Operationalizing Extended Cognition in Human–AI Systems

The extended mind framework (Clark & Chalmers, 1998) remains conceptually stable but lacks robust operationalization. While embodied and extended cognition have provided theoretical insights, they have struggled to empirically distinguish between the genuine cognitive components and the causally relevant background conditions

Recent advances in Artificial Intelligence, particularly large language models, introduce interactive systems that dynamically influence human cognition. These systems are adaptive, context-sensitive, and capable of reshaping reasoning processes, raising a central question: “Under what conditions does Artificial Intelligence become part of cognition rather than a tool?”

The mutual manipulability (Kaplan, 2012) criterion provides a mechanistic approach to demarcate cognitive boundaries. A component X is part of a mechanism M if the following conditions are provided:

- Bottom-up: manipulating X changes the behavior of M
- Top-down: manipulating M changes the behavior or state of X

In the application to Human– Artificial Intelligence Agents, that can be translated to the following experiment:

- Bottom-up: modifying model properties (memory, context, fine-tuning) alters human reasoning
- Top-down: modifying human goals or prompts alters model outputs and internal dynamics

To overcome binary classifications, cognition is modeled as a multidimensional integration space following Heersmink’s criteria of Integration with 8 relevant dimensions.

We propose a dual-criterion approach, one structural as the mutual manipulability establishes bidirectional causal coupling and another gradual one as the Integration is quantified across Herrsmink’s multiple dimensions.

Our hypothesis considers human–Artificial Intelligence agents qualifies as extended cognitive systems when the bidirectional causal coupling is present and a high integration is achieved across multiple dimensions.

The framework enables quantitative analysis of human–AI systems through:

- graph-based representations of reasoning processes.
- embedding-space dynamics (semantic evolution and convergence)
- complexity-based metrics.
- POMDP modeling the interaction dynamics.

These allow measuring the transformation of human cognition, the dependency and persistence effects and the emergent human–AI synergy.

Main implication leading to the cognitive extension becoming experimentally testable where AI systems can be designed to meet integration criteria ensuring the boundary of cognition can become a problem of system design and measurement.

Extended cognition can be reframed as a measurable property of human–AI systems, grounded in causal coupling and a degree of integration.

The key challenge is not defining cognition, but designing and evaluating systems where cognition is genuinely shared.