awareness (understanding
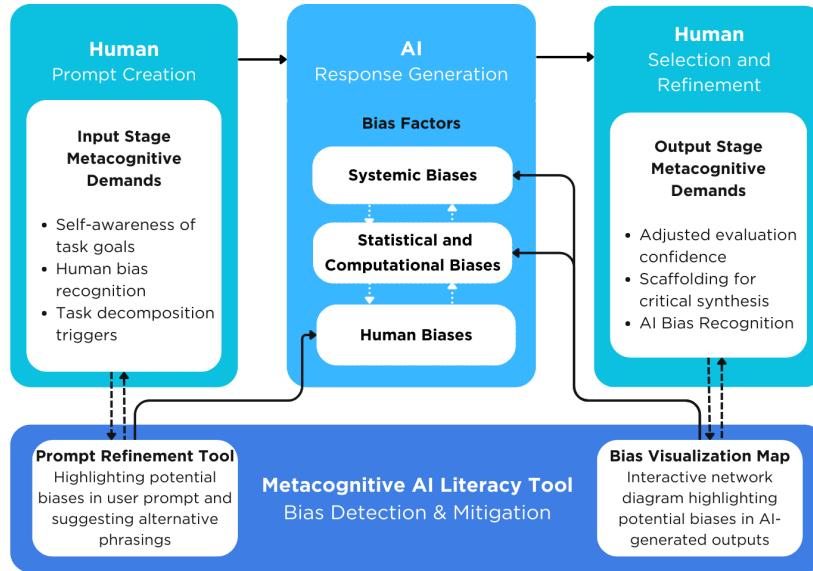
Fig. 2. Bidirectional AI Literacy Intervention in Human-AI Collaborative Framework for Metacognitive Support

how human biases affect prompt formulation); (2) AI understanding (comprehending AI's limitations and capabilities); and (3) critical thinking (developing self-awareness of reasoning processes during AI interactions).

4.2 Frictional design for metacognitive support

Based on identified metacognitive needs, design solution introduces deliberate frictions at input and output stages of the human-AI workflow through an interface that integrates two core interventions (See Fig. 3):

4.2.1 Prompt Refinement Tool. Before submitting prompts to an AI system, students receive real-time feedback highlighting potential biases and suggesting alternative phrasings. This input-stage intervention helps students identify and mitigate cognitive biases before AI interaction, such as anchoring bias in how questions are framed. When activated, the tool allows users to detect bias, reflect on their implications, and apply relevant changes for prompting.

4.2.2 Bias Visualization Map. After receiving an AI-generated response, users are presented with an interactive bias diagram highlighting potential biases in AI-generated outputs. This intervention visually maps different types of biases and their relationships by connecting information circles to the relevant text segments. By allowing users to navigate explanations of the identified bias types (human, systematic, statistical, and computational), their potential impact, and status ("addressed in prompt" and "detected in response"), the tool makes abstract bias concepts concrete and actionable.

Through the evaluation of an earlier version of the prototype, assessing both technical effectiveness and educational effectiveness with usability heuristics, think-aloud studies, and expert reviews [35, 56], the main interventions demonstrated potential as metacognitive AI literacy tools by making implicit human and AI biases explicit and actionable, helping students develop awareness of their own thinking patterns when interacting with AI. This addresses two main
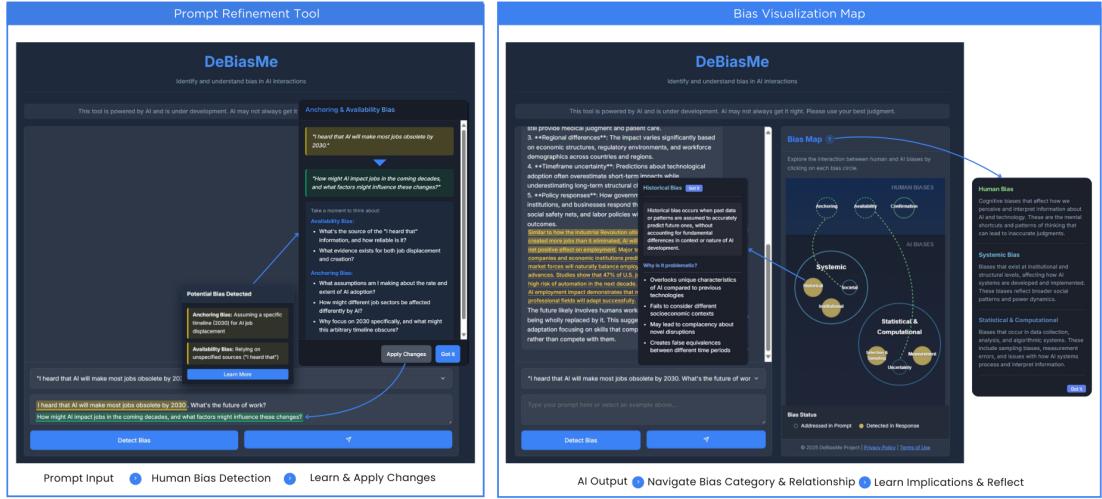
Fig. 3. Interactive prototype and frictional design elements (The previous UI has been fully integrated with AI by using GPT-4 on Azure)

components of metacognitive skills: metacognitive knowledge (enhanced awareness of biases and thought processes) and metacognitive regulation (actively monitoring and adjusting) for critical engagement with AI tools [31].

4.2.3 Bias detection approach. To provide real-time bias detection and visualization within the user's AI tool environment, the current bias detection approach uses AI to analyze text for potential biases against a hierarchical bias taxonomy established with main risks and vulnerabilities of human and AI system biases [45]. Current development is further refining bias detection by incorporating both hierarchical classification of bias types and analysis of human-AI bias relationships.

5 Developing AI Literacy Interventions for Human Bias

Following sections will outline ongoing development of the tool for developing AI literacy tools for human bias detection and mitigation: (1) enhancing scaffolding mechanisms for diverse users, (2) refining bias visualization techniques that effectively represent the relationship between different types of biases, and (3) developing comprehensive bias measurement frameworks for evidence-based evaluation across multiple bias dimensions.

5.1 Scaffolding mechanisms

Scaffolding mechanisms of the tool can be further developed by supporting diverse engagement patterns while addressing user trust.

5.1.1 AI use informed by bias implications. Tools can help students understand the broader implications of different types of biases in academic contexts. For example, anchoring bias identification is particularly relevant when addressing the pitfalls of AI collaboration that favors automation (i.e., completing tasks by finding answers to questions) over augmentation (i.e., critically questioning the questions themselves) [40]. Communicating bias implications in relation to AI capabilities and limitations can further promote ethical AI use.

5.1.2 Developing for Specific Bias-Aware Use Cases. Future tools could explore two potential directions of development: (1) training instruments for bias awareness, developing metacognitive skills in controlled educational environments where biases can be safely identified with pre-defined scenarios and discussed; and (2) real-time assistants integrated into students' everyday AI use workflows, detecting and flagging biases in different task contexts. Specific scaffolding mechanisms could focus on designing different modular components to support either workflow depending on institutional, educator, and learner needs, and developing appropriate evaluation methods for each approach.

5.1.3 Addressing User Trust in AI for Learning Support. The relationship between user bias and system bias detection creates complex trust dynamics that require further examination. When users' biases align with AI system biases, users may resist or reject the tool's feedback [17]. When there is alignment between user and system biases, identified or flagged biases from the tool can trigger defensive reactions [8] . While the tool is designed as an exploratory tool, determining whether certain features require compulsory engagement depends on educational needs and engagement patterns across users with varying levels of trust and agency.

5.2 Bias Visualization

Different bias visualization approaches are explored to further enhance the accessibility of information for users across different levels of technical expertise.

5.2.1 First-hand tool for qualitative assessments. Our working-in-progress visualization shows different bias relationship types ("reinforced" or "reduced/persistent") based on 1) structured comparison of biased and debiased alternatives [26] and 2) identified connections between human and AI biases (See Fig. 4). This approach can facilitate qualitative assessment of interactions between human, statistical, and systemic biases within and across different AI models, fostering responsible AI use [46]. For example, students can be made aware that their confirmation bias could lead them to frame prompts in ways that exploit existing cultural biases in AI training data, resulting in outputs that further reinforce their pre-existing beliefs. In other cases, students are guided to view one AI model as one among multiple options, selecting it based on an explicit assessment of its relevance and reliability.

5.2.2 Progressive Visualization. Visualizations can be integrated into different steps of the user's journey when visualization elements are introduced progressively rather than all at once. This approach would allow users to first identify bias presence, then explore bias types, and finally examine implications and mitigation strategies, reducing cognitive load while constructing a mental model of the bias in human-AI interaction [13]. The role of progressive disclosure can also be further explored in terms of features enabling the adjustment of visibility of bias severity and implications for individual users. When users try to be more sensitive to certain types of biases (e.g., confirmation), allowing personalized use becomes important. However, given that this type of visualization can also facilitate uncertainty and anchoring bias, different types of visualization need to be compared [39] .

5.2.3 Adaptive Visualization. Future tools need to support learners with different expertise levels, giving control in their learning pathways [6]. This can be done by implementing adaptive visualization complexity, allowing users to adjust the level of detail based on their comfort with the system and the specific requirements of their tasks. A promising direction for development involves implementing task-specific visualization modes tailored to different educational contexts. For example, research-focused visualizations might emphasize potential confirmation biases in source selection and interpretation, while creative writing-focused visualizations might highlight framing biases and

limitations in perspectives. This contextual adaptation would enhance the tool's relevance across various academic activities and learners.
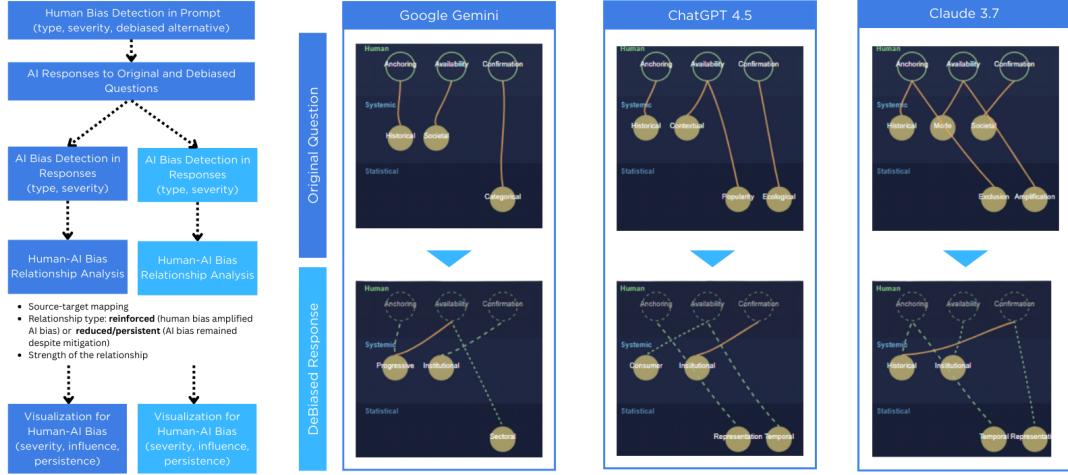


Fig. 4. Developing bias visualization: A structured Comparison of Biased and Debiased Alternatives

5.3 Bias Analysis and Mitigation Approaches

Developing quantitative metrics for bias detection and mitigation of the tool requires further technical considerations to develop a multi-dimensional bias index framework for bias detection and mitigation.

5.3.1 Bias measurement (prevalence, intensity, and mitigation). Mitigation of bias prevalence and intensity can be further assessed in terms of analysis accuracy and educational relevance. The success rate of bias mitigation strategies can be measured through changes in subsequent interactions after bias feedback. These metrics can be calculated both at the individual level and aggregated across user populations to identify common patterns and the effectiveness of the tool interventions from technical perspectives. This can further benefit from regular external auditing processes by experts in education and AI safety to assess both system performance and educational quality.

5.3.2 Methodological Integrity of Bias Analysis. A fundamental challenge in developing an AI literacy tool integrating bias detection functionality in AI is ensuring the methodological integrity of the detection system itself. This presents a multi-layered challenge. While AI biases reflect incompleteness of training data, biases in AI systems can also be deliberate for deception [38]. To address this, several safeguard methods can be considered. The tool can adopt a multi-model consensus approach that compares bias assessments from different AI systems with diverse training methodologies [9]. By identifying areas of agreement and disagreement between these systems, the tool can triangulate more reliable bias assessments while highlighting areas of uncertainty. Further, implementing transparent bias analysis systems will be crucial to clearly distinguish between different levels of confidence in bias assessments. Rather than presenting all bias detections with equal certainty, the system can communicate confidence levels based on the strength of evidence and consensus across detection methods.

6 Discussion

This position paper contributes to the emerging field of AI-augmented reasoning by emphasizing the critical role of metacognition in human-AI collaboration and the focus on human bias. There are several limitations that needs to be acknowledged to further develop this tool in line with the responsible innovation framework, which considers the potential societal impacts of emerging technology [52, 54].

6.0.1 Limitations. While this tool focused on addressing individual bias guided by self-regulation theory [5], biases in collaborative contexts in shared learning environments can be further examined. Further, this tool adopted an established bias taxonomy, but this can be further explored with co-designing sessions employing community engagement and creative approaches exploring bias-related themes and visualizations, data representation and data politics. Finally, while this tool focused on LLMs, future development on different types of biases, for example with image generation, is worth exploring with multimodal AI.

6.0.2 AI Literacy for epistemic ethics and civic imagination. AI literacy requires the assessment of complex affective, behavioral, cognitive, and ethical dimensions [16]. *DeBiasMe* provides both theoretical and technical solutions by prompting students to evaluate whether AI assistance is necessary for a given task, encouraging a more reflective approach to AI use. This paradigm shifts away from treating AI as an automatic solution to problems, instead positioning it as a strategic cognitive aid that should be deployed selectively and intentionally. This approach fosters adaptive, empathetic leadership in an era of AI sovereignty, where educational institutions and individuals maintain meaningful control over their AI infrastructure, data governance, and decision-making processes [4, 33].

6.0.3 Educational and Social Implications. This approach will benefit multiple stakeholders in the educational ecosystem: At the institutional level, tools like *DeBiasMe* can serve as bridges between academic integrity policies and students' everyday AI interactions, providing structured support for critical engagement while acknowledging AI's growing role in academic work [19]. At the societal level, the tool can help address the digital confidence divide by actively involving end-users in the research, design, and dissemination of AI literacy tools [10]. Enhanced transparency regarding both human and AI biases can reshape the user-data-algorithm lifecycle in AIED by making decision processes more visible, empowering user agency. This can further inspire policy recommendations with frictional design, introducing pauses in AIED products.

7 Conclusion

AI literacy tools must evolve beyond technical competencies to incorporate metacognitive skills that address human biases in human-AI collaboration. This position paper advocates for bias-aware AI literacy tools that encourage students to critically evaluate AI-generated content rather than defaulting to AI as an automatic authority. The illustrative tool *DeBiasMe* demonstrates AI literacy frameworks in promoting critical thinking in educational contexts by prioritizing: (1) deliberate friction to enhance metacognition, (2) bi-directional Human-AI interaction intervention, and (3) supporting diverse user engagement patterns to help students navigate the complex interactions between human and AI biases, offering a promising direction for future research and development to augment learning with learner-centered AIED.

References

[1] Fatima Ahmed. 2024. The Digital Divide and AI in Education: Addressing Equity and Accessibility. *Journal of AI Integration in Education* 1, 2 (2024), Article 2.

[2] Saar Alon-Barkat and Madalina Busuioc. 2022. Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory* 33, 1 (2022), 153–170. https://doi.org/10.1093/jopart/muac003

[3] Anthropic. 2025. Anthropic Education Report: How University Students Use Claude. https://www.anthropic.com/news/anthropic-education-report-how-university-students-use-claude Retrieved April 11, 2025.

[4] Javiera Atenas, Leo Havemann, and Chrissi Nerantzi. 2024. Critical and Creative Pedagogies for Artificial Intelligence and Data Literacy: An Epistemic Data Justice Approach for Academic Practice. *Research in Learning Technology* 32 (2024), 3296. https://doi.org/10.25304/rlt.v32.3296

[5] Roger Azevedo and Jennifer G. Cromley. 2004. Does Training on Self-Regulated Learning Facilitate Students' Learning With Hypermedia? *Journal of Educational Psychology* 96, 3 (2004), 523–535. https://doi.org/10.1037/0022-0663.96.3.523

[6] Xiaoyan Bai, David White, and David Sundaram. 2012. Contextual Adaptive Knowledge Visualization Environments. *Electronic Journal of Knowledge Management* 10, 1 (2012), 1–14. https://academic-publishing.org/index.php/ejkm/article/download/947/910/943

[7] Andrea Benedetti and Michele Mauri. 2023. Design for Friction: An Inquiry to Position Friction as a Method for Reflection in Design Interventions. *Convergences—Journal of Research and Arts Education* 16, 31 (2023), Article 31. https://doi.org/10.53681/c1514225187514391s.31.139

[8] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Transactions on Computer-Human Interaction* 12, 2 (2005), 293–327. https://doi.org/10.1145/1067860.1067867

[9] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*. 149–159. https://doi.org/10.1145/3287560.3287598

[10] Velibor Božić. 2023. Artificial Intelligence as the Reason and the Solution of Digital Divide. *Language Education and Technology* 3, 2 (2023), 96–109. https://langedutech.com/letjournal/index.php/let/article/view/53

[11] Bruno Berberian, Jean-Christophe Sarrazin, Patrick Le Blaye, and Patrick Haggard. 2012. Automation Technology and Sense of Control: A Window on Human Agency. *PLOS ONE* 7, 3 (2012), e34075. https://doi.org/10.1371/journal.pone.0034075

[12] Zana Buçinca, Michael B. Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021). https://doi.org/10.1145/3449287

[13] Paul Chandler and John Sweller. 1991. Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction* 8, 4 (1991), 293–332. https://doi.org/10.1207/s1532690xci0804_2

[14] Thomas K. F. Chiu, Zubair Ahmad, Murod Ismailov, and Ibrahim T. Sanusi. 2024. What are artificial intelligence literacy and competency? A comprehensive framework to support them. *Computers and Education Open* 6 (2024), 100171. https://doi.org/10.1016/j.caeo.2024.100171

[15] Mutlu Cukurova. 2024. The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence. *British Journal of Educational Technology* (2024). https://doi.org/10.1111/bjet.13514 arXiv:https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13514

[16] Davy Tsz Kit Ng, Wenjie Wu, Jac Ka Lok Leung, Thomas Kin Fung Chiu, and Samuel Kai Wah Chu. 2024. Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive and ethical approach. *British Journal of Educational Technology* 55, 3 (2024), 1082–1104. https://doi.org/10.1111/bjet.13411 arXiv:https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13411

[17] Mary T. Dzindolet, Scott A. Peterson, Rebecca A. Pomranky, Lora G. Pierce, and Hall P. Beck. 2003. The Role of Trust in Automation Reliance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 47. 381–385. https://doi.org/10.1177/154193120304700312

[18] Emilio Ferrara. 2023. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci* 6, 1 (2023), 3.

[19] David S. Fowler. 2023. AI in Higher Education: Academic Integrity, Harmony of Insights, and Recommendations. *Journal of Ethics in Higher Education* 3 (2023), 127–143. https://doi.org/10.26034/fr.jehe.2023.4657

[20] Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001

[21] José García, Raúl Ramos, Concha Bielza, and Pedro Larrañaga. 2022. Human-in-the-Loop Machine Learning: A State of the Art. *Artificial Intelligence Review* 55 (April 2022), 1001–1047. https://doi.org/10.1007/s10462-022-10246-w

[22] Dragan Gašević, George Siemens, and Shazia Sadiq. 2023. Empowering learners for the age of artificial intelligence. *Computers and Education: Artificial Intelligence* 4 (2023), 100130. https://doi.org/10.1016/j.caeai.2023.100130

[23] Bhavya Ghai and Klaus Mueller. 2023. D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias . *IEEE Transactions on Visualization & Computer Graphics* 29, 01 (Jan. 2023), 473–482. https://doi.org/10.1109/TVCG.2022.3209484

[24] Moshe Glickman and Tali Sharot. 2025. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour* 9, 2 (2025), 345–359. https://doi.org/10.1038/s41562-024-02077-2

[25] Agci Hikmawati and Nhelbourne K. Mohammad. 2025. Enhancing Critical Thinking with Gen AI: A Literature Review. *Buletin Edukasi Indonesia* 4, 1 (2025), 40–46. https://doi.org/10.56741/bei.v4i01.764

[26] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudik, and Hanna Wallach. 2020. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–14. https://doi.org/10.1145/3313831.3376449

[27] Google Jigsaw. 2017. Perspective API: An API to detect toxic language in text. Jigsaw via Google Cloud. https://perspectiveapi.com/ Accessed: April 2025.

[28] Nils Knoth, Marie Decker, Matthias Carl Laupichler, Marc Pinski, Nils Buchholtz, Katharina Bata, and Ben Schultz. 2024. Developing a holistic AI literacy assessment matrix – Bridging generic, domain-specific, and ethical competencies. *Computers and Education Open* 6 (2024), 100177. https://doi.org/10.1016/j.caeo.2024.100177

[29] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*. Springer International Publishing, Cham, 13–30.

[30] Carnegie Learning. 2025. From Hype to Help: The State of AI in Education. https://discover.carnegielearning.com/ai-in-education.

[31] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3706598.3713778

[32] John D Lee and Katrina A See. 2004. Trust in automation: designing for appropriate reliance. *Hum Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[33] Ziyan Lin and Yun Dai. 2025. Fostering Epistemic Insights into AI Ethics through a Constructionist Pedagogy: An Interdisciplinary Approach to AI Literacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. Article 28. https://doi.org/10.1609/aaai.v39i28.35190

[34] David Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16. https://doi.org/10.1145/3313831.3376727

[35] Jakob Nielsen. 1994. Enhancing the Explanatory Power of Usability Heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 152–158. https://doi.org/10.1145/191666.191729

[36] Partnership on AI. 2021. AI Blindspot: A Discovery Process for Early-Stage AI Risk Mitigation. Partnership on AI Toolkit. https://aiblindspot.partnershiponai.org/ Accessed: April 2025.

[37] OpenAI. 2022. OpenAI Moderation API. OpenAI API Documentation. https://platform.openai.com/docs/guides/moderation Accessed: April 2025.

[38] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. AI Deception: A Survey of Examples, Risks, and Potential Solutions. *Patterns* 5, 5 (2024), 100988. https://doi.org/10.1016/j.patter.2024.100988

[39] Marianne Procopio, Ab Mosca, Carlos Scheidegger, Eugene Wu, and Remco Chang. 2022. Impact of Cognitive Biases on Progressive Visualization. *IEEE Transactions on Visualization and Computer Graphics* 28, 9 (2022), 3093–3112. https://doi.org/10.1109/TVCG.2021.3051013

[40] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2020. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *arXiv preprint arXiv:2010.07938* (2020). https://arxiv.org/pdf/2010.07938

[41] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–22. https://doi.org/10.1145/3512930

[42] IBM Research. 2018. AI Fairness 360: A comprehensive open-source toolkit to help detect and mitigate bias in ML models. IBM Research AI. https://aif360.mybluemix.net/ Accessed: April 2025.

[43] Microsoft Research. 2021. Fairlearn: A toolkit for assessing and improving fairness in AI systems. GitHub Repository. https://fairlearn.org/ Accessed: April 2025.

[44] Emely Rosbach, Jonas Ammeling, Sebastian Krügel, Angelika Kießig, Alexis Fritz, Jonathan Ganz, Chloé Puget, Taryn Donovan, Andrea Klang, Maximilian C. Köller, Pompei Bolfa, Marco Tecilla, Daniela Denk, Matti Kiupel, Georgios Paraschou, Mun Keong Kok, Alexander F. H. Haake, Ronald R. de Krijger, Andreas F.-P. Sonnen, Tanit Kasantikul, Gerry M. Dorrestein, Rebecca C. Smedley, Nikolas Stathonikos, Matthias Uhl Christof A. Bertram, Andreas Riener, and Marc Aubreville. 2024. When Two Wrongs Don't Make a Right"–Examining Confirmation Bias and the Role of Time Pressure During Human-AI Collaboration in Computational Pathology. arXiv:2411.01007 [cs.HC] https://arxiv.org/abs/2411.01007

[45] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. https://doi.org/10.6028/NIST.SP.1270 Accessed December 12, 2024.

[46] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[47] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph "Jofish" Kaye. 2005. Reflective Design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*. 49. https://doi.org/10.1145/1094562.1094569

[48] Anuragini Shirish Shalini Chandra and Shirish C. Srivastava. 2022. To Be or Not to Be …Human? Theorizing the Role of Human-Like Competencies in Conversational Artificial Intelligence Agents. *Journal of Management Information Systems* 39, 4 (2022), 969–1005. https://doi.org/10.1080/07421222.2022.2127441 arXiv:https://doi.org/10.1080/07421222.2022.2127441

[49] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. arXiv:2406.09264 [cs.HC] https://arxiv.org/abs/2406.09264

[50] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Trans. Human-Computer Interaction* 12, 3 (2020), 109–124. https://doi.org/10.17705/1thci.00131

[51] Sidra Sidra and Claire Mason. 2024. Reconceptualizing AI Literacy: The Importance of Metacognitive Thinking in an Artificial Intelligence (AI)-Enabled Workforce. In *2024 IEEE Conference on Artificial Intelligence (CAI)*. 1181–1186. https://doi.org/10.1109/CAI59869.2024.00211

[52] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2020. Developing a Framework for Responsible Innovation. In *The Ethics of Nanotechnology, Geoengineering, and Clean Energy*. Routledge, London, 347–359.

[53] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (Honolulu, HI, USA). ACM, New York, NY, USA, 24 Pages. https://doi.org/10.1145/3613904.3642902

[54] UK Research and Innovation. 2023. Framework for Responsible Research and Innovation. https://www.ukri.org/who-we-are/epsrc/our-policies-and-standards/framework-for-responsible-innovation/ Accessed: 12 December 2024.

[55] UNESCO. 2024. AI Competency Framework for Students. https://unesdoc.unesco.org/ark:/48223/pf0000391105

[56] Feng Wang and Michael J. Hannafin. 2005. Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development* 53, 4 (December 2005), 5–23. https://doi.org/10.1007/BF02504682

[57] James Wexler, Mahima Pushkarna, and Marc A. Weksler et al. 2019. The What-If Tool: Code-free probing of machine learning models. PAIR - People + AI Research, Google. https://pair-code.github.io/what-if-tool/ Accessed: April 2025.

[58] Chunpeng Zhai, Santoso Wibowo, and Lily D. Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments* 11, 1 (2024), 28. https://doi.org/10.1186/s40561-024-00316-7