Making Bias Explorable for Metacognitive Scaffolding and Shared Regulation in Learning

Chaeyeon Lim
info@debiasme.com
DeBiasMe
South Korea & United Kingdom

Abstract

Generative AI in education creates a multi-agent environment where teachers, students, and AI systems collectively regulate learninging, yet most frameworks for AI literacy and tools for thought (Tf T) treat human-AI interaction as a dyadic cognitive exchange. This paper proposes collaborative bias mapping as a shared regulatory artifact: a representational structure that distributes metacognitive monitoring and control across teacher, student, and AI agents, while sustaining the relational conditions under which shared regulation becomes possible. Illustrative case study employing DeBiasMe AI literacy workshop with ten South Korean K-12 teachers shows how bias mapping could scaffold metacognitive engagement and relational practice. This proposal reframes bias not as a problem to be individually detected, but as a site of collective metacognitive engagement: explorable, negotiable, and generative of the kind of distributed sensemaking that AI-mediated education demands.

CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**; *Empirical studies in HCI* ; *Collaborative and social computing design and evaluation methods*; • **Applied computing → Education**; • **Social and professional topics → Computing literacy**.

Keywords

AI literacy, metacognition, shared regulation of learning, bias, multi-agent education, relational agency, tools for thought

ACM Reference Format:
Chaeyeon Lim. 2026. Making Bias Explorable for Metacognitive Scaffolding and Shared Regulation in Learning. *In Proceedings of Workshop on Tools for Thought at CHI 2026 (Tools for Thought at CHI 2026).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

As Generative AI (GenAI) tools become embedded in classrooms, they transform not only individual workflows but the regulatory process through which teachers and students jointly engage learning. Teachers use AI to generate materials, assess student work, and provide feedback; students use AI for inquiry, problem-solving, and content creation; AI systems mediate these activities with inherent biases and affordances. This creates a multi-agent ecosystem where the regulation of learning, traditionally understood as operating within individuals (self-regulation), between peers (co-regulation), or across groups (shared regulation) [5], must now account for the presence of computational systems that, through optimization for pattern-matching and measurable outputs, restructure the metacognitive demands placed on human agents in ways that neither teachers nor students have been prepared to engage.

This paper proposes the DeBiasMe workshop with collaborative bias mapping exercise, the identification and transformation of biases across stages of the AI life cycle, as a shared regulatory artifact: a representational structure that distributes metacognitive monitoring across teacher, student, and AI agents while sustaining the relational conditions under which shared regulation becomes possible.

This contribution builds on, and extends, existing frameworks in three directions. Current TfT research on cognitive augmenttation [1], metacognitive scaffolding [18], and the protection of human thought [15] has made significant progress at the level of dyadic human-AI interaction: how a single user can think more critically with AI, or how AI can scaffold one person's sensemaking. This paper extends this toward distributed and relational registers, drawing on Socially Shared Regulation of Learning (SSRL) [5], Human-AI Shared Regulation in Learning (HASRL) [2, 9], dis- tributed metacognition [8], and Edwards' relational agency [3]. As part of the AI literacy initiative, this extends the previously pro- posed bi-directional metacognitive framework [14], which mapped metacognitive demands at the input and output stages of human-AI interaction, into a multi-agent model in which those demands are distributed across and negotiated among human and AI agents.

2 Theoretical Foundations

2.1 Metacognitive Engagement and Shared Regulation in Human-AI Interactions

Our prior work [14] proposed a bi-directional metacognitive framework for AI literacy, identifying metacognitive demands for bias detection and mitigation at input and output stages of human-AI interactions. However, the use of AI in educational settings is rarely dyadic. Socially Shared Regulation of Learning (SSRL) describes how groups collectively plan, monitor, and adapt cognitive, motivational, and emotional processes toward shared goals in learning [5]. The emerging Human-AI Shared Regulation in Learning (HASRL) framework extends this to human-AI contexts where regulatory

Author's Contact Information: Chaeyeon Lim, info@debiasme.com, DeBiasMe

ACM 1557-7368/2026/3-ART111
https://doi.org/10.1145/nnnnnnn.nnnnnnn

processes are shared and coordinated across human and AI agent(s) [2, 9].

This multi-agent configuration requires a shift from individual to distributed metacognitive engagement. Following Iiskala et al. [8], metacognitive monitoring and regulation can be understood as processes that can be distributed across group members where the accuracy of one agent's metacognitive calibration (the correspondence between confidence and actual knowledge [11]) depends on and is shaped by other agents' regulatory contributions.

In AI-mediated education, this distribution becomes more complex: teachers must calibrate not only their own domain knowledge but their assessment of AI reliability; AI systems contribute computational processes such as confidence scoring, consistency checking, and self-correction routines that occupy a structural position analogous to metacognitive monitoring within the multi-agent regulatory system; and students must monitor not only their comprehension but the degree to which AI scaffolding is supporting their learning with the felt sense of knowing, uncertainty, and surprise, that grounds metacognitive experience and drives adaptive regulation in learning [6, 17, 18].

2.2 Relational Approach to Regulatory Challenge

The performance-learning tradeoff [16] emerges distinctly in multi-agent contexts in three different levels. First, it operates through metacognitive process when planning, monitoring, and reflection processes that support sensemaking can be bypassed [18]. Second, it operates with motivational factors: the effort required to regulate frustration, uncertainty, and confusion is both costly and valuable for achievement [7]. Third, it operates through relational channels: the social process of conversation and exploration that constitutes learning [17] is restructured when AI mediates the exchange. Tf T designed for multi-agent educational contexts must navigate all three channels simultaneously.

Edwards' [3] concept of relational agency makes this requirement explicit: it is the capacity to work with others to expand one's interpretations of a problem and the resources available to address it. Relational agency is not a property of individuals but of interactions, but it emerges when agents can align sufficiently on what matters while maintaining their distinct expertise. Edwards further proposes relational expertise as knowing how to work alongside others who bring different knowledge and commitments [4]. In the teacher-student-AI interaction, relational expertise requires understanding not just what AI can do, but what kind of agent it is and how its capacities differ from human judgment. This relational grounding matters for TfT design because bias mapping, as this paper proposes it, is not a cognitive exercise in identifying statistical patterns, but a relational practice of making visible each agent's interpretive framework, negotiating where expertise and authority reside, and coordinating regulatory action across agents whose goals may be asymmetric.

3 The DeBiasMe Workshop: Design and Framework

3.1 Designing for Distributed Metacognitive Engagement

This paper proposes a framework in which three agents, Teacher, Student, and AI, engage in shared regulatory processes distributed across four stages of the AI life cycle. Each agent contributes distinct metacognitive functions that, when coordinated, constitute shared regulation. Shared regulation emerges when these distributed functions coordinate: the teacher's regulatory demands create conditions under which the student's metacognitive engagement can develop, while AI's computational monitoring provides data that both teacher and student can interpret, verify, and contest. The DeBiasMe workshop is designed to make this distributed engagement visible and negotiable. Rather than presenting bias as a technical property to be individually detected, the workshop structures collaborative scenario analysis so that participants must reason across agent roles: whose monitoring function was active, where it failed, and what shared regulatory response would address it. Bias map- ping provides the shared representational space through which this coordination becomes visible and externalizes the distributed nature of regulatory demands.

3.2 Four Life Cycle Stages as Regulatory Sites

The DeBiasMe workshop organizes collaborative bias mapping across four stages of the AI life cycle (0: Input / Prompting, 1: Data Collection & Processing, 2: Model Training, 3: Output Interpretation), each constituting a distinct regulatory site with characteristic bias types and distributed metacognitive demands (Table 1). The life cycle structure presents bias as stage-specific rather than generic, creating the conditions for second-order monitoring: teachers and students must reason not only about what bias occurred but *where* in the AI life cycle it entered, and *how* it propagates across stages.

At the *input/prompting stage*, the workshop surfaces how prompt construction is a site of human cognitive bias, not neutral instruction. Participants examine how their own framing assumptions constrain what AI can produce, and what accountability this creates for prompt authors. At the *data collection & processing stage*, collaborative scenarios ask participants to identify where AI metrics reduce unmeasurable educational constructs to measurable proxies, and what interpretive authority this demands from human agents. At the *model training stage*, the workshop makes visible how training data encodes historical patterns of inclusion and exclusion, one that requires collective advocacy beyond individual awareness. At the *output interpretation stage*, participants collaboratively establish norms for when and how to contest AI outputs.

3.3 Risk-to-Competency Transformation as Design Mechanism

The DeBiasMe workshop's central design mechanism is the structured transformation of bias-related risks into metacognitive competencies. This transformation is deliberately collaborative: participants negotiate reframing in pairs and groups, contesting interpretations and assigning regulatory responsibilities, rather than working through pre-determined answers. Bias projection becomes

Table 1: Metacognitive Engagement and Regulatory Actions Across AI Life Cycle Stages

Stage	Bias Type	Teacher Regulatory Demand	Student Regulatory Demand
0	Human Cognitive Bias	Recognizing how one's evaluative frameworks and pedagogical assumptions constrain AI outputs	Recognizing how question framing limits exploratory scope
1	Data Bias	Identifying gaps between quantifiable AI metrics and educational values	Recognizing how AI feedback encodes proxies for measurable constructs
2	AI Bias	Contextualizing AI outputs within social, historical, and cultural values	Recognizing when AI feedback reproduces patterns of inclusion/exclusion in training data
3	Human Cognitive Bias	Designing structured opportunities for verification, challenge, and productive contestation of AI outputs	Maintaining metacognitive calibration and Emotional autonomy from AI validation

critical verification skill; measurement reduction becomes educational recontextualization; algorithmic opacity becomes a demand for explainability; over-reliance becomes human agency recovery. This mechanism is designed to (1) make the metacognitive work of bias awareness concrete and actionable and (2) build the interpretive alignment and trust on which shared regulation depends: participants are learning about bias to practice the shared regulatory engagement the framework requires.

4 Illustrative Case Study: DeBiasMe K-12 Teacher AI Literacy Workshop

4.1 Context

The case study employed a structured 2-hour AI literacy workshop with ten K-12 teachers (diverse subject areas) in Seoul, South Korea. The DeBiasMe workshop engages participants with AI bias concepts across the four life cycle stages through collaborative scenario analysis, using printed bias mapping materials that make the distributed nature of regulatory demands visible and collectively negotiable (Figure 1). This illustrative case study is part of the ongoing pilots and will inform empirical validation of the design framework.

4.2 Design in Practice

4.2.1 Collaborative Bias Mapping Across AI Life Cycle. The life cycle-stage structure produced the cross-stage metacognitive monitoring: participants traced how bias at the prompting stage propagated through to output interpretation, reasoning about compound regulatory failures rather than isolated instances.

4.2.2 Risk-to-competency transformation. The workshop's transformation mechanism invited teachers to generate competency reframing through negotiation and contestation rather, arriving at context-specific strategies that required the interpretive contributions of multiple agents.

4.2.3 Designing Relational Practices. The workshop format facilitated teachers to expand their interpretations of AI adoption challenges by working alongside others who bring different subject expertise and pedagogical commitments.

Figure 1: Teachers collaboratively mapping bias types and risk-to-competency transformation factors across the four AI life cycle stages using printed workshop materials. The physical arrangement of cards on life cycle stage made the distributed nature of regulatory demands visible and supported collective negotiation of where biases interact across stages.

5 Discussion and Future Work

5.1 Bias as Explorable, Negotiable, and Generative

The central theoretical contribution of this paper is the reconceptualization of collaborative bias mapping as a shared regulatory artifact, a representational structure that distributes metacognitive monitoring across teacher, student, and AI agents while sustaining the relational conditions under which shared regulation becomes possible. This reframes what bias means in AI-mediated educational contexts: not a technical property to be individually detected, but a site of collective metacognitive engagement, negotiable across agents with structurally different relationships to knowledge, ac- countability, and learning.

This contribution advances Tf T research in two directions. First, it shifts the unit of design from the individual user to the distributed

regulatory system, a shift that existing frameworks for metacognitive scaffolding and cognitive augmentation have not fully made, tending instead to treat human-AI interaction as dyadic even when the educational context is inherently multi-agent. Second, it foregrounds the relational conditions of shared regulation as a design concern, positioning relational scaffolding as a design function coordinate with cognitive scaffolding rather than secondary to it.

Ongoing DeBiasMe initiative and parallel pilots in Seoul (South Korea) and London (United Kingdom) will seek empirical validation of the proposed framework in both educational and professional contexts, testing whether this framework can generalize to other multi-agent contexts where regulatory asymmetry and AI-mediated oversight matter in practice.

Acknowledgments

I would like to thank University College London Edtech Labs and Learningspark for their support.

References

[1] Björklund, C. et al. 2020. Designing a tool for thought. Learning, Culture and Social Interaction, 27, 100443.

[2] Edwards, C. et al. 2024. Human-AI shared regulation in learning. British Journal of Educational Technology, 55(4), 1272-1290.

[3] Edwards, A. 2005. Relational agency: Learning to be a resourceful practitioner. International Journal of Educational Research, 43(3), 168-182.

[4] Edwards, A. 2010. Being an Expert Professional Practitioner: The Relational Turn in Expertise. Springer.

[5] Hadwin, A., Järvelä, S., & Miller, M. 2017. Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In Handbook of Self-Regulation of Learning and Performance (2nd ed., pp. 83-106). Routledge.

[6] Efklides, A. 2011. Interactions of metacognition with motivation and affect in self-regulated learning. Educational Psychology Review, 23(1), 6-25.

[7] Inzlicht, M., Shenhav, A., & Olivola, C. Y. 2018. The effort paradox: Effort is both costly and valued. Trends in Cognitive Sciences, 22(4), 337-349.

[8] Iiskala, T. et al. 2011. Socially shared metacognition of dyads of pupils in collaborative mathematical problem-solving processes. Learning and Instruction, 21(3), 379-393.

[9] Järvelä, S. et al. 2023. Human and artificial intelligence collaboration for socially shared regulation in learning. British Journal of Educational Technology, 54(5), 1057-1076.

[10] Kapur, M. 2016. Examining productive failure, productive success, and deliberate practice. Instructional Science, 44(2), 113-138.

[11] Koriat, A. 2007. Metacognition and consciousness. In The Cambridge Handbook of Consciousness (pp. 289-325). Cambridge University Press.

[12] Noddings, N. 2013. Caring: A Relational Approach to Ethics and Moral Education (2nd ed.). University of California Press.

[13] Long, D. & Magerko, B. 2020. What is AI literacy? Competencies and design considerations. In Proceedings of CHI'20 (pp. 1-16). ACM.

[14] Lim, C. 2025. DeBiasMe: Metacognitive AI literacy through interactive bias mapping. In Extended Abstracts of CHI'25. ACM.

[15] Guan, Z. K. et al. 2025. Tools for thought: Understanding, protecting, and augmenting human cognition with generative AI. In Extended Abstracts of CHI'25.

[16] Sarkar, A. 2024. AI should challenge, not obey. Communications of the ACM, 67(11), 39-41.

[17] Sharples, M. 2023. New AI Does Not Mean New Pedagogy. Routledge.

[18] Tankelevitch, L. et al. 2025. Metacognitive demands of human-AI interaction. In Proceedings of CHI'25. ACM.