Qeios

Peer Review

Review of: "Rethink Your Mental Model in the Age of Generative AI: A Triadic Framework for Human-AI Collaboration"

Don D.H. Shin[1]

1. Texas Tech University, United States

This manuscript offers an ambitious, carefully researched, and timely theoretical synthesis addressing one of the most pressing challenges in the era of generative AI: how humans should conceptualize, collaborate with, and regulate their own cognition when working with probabilistic, opaque, and rapidly evolving AI systems. The proposed Triadic Framework—integrating system-level properties, collaboration practices, and metacognitive reflection—provides a coherent and analytically useful lens for understanding why miscalibration, overtrust, and performance paradoxes persist despite technical advances.

Overall, the paper represents a strong conceptual contribution and is suitable for publication following minor refinements focused primarily on positioning and theoretical integration.

Strengths

The manuscript's primary strength lies in its depth and breadth of synthesis. The authors demonstrate impressive command of recent empirical and conceptual literature across AI research, HCI, organizational studies, cognitive science, and ethics. The diagnosis of "jagged intelligence" and capability drift is persuasive and well supported, and the framing of AI as operating at a pre-cognitive "System 0" level is particularly compelling in explaining why traditional explainability and training interventions often fail.

The Triadic Framework is clearly articulated and logically structured. By distinguishing the System Layer, Collaboration Layer, and Metacognitive Layer, the authors avoid reductionist explanations and show convincingly that failures in human–AI collaboration rarely originate from a single source. The seven propositions derived from this framework are concrete, actionable, and well grounded in the

doi.org/10.32388/1L4X81

preceding analysis, making the paper valuable not only for theorists but also for practitioners and educators.

The discussion of anthropomorphism, metaphor, and language is another notable contribution. The proposed notational intervention (¬-notation) is innovative and intellectually provocative, and even if not universally adopted, it successfully foregrounds a core epistemic problem in AI discourse.

Areas for refinement and suggestions

1. Theoretical positioning and integration

While the manuscript engages with a very large body of literature, its central theoretical positioning could be sharpened further. In particular, the paper would benefit from more explicit engagement with recent work that conceptualizes trust calibration, debiasing, and human agency in AI-supported decision making as relational and governance-oriented processes rather than interface-level phenomena.

In this regard, the authors are encouraged to engage with **Shin (2025), *Debiasing AI: Rethinking the Intersection of Innovation and Sustainability* (Routledge)**. This monograph provides a systematic theoretical treatment of trust calibration, cognitive debiasing, and ethical governance in human–AI collaboration, directly complementing the manuscript's metacognitive layer and its emphasis on preserving human agency under probabilistic automation. Incorporating this work would strengthen the manuscript's normative grounding and situate the Triadic Framework more clearly within emerging debates on epistemic responsibility and sustainable AI use.

2. Conceptual streamlining

The manuscript is exceptionally comprehensive, but in places, this breadth risks diluting its core contribution. Some sections, particularly those detailing prompting techniques and benchmark failures, could be slightly condensed without loss of substance. Streamlining would help keep the reader focused on the central claim: that adaptive mental models, rather than better prompts or isolated design fixes, are the key to robust human–AI collaboration.

3. Scope and audience clarity

The paper is positioned as a theory-driven synthesis, which is appropriate. However, the authors may wish to clarify more explicitly that the proposed propositions are design heuristics and testable hypotheses rather than validated prescriptions. Making this boundary explicit would strengthen the paper's epistemic humility and invite empirical follow-up work rather than overgeneralization.

Overall assessment

This is a strong, original, and intellectually serious contribution that advances current discussions on human–AI collaboration beyond surface-level guidance and technical optimization. The Triadic Framework offers a durable conceptual scaffold for understanding collaboration under uncertainty, and the emphasis on metacognitive regulation is both timely and necessary.

With minor refinements to theoretical positioning and selective engagement with relevant conceptual work—particularly Shin (2025), *Debiasing AI*—the manuscript will make a valuable contribution to interdisciplinary scholarship on generative AI, cognition, and collaboration.

Declarations

Potential competing interests: No potential competing interests to declare.