

The Metacognition Paradox

Dr. Tessa Forshaw, Next Level Lab, Harvard University

Main theme(s): Design Strategy

Target domain(s): Education, AI-assisted cognitive work

Cognitive 'target(s)': Metacognition

Type of contribution & main idea

(1) Critique of how current convergent AI design undermines metacognition, and (2) proposal of a design framework (seven principles) for AI tools that restore the conditions metacognition requires.

The AI tool's key characteristics

AI tools designed to converge on answers unintentionally dismantle the conditions that make metacognition possible: explicit prompts, dedicated time, occasions that trigger it, and emotional signals that sustain it. This is the Metacognition Paradox. Working with AI demands more metacognition of users not less. Users must monitor not just their own thinking but the AI's reasoning. This is 'double metacognition': the requirement to be metacognitive about both one's own thinking and the AI's reasoning under conditions in which AI tools simultaneously erode the supports that typically foster metacognitive regulation.

This paper reframes metacognition in AI-mediated work as a design problem rather than solely an individual cognitive one. Seven design principles address the paradox by restoring the four enabling conditions that current AI tends to remove.

Seven design principles:

1. Elicit Before Generating
2. Show the Reasoning
3. Preserve the Mesy Middle
4. Pause at Decision Points
5. Surface Uncertainty
6. Resist Premature Closure
7. Make Help-Seeking Metacognitive

Note: This is a theoretical contribution. The principles are derived from existing metacognition research and emerging evidence on AI's cognitive effects. They have not yet been tested as a unified framework.

What you would like to discuss

Which principles matter most, how they should adapt across user populations and expertise levels, and how to measure metacognitive outcomes in practice.

Interaction & outcomes

How people use and interact with the AI tool for a specific activity (first column) and what outcomes this provides, enables, or leads to (second and third column).

Interaction design/usage strategy

Each principle is designed to support eliciting the enabling conditions that many current AI tools undermine. Elicit Before Generating and Pause at Decision Points restore explicit prompting. Preserve the Mesy Middle restores dedicated time and space. Show the Reasoning and Surface Uncertainty create occasions for noticing. Resist Premature Closure restores emotional signals as metacognitive data.

Objective value

The 'objective' quality of the process and the outcomes - i.e. what makes them 'good' or desirable (e.g. critical thinking, understanding, new insights, learning).

The interaction pattern sustains metacognitive engagement by creating occasions for metacognition that users typically fail to notice on their own (Langr, 1989; Perkins et al., 1993). Preserving intermediary work maintains the cognitive involvement that drives stronger outcomes in quality and originality (Hargrove & Niefeld, 2015). This could be observed through how often users modify AI output or catch flawed assumptions.

Perceived value & UX

How the process of using the AI tool and its outcomes is experienced by users (e.g. how they experience the cognitive work they are doing).

Users accustomed to fast AI answers will likely find this agent more demanding. Users expecting AI to reduce effort may find the prompts frustrating. People who already use metacognitive strategies will probably take to it naturally. This could be assessed through how users react to the pauses, whether they engage or skip past them.

Intermediary task outcomes

The intermediary outcomes are where double metacognition shows up. Users have their own perspective on record before they see AI output, which creates an independent reference point for evaluation. Visible reasoning gives users something to evaluate beyond the final answer. Structured pauses create space for monitoring and adjustment during the work, not just before or after it. These are the moments where awareness, evaluation, and control happen in real time.

The intermediary outcomes are valuable because metacognition operates during the work, not just before or after (Forshaw, 2025). Visible AI reasoning gives users something to be metacognitive about (Tankelevitch et al., 2024). Friction restores emotional signals like confusion and doubt that trigger monitoring (Jia et al., 2019; Kavousi et al., 2020b).

Users will appreciate the intermediary outcomes when they can see the friction actually helped them. The challenge is that metacognition and cognition compete for the same attentional resources (Cuxinello & Gruber, 2022), and under time pressure people accept whatever looks okay (Butler & Roberto, 2018). Low-stakes tasks are where this friction will feel least worthwhile. This could be assessed by tracking engagement with prompts across different task types.

Final task outcomes

Task outputs would reflect real evaluation rather than acceptance without consideration. The key question is whether metacognitive accuracy improves, meaning whether people's sense of how well they are doing starts to match how well they are actually doing.

AI currently makes people perform better while making them worse at knowing how well they are doing (Fernandes et al., 2026). What makes the final outcome of using a tool built with these principles good is whether that gap closes and performance and awareness match up.

Users may not notice the difference this agent makes. The confidence-accuracy gap (Fernandes et al., 2026) means people already misjudge their performance with AI, and knowing more about AI makes this worse, not better. The contribution would most likely become visible through comparison over time, or through external evaluation that reveals quality differences users do not detect themselves.

How do people's goals interact with the AI tool's goals? Is there a tension?

Since this is a design framework, not a specific tool, the tension described here is theorised from the literature. It would need to be validated empirically for any specific implementation.

Users often want the AI to produce a finished result quickly, e.g. finishing an assignment before a deadline and the framework's principles add time and friction that appears to work against these goals.

But they also want things like to actually learn, develop skill, and do good work, and the framework's goals align with these.

This is a central tension the paper discusses. Users optimising for immediate productivity may resist the features that support long-term metacognitive outcomes. The impulse to smooth away friction is itself the convergent tendency that undermines metacognition.

How do you expect people to continue using this AI tool?

People who are building AI tools and want them to support thinking, not just produce answers. Educators setting up AI for their students where the goal is learning, not just getting the assignment done. Organizations figuring out guidelines for how their teams should work with AI, especially when the stakes are high. Basically, anyone making decisions about how an AI tool behaves - whether they're writing the code or writing the rules around it.

How to proceed with this work/idea?

I will use the workshop feedback to refine prioritisation and sequencing of principles. Then I want to prototype with 2-3 principles and run user studies measuring both the impact on metacognition and on performance

What would you like to take away from the workshop?

Feedback on the framework, design insights from adjacent work, and collaborators interested in prototyping and testing the principles across domains. I would also like to build a community of people interested in similar questions and topics - bridging HCI and the learning sciences.

Key references (e.g. of main theories, empirical evidence, measurement methods etc.)

- [1] Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmitt, A., Kosch, T., Shen, C., & Welch, R. (2026). AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior*, 175, 108779.
- [2] Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new arena of cognitive-development inquiry. *American Psychologist*, 34(10), 906-911.
- [3] Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- [4] Sun, S., Li, Z. A., Foo, M.-D., Zhou, J., & Lu, J. G. (2025). How and for whom using generative AI affects creativity: A field experiment. *Journal of Applied Psychology*, 110(12), 1561-1572.
- [5] Tankelevitch, L., Kavenigi, V., Simbute, A., Scott, A. E., Sarkar, A., & Sellen, A. J. (2024). The metacognitive demands and opportunities of generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24), 4634.