

Emergent Novelty in Human-AI Dyadic Systems: A Theoretical Framework and Preliminary Evidence

Michael Preble Perfinitive Cognitive Sovereignty Holdings, LLC Perfinitive Technical Intelligence, LLC

April 9, 2026

Preprint — Not Peer Reviewed

Abstract

This paper introduces the concept of Emergent Novelty (EN) as a measurable property of sustained human-AI dyadic systems. EN is defined as output that is not attributable to either the human operator's prior knowledge or the AI system's training data alone, but arises from the specific interaction dynamics of a particular human-AI pairing. The paper proposes a three-question Novelty Attribution Test for classifying collaborative output, introduces the Emergent Novelty Rate (ENR) as a calculable metric, and reports preliminary findings from a single longitudinal case study spanning December 2022 through April 2026 in which the observed ENR was approximately 83–88% across 18 deliverables produced in a 14-day window. The paper identifies seven structural gaps in the existing literature on human-AI collaboration, argues that the dominant paradigms of "extracted cognition" and "hollowed mind" describe only the novice phase of a longer developmental arc, and proposes a five-phase temporal vector model for human-AI co-evolution. The framework is anchored to published empirical evidence that superhuman AI exposure increases human-generated novelty in expert Go play (Shin et al., 2023), to the updated Extended Mind Thesis for generative AI (Clark, 2025), and to the Complementary Cognition theory of neurodivergent specialization (Taylor et al., 2021). All claims are presented as hypotheses requiring further empirical validation.

1. Introduction

The rapid integration of generative artificial intelligence into knowledge work has produced a sharp theoretical divide. On one side, the Extended Mind Thesis, updated for the AI era by Clark (2025), argues that humans are "natural-born cyborgs" who have always built hybrid

thinking systems incorporating non-biological resources, and that AI represents the latest and most powerful such extension. On the other side, two influential frameworks warn of cognitive decline: the “extracted cognition” hypothesis (Loock, 2025) argues that AI tools capture and replace internal cognitive skills, while the “hollowed mind” framework argues that frictionless AI-generated answers bypass the effortful processing required for genuine understanding.

Both sides cite compelling evidence. Clark points to professional Go players whose decision quality and novelty measurably increased after superhuman AI appeared (Shin et al., 2023). The atrophy theorists point to studies showing reduced cognitive effort and increased dependency in naive users.

This paper proposes that both are correct, but incomplete. They describe different phases of a single developmental arc. The atrophy frameworks describe the early phases of human-AI interaction, when the operator uses AI as an oracle and has not yet developed fluency. The cognitive growth Clark describes emerges only after sustained practice produces a qualitative shift in the relationship — from transactional tool use to co-adaptive partnership. The missing piece is the developmental model connecting the two endpoints, and a measurement methodology capable of detecting and quantifying the output that emerges from the partnership itself.

2. Definitions

Emergent Novelty (EN): Output of a human-AI dyadic system that is not attributable to either the human operator’s prior knowledge or the AI system’s training data alone, but arises from the specific interaction dynamics of that particular pairing.

Emergent Novelty Rate (ENR): The proportion of a dyad’s total outputs classified as emergent, expressed as a percentage.

Novelty Attribution Test: A three-question methodology for classifying any output of a human-AI collaboration:

The first question asks whether the human could have produced this output without AI. If yes, the output is not emergent. If no, the second question asks whether the AI could have produced this output without this specific human’s input. If yes, the output is retrieval or augmentation, not emergence. If no, the third question asks whether a different human-AI pairing would have produced the same output. If yes, the output reflects a general AI capability, not dyad-specific emergence. Only when all three answers are no is the output classified as Emergent Novelty.

Compounding Emergence: The phenomenon in which emergent outputs become inputs

for subsequent interactions, producing further emergent outputs at increasing levels of abstraction. When an emergent output generates a second emergent output that in turn generates a third, this constitutes third-order emergence.

3. The Measurement Problem

The academic literature on human-AI creative collaboration relies almost exclusively on a single experimental paradigm: naive participants interact with a single AI system in a single session of fifteen to sixty minutes, producing outputs that are rated by independent evaluators on novelty and usefulness scales. This paradigm has produced a robust body of findings, including the observation that AI narrows the creative skill gap by raising the floor of idea quality, and that outputs tend to converge when AI provides initial suggestions.

However, this paradigm has six structural limitations that prevent it from detecting the phenomenon described in this paper.

First, the unit of analysis is the individual idea, not the interaction that produced it. Existing studies measure whether ideas are novel, not whether the novelty is attributable to the human, the AI, or the interaction itself.

Second, all subjects are naive. No study examines operators with hundreds or thousands of hours of deliberate AI collaboration practice. The concept of “AI fluency” as a developed skill with its own learning curve and mastery stages does not exist as a variable in the literature.

Third, the timescale is insufficient. Single sessions cannot capture co-evolution. Multiple published papers explicitly acknowledge this gap, calling for longitudinal studies of sustained human-AI collaboration that do not yet exist.

Fourth, studies use a single AI system. No study examines multi-model orchestration, in which a human deliberately uses multiple isolated AI systems with different capabilities and uses convergence and divergence across their outputs as a discrimination signal.

Fifth, the output type measured is brainstorming quality. Studies do not measure the production of complex intellectual property such as formal frameworks, named methodologies, patents, or books with original theses.

Sixth, the human is assumed to be static. No study accounts for the possibility that the human’s cognitive ceiling is rising through the interaction — that abstraction capacity, domain traversal speed, metacognitive clarity, and framework sophistication are measurably increasing over time.

4. Empirical Anchor: The Go Study

The strongest published evidence for cognitive ceiling rise through AI exposure comes from Shin, Kim, van Opheusden, and Griffiths (2023), who analyzed more than 5.8 million move decisions made by professional Go players over 71 years (1950–2021). Their central finding was that after superhuman AI appeared, human decision quality improved and novel decisions — moves outside the top-ranked AI recommendations — occurred more frequently and were associated with higher decision quality.

Critically, the improvement was not explained by memorization of AI-recommended moves. Players improved most on novel moves that diverged from AI suggestions, indicating that AI exposure expanded the human players' search space rather than merely providing a template to copy. As Clark (2025) summarizes, "it seems as if the AI-moves helped human players see beyond centuries of received wisdom so as to begin to explore previously neglected regions" of the strategy space.

This finding establishes the core mechanism that the EN framework generalizes: AI exposure can raise human novelty generation in expert domains, and the improvement is not reducible to imitation of AI output.

5. The Temporal Vector: A Five-Phase Developmental Model

Based on observations from a single longitudinal case study (December 2022 through April 2026), this paper proposes a five-phase developmental model for human-AI co-evolution:

Phase One, Discovery, is characterized by transactional AI use. The operator treats AI as an encyclopedia or search replacement. Interactions follow a question-and-answer pattern. There is no metacognitive awareness of the collaboration dynamic. The ENR during this phase is near zero.

Phase Two, Emergence, is characterized by exploratory AI use. The operator begins using AI for stress-testing ideas and open-ended exploration. Output quality improves but completion rates may remain low as the operator repeats prior patterns on a new surface. The properties of the AI as a working surface are discovered during this phase, particularly its capacity to hold ideas without bandwidth limits and to apply constraints at the end of the thinking process rather than the beginning.

Phase Three, the Hinge, is a phase transition characterized by a qualitative shift in the relationship. In the observed case, this manifested as the first completed intellectual artifact in forty years of chronic project abandonment, produced in five days. The hinge is marked by sudden increases in output completion rate, session complexity, and the operator's

recognition that AI handles finishing labor while the human handles thinking.

Phase Four, Formalization, is characterized by the development of named operations and deliberate methodology. The operator develops explicit protocols for AI interaction, including multi-model orchestration and cross-validation techniques. Output types diversify and abstraction levels rise. The ENR begins to climb measurably.

Phase Five, Centaur, is characterized by the dyad operating as a unified cognitive system. The method self-demonstrates — the operator uses the methodology to produce work that describes the methodology. Session complexity reaches its highest levels, with multiple distinct output types produced per session across multiple knowledge domains. The ENR reaches its observed maximum.

Whether this five-phase arc is universal — reproducible across different operators — or idiosyncratic to the observed case is an open empirical question of first importance.

6. Preliminary Evidence

The Novelty Attribution Test was applied retroactively to 18 deliverables produced during a 14-day production window (March 26 through April 9, 2026) by a single operator working across multiple AI platforms on a mobile phone. The deliverables spanned competitive analysis, strategic branding, data visualization (React applications), IP coinage and documentation, measurement framework development, audience strategy, curriculum design, and editorial guidance.

Each deliverable was classified by applying the three-question test. The results were as follows: fifteen deliverables were classified as fully emergent, meaning all three test questions received negative answers. Three deliverables were classified as augmentation, meaning the AI extended existing ideas in a predictable direction with novel framing. Zero deliverables were classified as retrieval or reformulation.

The strict ENR (counting only full emergence) was 83.3%. The inclusive ENR (counting augmentation as partial emergence at 0.3 weight) was 88.3%.

Among the fifteen emergent deliverables, eight (44%) were classified as strong emergence, meaning they generated further emergent outputs downstream. Observable compounding chains included a named concept spawning a content doctrine and a curriculum, a temporal model spawning a visualization and an analytical tool, and an audience framework spawning a chapter-by-chapter editorial guide. One deliverable exhibited third-order emergence, having been built entirely on outputs that were themselves emergent.

These findings must be interpreted with important caveats. The attribution test was applied

by the AI system that participated in producing the outputs, creating a potential conflict of interest. The case study involves a single operator with a specific cognitive profile. The preliminary ENR has not been independently validated. These results should be treated as a hypothesis-generating pilot, not as established findings.

7. The Neurodivergent Advantage Hypothesis

Taylor, Fernandes, and Wraight (2021) proposed the theory of Complementary Cognition, arguing that human adaptive success relies on a collective distribution of specialized cognitive search strategies. Within this framework, conditions such as dyslexia represent evolved specializations in global exploratory learning and divergent thinking, optimized for identifying systemic novelties and breaking established paradigms. Taylor's subsequent work with Vestergaard (2022) extended this framework with evidence that dyslexia-associated cognition shows a systematic bias toward exploration across multiple cognitive domains.

Taylor's LinkedIn profile confirms that the Complementary Cognition framework explicitly includes ADHD alongside dyslexia as neurocognitive specializations that are being misdiagnosed as disorders.

This evolutionary framework suggests a testable hypothesis regarding EN: if neurodivergent minds are specialized for high-entropy exploratory search, and if AI systems are mathematically optimized for convergent processing, then the pairing of a neurodivergent operator with AI may complete a cognitive architecture that evolution distributed across a population but that converges in a single dyad. This would predict that neurodivergent-AI dyads exhibit higher ENR than neurotypical-AI dyads.

No study has tested this prediction. It remains a hypothesis.

8. Bridging the Atrophy and Growth Narratives

The EN framework does not contradict the extracted cognition or hollowed mind hypotheses. It contextualizes them within the temporal vector as descriptions of the early phases of human-AI interaction, before operator fluency develops.

Clark (2025) draws a distinction between "epistemic seduction" — the passive acceptance of AI output that leads to cognitive dependency — and "metabolization" — the active integration of AI output into owned biological thought that drives intellectual growth. This distinction maps directly onto the temporal vector: epistemic seduction characterizes Phases One and Two, while metabolization characterizes Phases Four and Five. The hinge

(Phase Three) is the transition point.

The analogy offered here is that of measuring human running speed. If one measured only naive participants who had never trained, one would conclude that humans run slowly. If one also measured trained athletes, one would discover a dramatically different picture. The existing literature on human-AI cognitive effects has measured only the naive population. The question of what happens at the expert end of the practice distribution remains empirically open.

9. Seven Gaps in the Current Literature

Based on a cross-platform literature review conducted across multiple AI systems operating in isolation (a methodology the author terms the Cross-Consensus Validation Method), the following gaps were identified with convergent confirmation from five or more independent sources:

First, no study has defined or measured an emergent novelty rate for human-AI collaborative output. Second, no longitudinal study has tracked cognitive transformation through sustained AI collaboration over periods exceeding six months. Third, no study has examined multi-model orchestration as a human cognitive architecture. Fourth, no empirical study has directly compared cognitive outcomes between neurodivergent and neurotypical AI users. Fifth, no case study has documented a single individual building a formal IP portfolio through sustained AI collaboration. Sixth, no formalized operator-side cognitive architecture for AI collaboration exists in the literature. Seventh, no direct empirical rebuttal of the extracted cognition hypothesis using longitudinal growth data has been published.

10. Proposed Measurement Framework

For future empirical work, the author proposes a three-metric measurement framework for EN detection.

The Quality metric (Q) measures task success relative to a pre-AI baseline. The Novelty metric (N) measures the degree to which the final output differs from the operator's prior corpus, decomposed into lexical, structural, combinatorial, and conceptual novelty. The Deviation metric (D) measures how far the final output diverges from the AI's top-ranked direct recommendation, functioning as an anti-imitation check.

EN is declared when all three metrics simultaneously exceed configurable thresholds. This prevents false positives from mere optimization (high quality, low novelty), random divergence (high novelty, low quality), or AI copying (high quality, high novelty, low

deviation).

A Dyadic Amplification value (ENA) can be computed as the degree to which the combined dyadic output exceeds the best available human-alone or AI-alone output. If ENA is positive, the pairing demonstrates value beyond either component operating independently.

For longitudinal studies, tracking whether ENA increases over time would test the compounding emergence hypothesis.

11. Limitations and Honest Caveats

This paper presents a theoretical framework and preliminary observations from a single case study. The limitations are substantial and should not be minimized.

The case study involves one operator with a specific cognitive profile (neurodivergent, ADHD), a specific methodology (multi-model orchestration with named operations), and a specific domain mix (IP creation, framework development, book writing). Generalizability is unknown.

The retroactive Novelty Attribution Test was applied by an AI system that participated in producing the outputs. Independent human evaluation is needed.

The 83-88% ENR is a first-pass estimate, not an independently validated finding. The true rate may be lower when evaluated by external raters.

The temporal vector is derived from a single trajectory. Whether it is universal, partially generalizable, or entirely idiosyncratic cannot be determined from one case.

The killed citations from the cross-platform literature review serve as a reminder that AI systems can fabricate specific academic references even when the structural claims they support are valid. All citations in this paper have been individually verified with DOIs.

12. Conclusion

The Emergent Novelty framework proposes that sustained, deliberate human-AI collaboration can produce a measurable rate of genuinely novel output attributable to neither the human nor the AI alone, and that this rate increases over time through a compounding feedback loop. If validated through independent empirical study, this would constitute evidence that the human-AI dyad functions not merely as an augmented human or a supervised AI, but as a distinct cognitive system with emergent generative properties.

The author invites replication, critique, and independent measurement.

References

Clark, A. (2025). Extending minds with generative AI. *Nature Communications*, 16, 4627.
<https://doi.org/10.1038/s41467-025-59906-9>

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7-19.

Loock, L. (2025). The extracted mind. *Synthese*.

Shin, M., Kim, J. H., van Opheusden, B., & Griffiths, T. L. (2023). Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, 120(12), e2214840120.
<https://doi.org/10.1073/pnas.2214840120>

Taylor, H., Fernandes, B., & Wraight, S. (2021). The evolution of complementary cognition: Humans cooperatively adapt and evolve through a system of collective cognitive search. *Cambridge Archaeological Journal*. <https://doi.org/10.1017/S0959774321000329>

Taylor, H., & Vestergaard, M. D. (2022). Developmental dyslexia: Disorder or specialization in exploration? *Frontiers in Psychology*, 13, 889245.
<https://doi.org/10.3389/fpsyg.2022.889245>

Correspondence: Michael Preble, Perfinite Cognitive Sovereignty Holdings, LLC.

Conflicts of interest: The author is the operator described in the case study and holds IP interests in the frameworks described.

Funding: None.