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

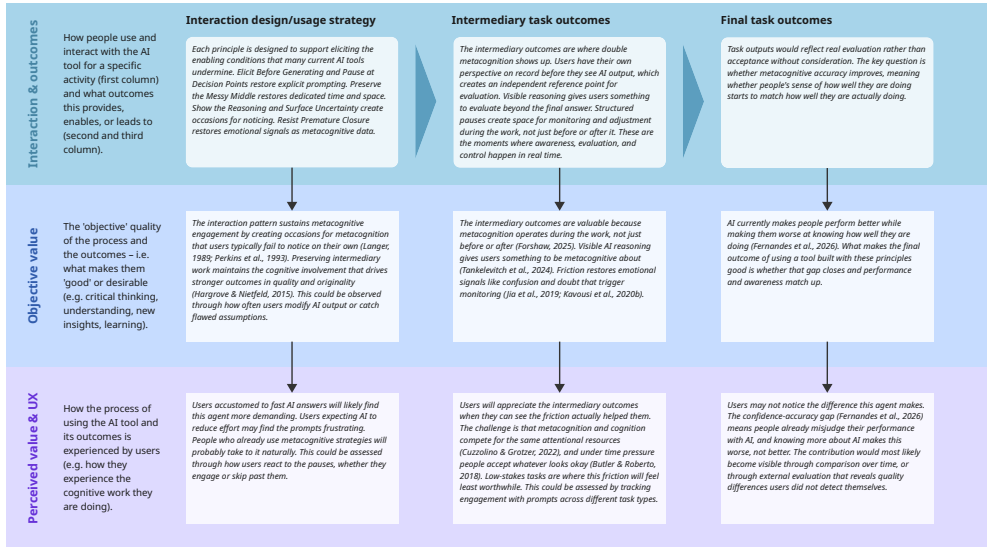
AI tools designed to converge on answers unintentionally dismantle the conditions that make metacognition possible: explicit prompts, dedicated time, occasions that trigger it, and emotional signals that sustain it. This is the Metacognition Paradox. Working with AI demands more metacognition od users 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 Messy 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.



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.

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.

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., Schmidt, A., Kosch, T., Shen, C., & Welsch, 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 area of cognitive-developmental 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), 1561–1573.
[5] Tankelevitch, L., Kewenig, V., Simkute, 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). ACM.

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