The Metacognitive Paradox of AI-Assisted Creativity: A Theoretical Extension

Jonathan H. Westover *

Posted Date: 2 March 2026

doi: 10.20944/preprints202603.0121.v1

Keywords: generative AI; creativity; metacognition; large language models; cognitive offloading; human-AI interaction; automation; field experiment

 doi:10.20944/preprints202603.0121.v1

Article

The Metacognitive Paradox of AI-Assisted Creativity: A Theoretical Extension

Jonathan H. Westover

Nexus Institute for Work & AI, Catalyst Center for Work Innovation; jon.westover@gmail.com

Abstract

The rapid integration of large language models (LLMs) into organizational workflows raises fundamental questions about the long-term effects of AI assistance on human creative capabilities. This article provides a comprehensive theoretical extension of Sun et al.'s (2025) field experiment, which demonstrated that LLM assistance enhances creative output during use but diminishes subsequent independent creativity—particularly for individuals with lower metacognitive ability. We develop the metacognitive paradox framework to explain this phenomenon: AI tools designed to augment human creativity may inadvertently suppress the very cognitive processes that sustain independent creative capacity. Drawing on metacognition theory, cognitive offloading research, automation and human factors literatures, and the componential model of creativity, we articulate the mechanisms through which LLM assistance affects creative cognition, specify temporal dynamics and boundary conditions, and generate testable propositions for future research. We explicitly compare our framework against alternative theoretical explanations—including cognitive load theory, motivational accounts, and expertise development perspectives—and provide methodological guidance for testing our propositions. Our analysis reveals that the relationship between AI assistance and human creativity is neither uniformly beneficial nor detrimental but contingent upon individual metacognitive capabilities, patterns of tool use, task characteristics, and organizational context. We conclude with implications for theory, research methodology, organizational practice, and the ethical dimensions of AI-augmented work.

Keywords: generative AI; creativity; metacognition; large language models; cognitive offloading; human-AI interaction; automation; field experiment

---

Introduction

The emergence of generative artificial intelligence, particularly large language models (LLMs), represents a watershed moment in the relationship between human cognition and technological augmentation. Organizations across industries have rapidly integrated these tools into knowledge work, driven by their demonstrated capacity to generate text, solve problems, and produce creative outputs that often rival human performance (Zhao et al., 2023). This technological shift has prompted fundamental questions about how AI assistance affects not merely immediate task performance but the underlying cognitive capabilities that enable human creativity and innovation.

Sun, Li, Foo, Zhou, and Lu (2025) provide crucial empirical evidence addressing these questions through a rigorously designed field experiment. Their study revealed a striking pattern: while LLM assistance enhanced creative performance during use, it subsequently diminished independent creativity when the AI tool was no longer available. Moreover, this effect was moderated by individual differences in metacognitive ability, with individuals possessing lower metacognitive skills experiencing the most pronounced decrements in subsequent creative performance. These findings carry profound implications for how organizations deploy AI tools and how individuals develop and maintain creative capabilities in an increasingly AI-augmented workplace.

The present article offers an extensive theoretical analysis and extension of Sun et al.'s (2025) findings. We develop what we term the *metacognitive paradox framework*—the phenomenon whereby

 doi:10.20944/preprints202603.0121.v1

tools designed to enhance human creativity may inadvertently undermine the cognitive processes that sustain independent creative capacity. This paradox emerges because the very efficiency gains that make AI assistance attractive—reduced cognitive effort, accelerated idea generation, diminished uncertainty—simultaneously reduce opportunities for the metacognitive engagement that maintains and develops creative skills over time.

Our analysis proceeds in several stages. We first situate Sun et al.'s study within the broader literature on creativity, technology, and cognition, identifying the theoretical gap their work addresses. We then develop our metacognitive framework in detail, articulating the mechanisms through which LLM assistance may both enhance and erode creative capabilities. Crucially, we compare this framework against alternative theoretical explanations, including cognitive load theory, motivational accounts, and expertise development perspectives, to demonstrate its unique explanatory value. We propose specific boundary conditions and moderating factors that shape these dynamics, with particular attention to temporal considerations. We provide methodological guidance for testing our propositions and address operationalization challenges. Finally, we discuss implications for organizational practice, research methodology, and the ethical dimensions of AI-augmented creative work.

Theoretical Background

Creativity in Organizational Contexts

Creativity—the generation of ideas, solutions, or products that are both novel and useful—constitutes a cornerstone of organizational adaptation and competitive advantage (Amabile, 1988; Zhou & Shalley, 2003). The componential model of creativity identifies three essential elements: domain-relevant skills encompassing technical knowledge and expertise, creativity-relevant processes including cognitive styles and heuristics for novel ideation, and intrinsic task motivation reflecting genuine interest and engagement (Amabile & Pratt, 2016). This tripartite framework has proven remarkably durable, providing a foundation for understanding how individual, social, and contextual factors combine to influence creative outcomes.

Research on creativity in organizations has increasingly recognized its dynamic and contextual nature (Anderson et al., 2014). Creative performance fluctuates over time in response to changing demands, resources, and psychological states. The journey from initial idea generation through development, championing, and implementation involves distinct phases with different requirements and challenges (Perry-Smith & Mannucci, 2017). This temporal perspective proves essential for understanding how technological interventions might differentially affect various stages of the creative process.

The introduction of AI tools into creative workflows represents both an opportunity and a challenge for established creativity frameworks. On one hand, LLMs can provide domain knowledge, suggest novel combinations, and reduce the cognitive burden of routine aspects of creative tasks. On the other hand, these same capabilities may alter the fundamental nature of creative cognition in ways that existing theories have not fully anticipated. Understanding these dynamics requires integration across multiple theoretical perspectives.

Metacognition and Creative Performance

Metacognition—cognition about cognition—encompasses the knowledge, monitoring, and regulation of one's own cognitive processes (Flavell, 1979). In creative contexts, metacognition manifests as awareness of one's creative strengths and limitations, monitoring of idea quality during generation, and strategic regulation of creative approaches based on task demands and feedback (Davidson & Sternberg, 1998). These metacognitive processes enable individuals to navigate the uncertainty inherent in creative work, recognizing when familiar approaches prove insufficient and novel strategies become necessary.

Sun (2024) advanced a componential and functional framework for metacognition that identifies three core elements: metacognitive knowledge (understanding of cognitive processes and strategies), metacognitive monitoring (ongoing assessment of cognitive states and progress), and metacognitive

 doi:10.20944/preprints202603.0121.v1

control (strategic regulation based on monitoring outcomes). This framework illuminates how metacognition operates in complex work environments and suggests mechanisms through which external tools might interact with internal cognitive processes.

The relationship between metacognition and creativity operates through several pathways (Veenman et al., 2004). First, metacognitive monitoring enables recognition of impasses and fixation states that impede creative progress. Second, metacognitive knowledge provides access to strategies for overcoming these obstacles, such as analogical reasoning or deliberate incubation. Third, metacognitive control allows flexible adjustment of approaches based on ongoing assessment of progress. Together, these processes enable the adaptive cognition that characterizes effective creative performance.

Importantly, metacognitive skills develop through practice and experience (Winne & Nesbit, 2010). Repeated engagement with challenging cognitive tasks builds both the knowledge base and the monitoring capabilities that constitute metacognitive expertise. This developmental perspective raises important questions about how technological assistance might affect the acquisition and maintenance of metacognitive skills over time.

Cognitive Offloading and Extended Cognition

The concept of cognitive offloading provides a complementary lens for understanding human-technology interaction in creative work. Cognitive offloading refers to the use of external resources—including physical tools, environmental features, and other people—to reduce the cognitive demands of a task (Risko & Gilbert, 2016). This strategy proves adaptive when cognitive resources are limited or when external resources provide more reliable or efficient storage and processing.

The extended mind thesis (Clark & Chalmers, 1998) proposes that cognitive processes can genuinely extend beyond the boundaries of the individual brain to incorporate external tools and resources. From this perspective, a person using a notebook to store information or a calculator to perform computations is not merely supplementing cognition but extending it. The cognitive system, properly understood, includes the external resources that reliably contribute to cognitive performance.

However, the extended mind framework raises important questions about the locus of cognitive capabilities. If cognition extends to include external tools, what happens when those tools are removed? Does the individual retain the cognitive capabilities that were previously distributed across person and technology? These questions become particularly pressing when the external resource is an AI system capable of performing sophisticated cognitive operations that would otherwise require substantial mental effort.

Recent research suggests that cognitive offloading, while beneficial for immediate performance, may carry costs for long-term cognitive development. When individuals routinely rely on external resources for cognitive tasks, they may experience reduced development of internal capabilities (Risko & Gilbert, 2016). This pattern suggests potential tensions between short-term performance optimization through offloading and long-term capability development through effortful internal processing.

Automation, Skill Degradation, and Human-AI Interaction

The human factors and automation literatures have extensively studied how technological assistance affects human skill maintenance and performance—research directly relevant to understanding AI's effects on creativity. Parasuraman and Riley (1997) identified patterns of automation use, misuse, disuse, and abuse that illuminate the complex dynamics of human-technology interaction. Their framework highlights how operators may become overly reliant on automated systems (misuse) or fail to use automation appropriately (disuse), with consequences for both performance and safety.

Particularly relevant is research on skill degradation in automated systems. Endsley and Kiris (1995) documented the "out-of-the-loop" performance problem, whereby operators who rely on automation lose situational awareness and struggle to resume manual control when needed. This research, primarily conducted in aviation and industrial control contexts, demonstrates that skills can

 doi:10.20944/preprints202603.0121.v1

atrophy relatively quickly when not regularly exercised, and that regaining proficiency after periods of automation-supported performance requires substantial retraining.

The concept of automation complacency (Parasuraman & Manzey, 2010) describes the tendency for operators to reduce their monitoring and vigilance when automated systems are functioning. This reduced engagement can lead to failures in detecting automation errors and degraded ability to intervene when necessary. Applied to creative work, automation complacency might manifest as reduced critical evaluation of AI-generated outputs and diminished engagement with the creative process itself.

Sheridan and Verplank's (1978) taxonomy of automation levels provides a framework for understanding different types of AI assistance. At lower levels, automation merely offers suggestions while humans retain decision authority. At higher levels, automation executes decisions with limited human oversight. The metacognitive implications likely differ across these levels, with higher automation potentially producing greater disengagement and more rapid skill degradation.

Emerging frameworks on human-AI teaming (Seeber et al., 2020) conceptualize the relationship between humans and AI systems as collaborative rather than merely assistive. This perspective emphasizes mutual adaptation, shared mental models, and complementary capabilities. However, it also raises questions about how human capabilities develop and persist when substantial cognitive work is delegated to AI teammates.

The Sun et al. (2025) Field Experiment

Sun et al.'s (2025) study represents a significant methodological advance in understanding AI effects on creativity. Conducted with 656 employees of an online education company, the experiment employed random assignment to LLM assistance versus control conditions with a design that enabled assessment of both concurrent and subsequent effects on creative performance. Participants completed creative writing tasks that were evaluated by both supervisors and external raters blind to condition.

The results revealed a nuanced pattern of effects. Individuals assigned to use LLM assistance demonstrated enhanced creative performance on the initial task compared to those working without AI support. However, when subsequently asked to complete a creative task without LLM access, those who had previously used AI assistance showed diminished creative performance relative to their baseline capabilities and to the control group's subsequent performance.

Critically, Sun et al. (2025) found that metacognitive ability moderated these effects. Individuals with higher metacognitive ability maintained stronger subsequent creative performance after LLM use, while those with lower metacognitive ability experienced more pronounced decrements. The authors measured metacognitive ability using an adapted scale assessing participants' tendencies to plan, monitor, and evaluate their cognitive processes during complex tasks.

Mediation analyses suggested that reduced metacognitive engagement during LLM-assisted work partially explained the subsequent creativity decrements. When individuals relied on AI assistance, they reported less active monitoring and regulation of their own thinking processes. This reduced metacognitive engagement appeared to have carry-over effects, diminishing subsequent independent creative performance.

The Metacognitive Paradox Framework

Defining the Paradox

We introduce the term *metacognitive paradox* to describe the phenomenon whereby AI tools designed to enhance human creativity may inadvertently suppress the cognitive processes that sustain independent creative capacity. The paradox emerges from a fundamental tension: the efficiency gains that make AI assistance attractive simultaneously reduce opportunities for the metacognitive engagement that maintains and develops creative skills.

The paradoxical quality of this phenomenon operates at multiple levels. At the individual level, rational behavior (using available tools to enhance performance) produces seemingly irrational outcomes (diminished capability over time). At the organizational level, investments in AI tools to

 doi:10.20944/preprints202603.0121.v1

boost creativity may paradoxically undermine the human creative capital that organizations depend upon. At the societal level, technologies developed to augment human capabilities may, if not carefully managed, erode those very capabilities.

This paradox differs from simple skill atrophy or disuse. The mechanism is not merely that creative skills deteriorate from lack of practice. Rather, the metacognitive processes that enable flexible, adaptive creativity—processes of monitoring, evaluation, and strategic adjustment—become externalized to the AI system. When the AI is removed, individuals must reconstitute these processes, a task for which they may be less prepared than if they had maintained continuous metacognitive engagement.

Articulating the Core Mechanism

The core mechanism underlying the metacognitive paradox involves the externalization of metacognitive functions to AI systems. When individuals engage in creative work without AI assistance, they must continuously monitor their progress, evaluate the quality of emerging ideas, recognize obstacles and impasses, and adjust their strategies accordingly. These metacognitive operations are cognitively demanding but serve essential functions: they build metacognitive knowledge, refine monitoring capabilities, and develop flexible control strategies.

When LLM assistance becomes available, individuals can offload substantial portions of these metacognitive operations to the AI system. Rather than monitoring their own idea generation, they can evaluate AI-generated suggestions. Rather than recognizing impasses, they can query the AI for additional options. Rather than strategically adjusting their approach, they can refine their prompts or accept AI recommendations. This offloading reduces immediate cognitive burden while maintaining or enhancing task performance.

However, this externalization comes at a cost. The metacognitive processes that would normally be exercised and refined during creative work are instead bypassed. Over time—potentially even over relatively short periods of consistent AI use—the individual's metacognitive engagement becomes attenuated. When subsequently required to work without AI assistance, the individual must reconstitute these metacognitive processes, often finding that their monitoring is less acute, their strategic knowledge less accessible, and their control less flexible than before AI exposure.

We distinguish between *metacognitive suppression* (temporary reduction in metacognitive engagement during AI-assisted work) and *metacognitive erosion* (more persistent changes in metacognitive capability resulting from extended or intensive AI use). The Sun et al. (2025) findings, observed over approximately eight days, likely reflect a combination of these processes. Suppression may occur rapidly and recover quickly once AI assistance is removed, while erosion involves more gradual changes that require more extensive intervention to reverse.

This distinction has important implications for intervention. If the observed effects primarily reflect suppression, brief periods of unassisted work may be sufficient to restore metacognitive engagement. If they reflect erosion, more intensive remediation—including explicit metacognitive training—may be necessary.

Comparison with Alternative Theoretical Explanations

To establish the unique explanatory value of the metacognitive paradox framework, we must consider alternative theoretical accounts that might explain Sun et al.'s (2025) findings.

Cognitive Load Theory

Cognitive load theory (Sweller, 1988) suggests that learning and performance depend on the management of working memory demands. From this perspective, LLM assistance might enhance performance by reducing extraneous cognitive load, freeing resources for more productive cognitive operations. However, this theory also predicts that removing the AI assistance would simply return cognitive load to baseline levels, not produce decrements below baseline performance.

Cognitive load theory could potentially explain diminished subsequent performance if germane load (cognitive effort devoted to learning and skill development) was reduced during AI-assisted work. Without the effortful processing that builds schemas and procedural knowledge, individuals might fail to consolidate learning from the creative task. However, this explanation focuses on

 doi:10.20944/preprints202603.0121.v1

knowledge acquisition rather than the metacognitive monitoring and control processes that our framework emphasizes.

The metacognitive paradox framework provides a more specific mechanism than cognitive load theory: it identifies the particular cognitive processes (metacognitive monitoring and control) that are disrupted and explains why this disruption persists beyond the period of AI assistance. Cognitive load theory is compatible with our framework but does not fully capture the metacognitive dynamics we describe.

Motivational Explanations

Motivational theories offer another alternative explanation. Learned industriousness theory (Eisenberger, 1992) proposes that effortful processing builds generalized persistence and motivation for cognitive work. If LLM assistance reduces the effort required for creative tasks, it might diminish the learned industriousness that sustains subsequent creative effort.

Similarly, self-determination theory (Deci & Ryan, 2000) emphasizes autonomy, competence, and relatedness as fundamental psychological needs. AI assistance might satisfy or undermine these needs in complex ways. If individuals experience AI-assisted work as diminishing their sense of competence or autonomy, subsequent motivation for creative work might suffer.

We acknowledge that motivational mechanisms likely contribute to the observed effects. However, the Sun et al. (2025) findings specifically implicate metacognitive ability as a moderator, suggesting that cognitive rather than purely motivational processes are centrally involved. Individuals with high metacognitive ability, who presumably experienced similar motivational dynamics as their lower-metacognitive-ability counterparts, did not show the same decrements in subsequent performance. This pattern is more consistent with a metacognitive than a purely motivational explanation.

Expertise Development Perspectives

Expertise development theories (Ericsson et al., 1993) emphasize the role of deliberate practice in building expert performance. Deliberate practice involves focused effort on challenging tasks with immediate feedback, typically guided by coaches or mentors. From this perspective, AI assistance might undermine expertise development by reducing the deliberate practice that would otherwise occur during creative work.

This explanation has merit, particularly for understanding longer-term effects of AI assistance on creative skill development. However, the Sun et al. (2025) study observed effects over a relatively brief period (approximately eight days), suggesting that mechanisms beyond gradual skill development are involved. The metacognitive paradox framework accounts for these rapid effects by focusing on the immediate disengagement of metacognitive processes rather than the longer-term accumulation of expertise.

Summary of Theoretical Comparison

The metacognitive paradox framework offers several advantages over these alternative explanations:

1. *Specificity:* It identifies specific cognitive processes (metacognitive monitoring and control) that are affected, rather than invoking general mechanisms like cognitive load or motivation.
2. *Mechanism precision:* It explains how these processes become externalized to AI systems, not merely reduced or suppressed.
3. *Moderator explanation:* It accounts for why metacognitive ability moderates the effects, a pattern less easily explained by cognitive load or motivational theories.
4. *Temporal dynamics:* It distinguishes between suppression and erosion, providing a framework for understanding both rapid and gradual effects.
5. *Integration with AI research:* It connects to the broader literature on automation and skill degradation, providing a more comprehensive account of human-AI interaction.

We do not claim that the metacognitive paradox framework entirely supersedes these alternative explanations. Rather, we propose that metacognitive processes represent the primary mechanism,

 doi:10.20944/preprints202603.0121.v1

while cognitive load, motivational, and expertise development factors operate as complementary influences that may amplify or attenuate the core metacognitive dynamics.

Extending the Framework: Propositions and Boundary Conditions

Building on this theoretical foundation, we develop a series of propositions that extend Sun et al.'s (2025) findings and guide future research. These propositions address mechanisms, moderators, temporal dynamics, and contextual factors.

Metacognitive Processes as Mediating Mechanisms

The first set of propositions addresses the specific metacognitive processes through which LLM assistance affects subsequent creative performance.

Proposition 1a: Strategic monitoring—the ongoing assessment of progress toward creative goals—mediates the relationship between LLM assistance and subsequent creative performance. Specifically, reduced strategic monitoring during LLM-assisted work leads to diminished monitoring capabilities in subsequent unassisted work.

This proposition extends Sun et al.'s (2025) findings by specifying one component of the broader metacognitive mechanism. Strategic monitoring enables recognition of when current approaches are failing and new strategies are needed. When LLM assistance provides continuous suggestions and alternatives, the need for internal monitoring decreases, and this monitoring capability may become attenuated.

Proposition 1b: Metacognitive knowledge—understanding of effective creative strategies and when to deploy them—is less actively accessed and refined during LLM-assisted creative work, resulting in less available and less flexible strategic knowledge during subsequent unassisted work.

Creative work typically builds metacognitive knowledge through trial and error: attempting strategies, evaluating their effectiveness, and encoding this information for future use. LLM assistance may short-circuit this learning process by providing effective outputs without requiring the strategic exploration that builds metacognitive knowledge.

Proposition 1c: The relationship between LLM assistance intensity and subsequent creative performance follows a non-linear pattern, with moderate assistance potentially enhancing subsequent performance while intensive assistance diminishes it.

This proposition reflects the idea that some level of external scaffolding may support metacognitive development (cf. Wood et al., 1976), while excessive support may produce the disengagement effects we describe. The optimal level of assistance likely varies across individuals and contexts, a boundary condition we elaborate below.

Individual Differences in Vulnerability and Resilience

The second set of propositions addresses individual differences that moderate the effects of LLM assistance on creative performance.

Proposition 2a: Individuals with higher baseline metacognitive ability maintain more stable metacognitive engagement during LLM-assisted work and demonstrate greater resilience in subsequent unassisted creative performance.

This proposition replicates and extends Sun et al.'s (2025) core finding regarding metacognitive ability as a moderator. We propose that high-metacognitive-ability individuals not only possess stronger metacognitive skills but also maintain greater awareness of their own cognitive processes during AI-assisted work. This awareness may enable them to recognize when metacognitive disengagement is occurring and take corrective action.

Proposition 2b: Need for cognition—the tendency to engage in and enjoy effortful cognitive activity (Cacioppo & Petty, 1982)—moderates the effects of LLM assistance on subsequent creative performance. Individuals high in need for cognition experience smaller decrements in subsequent creativity following LLM use.

High need for cognition may lead individuals to maintain cognitive engagement even when external assistance is available, thereby preserving metacognitive functioning. These individuals may

 doi:10.20944/preprints202603.0121.v1

find purely receptive interaction with AI unsatisfying and actively seek opportunities for mental effort.

Proposition 2c: Domain expertise moderates the relationship between LLM assistance and subsequent creative performance. Experts experience smaller decrements than novices because their well-developed domain knowledge provides a stable foundation for metacognitive operations even when AI assistance is removed.

Experts possess rich, well-organized knowledge structures that support metacognitive monitoring and control (Chi et al., 1988). This cognitive infrastructure may be more resistant to disruption from AI assistance than the less developed knowledge structures of novices.

Temporal Dynamics of Metacognitive Effects

The third set of propositions addresses the temporal dynamics of LLM effects on creativity, distinguishing between immediate, short-term, and longer-term patterns.

Proposition 3a: Metacognitive suppression (temporary reduction in metacognitive engagement) occurs rapidly upon initiation of LLM assistance, while metacognitive erosion (more persistent changes in metacognitive capability) develops gradually with sustained or repeated AI use.

This proposition suggests that the effects observed in Sun et al.'s (2025) study may combine rapid suppression with incipient erosion. Distinguishing these processes has important implications for understanding recovery trajectories and designing interventions.

Proposition 3b: Recovery of metacognitive engagement following LLM use depends on the duration and intensity of prior AI assistance. Brief, moderate assistance permits rapid recovery, while extended, intensive assistance requires more prolonged periods of unassisted work or explicit metacognitive training for full recovery.

The dose-response relationship between AI assistance and metacognitive effects remains poorly specified. This proposition suggests that both duration and intensity matter, with their combined effect determining the difficulty of recovery.

Proposition 3c: Periodic interruption of LLM assistance with required unassisted work ("metacognitive maintenance intervals") can prevent or attenuate the erosion of metacognitive capabilities even during extended periods of AI tool availability.

This proposition has direct practical implications, suggesting that organizations can mitigate metacognitive erosion through structured alternation between assisted and unassisted work. The optimal frequency and duration of these intervals remains an empirical question.

Task and Context Characteristics

The fourth set of propositions addresses task and contextual factors that moderate the effects of LLM assistance.

Proposition 4a: Task complexity moderates the relationship between LLM assistance and subsequent creative performance. For complex creative tasks requiring substantial metacognitive engagement, the effects of AI assistance on subsequent performance are larger than for simpler tasks.

Complex tasks typically demand more metacognitive resources—more monitoring, more strategic adjustment, more recognition of impasses. Consequently, the externalization of these processes to AI systems may have more pronounced effects for complex than for simple tasks.

Proposition 4b: The type of AI assistance matters: generative assistance (AI produces complete solutions or products) produces larger metacognitive effects than evaluative assistance (AI assesses or critiques human-generated content).

Generative assistance may allow more complete externalization of metacognitive processes than evaluative assistance, which still requires the individual to produce content for evaluation. This proposition connects to Sheridan and Verplank's (1978) taxonomy of automation levels.

Proposition 4c: Organizational norms regarding AI use moderate individual-level metacognitive dynamics. In organizations that emphasize critical engagement with AI outputs and maintain expectations for independent creative capability, individuals experience smaller decrements in subsequent creativity following LLM use.

 doi:10.20944/preprints202603.0121.v1

Organizational context shapes how individuals approach AI tools. Organizations that frame AI as a complement to, rather than substitute for, human cognition may foster more reflective engagement that preserves metacognitive functioning.

Proposition 4d: Psychological safety (Edmondson, 1999) moderates the relationship between LLM assistance and subsequent creative performance. In psychologically safe environments, individuals are more willing to acknowledge uncertainty, seek feedback, and maintain metacognitive engagement even when AI assistance is available.

Psychological safety may encourage the kind of reflective practice that sustains metacognitive capability. When individuals feel safe to admit what they do not know and to seek help in developing their own capabilities, they may be more resistant to the complacency that can accompany AI assistance.

Creative Process Stages

The fifth proposition addresses how LLM effects may differ across stages of the creative process.

Proposition 5: The effects of LLM assistance on metacognitive engagement and subsequent creative performance differ across stages of the creative process. Assistance during early-stage ideation produces different effects than assistance during later-stage refinement and implementation, with early-stage assistance potentially having larger effects on the development of generative metacognitive skills.

Perry-Smith and Mannucci's (2017) model of the creative process identifies distinct stages with different requirements. Early-stage ideation may be particularly dependent on generative metacognitive processes—exploration, incubation, insight—while later stages involve more evaluative and refinement processes. LLM assistance that displaces generative processes may have different consequences than assistance that supports evaluative processes.

Methodological Considerations and Operationalization

Addressing the peer reviewers' concerns about methodology and operationalization, we provide guidance for future research testing our framework.

Research Design Considerations

Testing the metacognitive paradox framework requires careful attention to research design. Several approaches merit consideration:

Longitudinal field experiments extending the Sun et al. (2025) paradigm can examine temporal dynamics by varying the duration and intensity of AI assistance and measuring outcomes at multiple follow-up points. These designs can distinguish suppression from erosion and identify recovery trajectories.

Experience sampling methods can capture metacognitive processes during AI-assisted and unassisted work. Repeated brief assessments of monitoring, evaluation, and strategic adjustment can reveal how these processes fluctuate across conditions and over time.

Think-aloud protocols during creative tasks can provide rich process data on metacognitive engagement. Comparing protocols during AI-assisted versus unassisted work can reveal the specific metacognitive operations that are externalized to AI systems.

Behavioral trace analysis of interaction logs with AI systems (e.g., query patterns, revision behaviors, time allocation) can provide unobtrusive indicators of metacognitive engagement. Complex, iterative queries may indicate more active metacognitive processing than simple, one-shot prompts.

Randomized withdrawal designs can examine recovery processes by randomly varying when and whether AI assistance is removed. These designs can identify the factors that facilitate or impede recovery of metacognitive functioning.

Construct Operationalization

The key constructs in our framework require careful operationalization:

Metacognitive engagement during AI-assisted work can be measured through:

 doi:10.20944/preprints202603.0121.v1

- Self-report scales adapted from existing metacognitive measures (e.g., Schraw & Dennison, 1994), modified to reference AI-assisted work contexts
- Behavioral indicators including query complexity, revision frequency, and time spent evaluating AI outputs before acceptance
- Think-aloud coding for monitoring and evaluation statements
- Physiological indicators of cognitive effort (e.g., pupillometry, EEG measures)

Metacognitive suppression versus erosion can be distinguished through:

- Temporal patterns of recovery following AI removal (rapid recovery suggests suppression; slow or incomplete recovery suggests erosion)
- Performance on transfer tasks requiring metacognitive skills not directly exercised during the focal creative task
- Measures of metacognitive knowledge accessibility and flexibility

Habitual versus strategic LLM engagement can be assessed through:

- Behavioral patterns in AI use (consistent vs. variable query types, reflection time before querying)
- Self-report measures of intentionality and goal-directedness in AI use
- Response to prompts encouraging reflection on AI use strategies

Causal Identification Challenges

Several challenges complicate causal inference in this domain:

Temporal confounds: Changes in metacognitive functioning over time may reflect maturational processes, practice effects, or historical events rather than AI assistance effects. Control conditions and multiple baseline designs can address this concern.

Selection effects: Even with random assignment to AI assistance conditions, individuals may self-select into different patterns of AI use. Encouragement designs and instrumental variable approaches may help isolate causal effects.

Mediation analysis limitations: Testing the metacognitive mechanism requires establishing that (a) AI assistance affects metacognitive engagement, (b) metacognitive engagement affects subsequent creativity, and (c) controlling for metacognitive engagement reduces the AI assistance effect. Experimental manipulation of the mediator provides stronger causal evidence than correlational mediation analysis (Spencer et al., 2005).

Measurement reactivity: Measuring metacognitive processes may alter those processes. Unobtrusive behavioral measures and within-person designs that separate measurement from treatment can reduce reactivity concerns.

Discussion

Theoretical Implications

The metacognitive paradox framework advances understanding of human-AI interaction in creative work by providing a specific, mechanistic account of how AI assistance affects human cognitive capabilities. Rather than treating AI effects as uniformly positive or negative, the framework specifies the conditions under which assistance enhances or diminishes human creativity and articulates the cognitive processes through which these effects occur.

The framework contributes to metacognition theory by identifying a new context—AI-assisted work—in which metacognitive processes operate and can be disrupted. While prior research has examined metacognition in educational and problem-solving contexts, the introduction of AI assistance creates novel dynamics that existing theories have not fully addressed. The externalization of metacognitive functions to AI systems represents a qualitatively different phenomenon from simple cognitive offloading, with potentially more consequential implications for long-term cognitive development.

For creativity research, the framework highlights the importance of considering not just creative outputs but the cognitive processes that generate them. Evaluating AI effects on creativity solely in terms of immediate output quality may miss important consequences for the sustainability of human

 doi:10.20944/preprints202603.0121.v1

creative capability. The framework suggests a need for longitudinal perspectives that track creative skill development and maintenance over time.

The framework also contributes to the growing literature on automation and human factors by extending concepts like automation complacency and skill degradation to cognitive and creative work contexts. While much of this research has focused on procedural tasks in high-reliability contexts (aviation, industrial control), the present analysis suggests that similar dynamics operate in knowledge work—with potentially broader implications given the centrality of creativity to organizational adaptation and competitiveness.

Implications for the Extended Mind Debate

Our framework engages with the extended mind thesis (Clark & Chalmers, 1998) in nuanced ways. We do not dispute that cognitive processes can genuinely extend to incorporate external tools and resources. However, we emphasize that the distribution of cognition across person and technology has implications for what happens when that distribution changes.

If metacognitive processes become distributed across person and AI system, removing the AI system does not simply return the individual to their prior state. Rather, the individual must reconstitute processes that have been externalized, and they may find their capacity for this reconstitution diminished. The extended mind, once contracted, may not readily re-expand to its former boundaries.

This observation suggests a refinement of the extended mind thesis: cognitive extension is not without costs, and those costs become apparent when extension is disrupted. For sustainable cognitive augmentation, the human component of the extended system must maintain sufficient capability to function independently when technology becomes unavailable—whether due to technical failure, policy change, or shifting task demands.

Practical Implications for Organizations

The metacognitive paradox framework carries significant implications for how organizations deploy AI tools for creative work.

Training and onboarding: Organizations should prepare employees for AI tool use not just in terms of technical operation but in terms of maintaining metacognitive engagement. Training programs should emphasize the importance of reflective practice, critical evaluation of AI outputs, and periodic unassisted work.

Work design: Rather than maximizing AI tool availability, organizations should consider structured alternation between AI-assisted and unassisted work. "Metacognitive maintenance intervals"—protected periods for independent creative work—may help preserve human creative capabilities.

Individual differences: Recognition that individuals differ in their vulnerability to metacognitive erosion suggests value in differentiated AI deployment strategies. Individuals with lower baseline metacognitive ability may benefit from additional scaffolding, monitoring, and training to maintain metacognitive engagement during AI-assisted work.

Performance evaluation: Evaluation systems should assess not just immediate creative outputs but sustained creative capability. Organizations might monitor for signs of increasing AI dependence and take corrective action when individuals show diminished independent creative performance.

AI system design: The findings suggest value in AI systems designed to support rather than supplant human metacognition. Systems that prompt users to evaluate outputs, explain reasoning, and reflect on alternatives may preserve metacognitive engagement better than systems optimized purely for output quality.

Ethical Considerations

The metacognitive paradox raises important ethical questions that deserve acknowledgment.

Equity implications: If AI assistance is differentially beneficial for high-skill individuals (as Sun et al.'s findings suggest), widespread AI deployment may exacerbate rather than reduce skill-based inequalities. Organizations should consider whether AI tools are widening gaps between high- and low-skill employees and take steps to ensure equitable benefit distribution.

 doi:10.20944/preprints202603.0121.v1

Responsibility for skill maintenance: When AI use leads to skill erosion, who bears responsibility? Individuals may have limited awareness of their own capability changes. Organizations that mandate or encourage AI use may bear some responsibility for resulting skill effects. Technology developers might consider designing systems that minimize negative cognitive consequences.

Informed consent: Employees may not fully understand the potential cognitive consequences of AI tool use. Ethical deployment may require informing workers about possible effects on their skills and providing options for limiting AI use.

Long-term workforce implications: Widespread dependence on AI assistance could create workforce vulnerabilities—pools of workers who cannot function effectively without AI support. This dependence raises concerns about technological lock-in and reduced organizational resilience.

Limitations and Future Directions

Several limitations of the present analysis suggest directions for future research.

First, our framework is primarily derived from a single, albeit rigorous, field experiment. Replication across different organizational contexts, creative domains, and AI systems is essential. The dynamics observed in creative writing may differ from those in visual design, scientific problem-solving, or product development.

Second, the temporal dynamics of metacognitive erosion and recovery remain underspecified. Longitudinal research with extended follow-up periods can better characterize the time course of these effects and identify factors that accelerate or retard recovery.

Third, our analysis focuses on individual-level processes. Much organizational creativity is collaborative, and future research should examine how AI assistance affects team-level metacognitive processes and collective creativity.

Fourth, individual differences beyond metacognitive ability warrant investigation. Cognitive styles, personality factors, prior technology experience, and domain expertise may all moderate the effects of AI assistance on metacognition and creativity.

Fifth, the generalizability of findings across different AI systems deserves examination. ChatGPT, Claude, and specialized creative AI tools differ in their capabilities, interfaces, and interaction patterns. These differences may affect the metacognitive dynamics we describe.

Sixth, the relationship between metacognitive erosion and creative identity remains unexplored. Sustained AI use may affect not only cognitive capabilities but also individuals' self-concepts as creative professionals, with potential consequences for motivation and career development.

Toward a Research Agenda

We propose a coordinated research agenda to advance understanding of the metacognitive paradox:

Immediate priorities:

- Replication of Sun et al. (2025) findings across different creative domains and organizational contexts
- Development and validation of measures specifically designed to assess metacognitive engagement during AI-assisted work
- Experimental studies varying the type and intensity of AI assistance to identify dose-response relationships

Medium-term goals:

- Longitudinal studies with extended follow-up to characterize temporal dynamics of metacognitive erosion and recovery
- Intervention studies testing strategies for maintaining metacognitive engagement during AI-assisted work
- Examination of organizational factors that shape how individuals engage with AI tools

Longer-term objectives:

- Development of AI systems designed to support rather than supplant human metacognition
- Investigation of developmental trajectories—how AI assistance during early career stages affects creative skill acquisition

 doi:10.20944/preprints202603.0121.v1

- Examination of societal implications of widespread cognitive dependence on AI systems

Conclusion

The introduction of generative AI into creative work represents a profound technological shift with far-reaching implications for human cognition and capability. Sun et al.'s (2025) field experiment provides crucial evidence that these implications are more complex than early enthusiasm suggested: AI assistance enhances immediate performance while potentially undermining the cognitive foundations of independent creativity.

The metacognitive paradox framework developed in this article offers a theoretical lens for understanding these dynamics. By identifying the externalization of metacognitive functions to AI systems as the core mechanism, the framework explains both why AI assistance produces immediate benefits and why those benefits may come at a cost to sustained creative capability. The moderating role of metacognitive ability suggests that these costs fall disproportionately on those least equipped to bear them.

Importantly, the framework also points toward solutions. Intentional metacognitive engagement, strategic patterns of AI use, organizational support for independent creative work, and AI systems designed to preserve human metacognition can all help realize the benefits of AI assistance while mitigating its costs. The goal is not to reject AI tools but to integrate them thoughtfully in ways that augment rather than replace human cognitive capabilities.

As AI systems become increasingly capable and ubiquitous, understanding their effects on human cognition becomes ever more critical. The present analysis suggests that this understanding requires moving beyond simple questions of whether AI helps or hurts performance to examine the specific cognitive processes through which AI affects human capability. Only with such understanding can we chart a course toward AI integration that genuinely enhances rather than diminishes human potential.

Appendix: Summary of Propositions

Proposition	Statement	Key Variables
1a	Strategic monitoring mediates the relationship between LLM assistance and subsequent creative performance	LLM assistance → Strategic monitoring → Creativity
1b	Metacognitive knowledge access and refinement is reduced during LLM-assisted work	LLM assistance → Metacognitive knowledge → Creativity
1c	Non-linear (inverted-U) relationship between LLM assistance intensity and subsequent performance	Assistance intensity (moderate vs. intensive)
2a	Higher metacognitive ability leads to greater resilience in subsequent creative performance	Metacognitive ability as moderator
2b	Need for cognition moderates LLM effects on subsequent creativity	Need for cognition as moderator
2c	Domain expertise moderates LLM effects; experts show smaller decrements	Domain expertise as moderator
3a	Metacognitive suppression occurs rapidly; erosion develops gradually	Time course of effects
3b	Recovery depends on duration and intensity of prior AI assistance	Dose-response relationship
3c	Periodic unassisted work intervals prevent metacognitive erosion	Intervention design
4a	Task complexity moderates effects; larger effects for complex tasks	Task complexity as moderator
4b	Generative AI assistance produces larger effects than evaluative assistance	Type of AI assistance

 doi:10.20944/preprints202603.0121.v1

Proposition	Statement	Key Variables
4c	Organizational norms emphasizing critical AI engagement reduce decrements	Organizational context
4d	Psychological safety moderates effects by encouraging reflective engagement	Psychological safety as moderator
5	Effects differ across creative process stages; early-stage effects may be larger	Creative process stage

References

1. Amabile, T. M. (1988). A model of creativity and innovation in organizations. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 10, pp. 123–167). JAI Press.
2. Amabile, T. M., & Pratt, M. G. (2016). The dynamic componential model of creativity and innovation in organizations: Making progress, making meaning. *Research in Organizational Behavior*, *36*, 157–183.
3. Anderson, N., Potočnik, K., & Zhou, J. (2014). Innovation and creativity in organizations: A state-of-the-science review, prospective commentary, and guiding framework. *Journal of Management*, *40*(5), 1297–1333.
4. Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131.
5. Chi, M. T. H., Glaser, R., & Farr, M. J. (Eds.). (1988). *The nature of expertise*. Lawrence Erlbaum Associates.
6. Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.
7. Davidson, J. E., & Sternberg, R. J. (1998). Smart problem solving: How metacognition helps. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 47–68). Lawrence Erlbaum Associates.
8. Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268.
9. Edmondson, A. C. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(2), 350–383.
10. Eisenberger, R. (1992). Learned industriousness. *Psychological Review*, *99*(2), 248–267.
11. Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*(2), 381–394.
12. Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406.
13. Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911.
14. Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410.
15. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.
16. Perry-Smith, J. E., & Mannucci, P. V. (2017). From creativity to innovation: The social network drivers of the four phases of the idea journey. *Academy of Management Review*, *42*(1), 53–79.
17. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688.
18. Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*(4), 460–475.
19. Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, *57*(2), Article 103174.
20. Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Technical Report). MIT Man-Machine Systems Laboratory.
21. Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*(6), 845–851.

 doi:10.20944/preprints202603.0121.v1

22. Sun, S. (2024). A componential and functional framework for metacognition: Implications for research in personnel and human resources management. In M. R. Buckley, A. R. Wheeler, J. E. Baur, & J. R. B. Halbesleben (Eds.), *Research in personnel and human resources management* (Vol. 42, pp. 45–73). Emerald Publishing Limited.
23. 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.
24. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.
25. Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction*, *14*(1), 89–109.
26. Winne, P. H., & Nesbit, J. C. (2010). The psychology of academic achievement. *Annual Review of Psychology*, *61*, 653–678.
27. Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100.
28. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., & Wen, J.-R. (2023). *A survey of large language models*. arXiv preprint arXiv:2303.18223.
29. Zhou, J., & Shalley, C. E. (2003). Research on employee creativity: A critical review and directions for future research. In J. J. Martocchio & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 22, pp. 165–217). Emerald Group Publishing Limited.