TOWARD INTENTIONAL AGENCY IN ARTIFICIAL SYSTEMS: INTERPLAY OF INPUT, OUTPUT, AND PROCESSING

Artem S. Yashin (1,2)

1 – MEG Moscow State University of Psychology and Education; 2 – NRC "Kurchatov Institute"
Correspondence: yashinart1996@gmail.com

This poster

INTENTIONAL AGENCY IN AI?

Intentional agency depends on an agent's internal architecture, not merely on whether observers adopt the intentional stance (Dennett 1971).

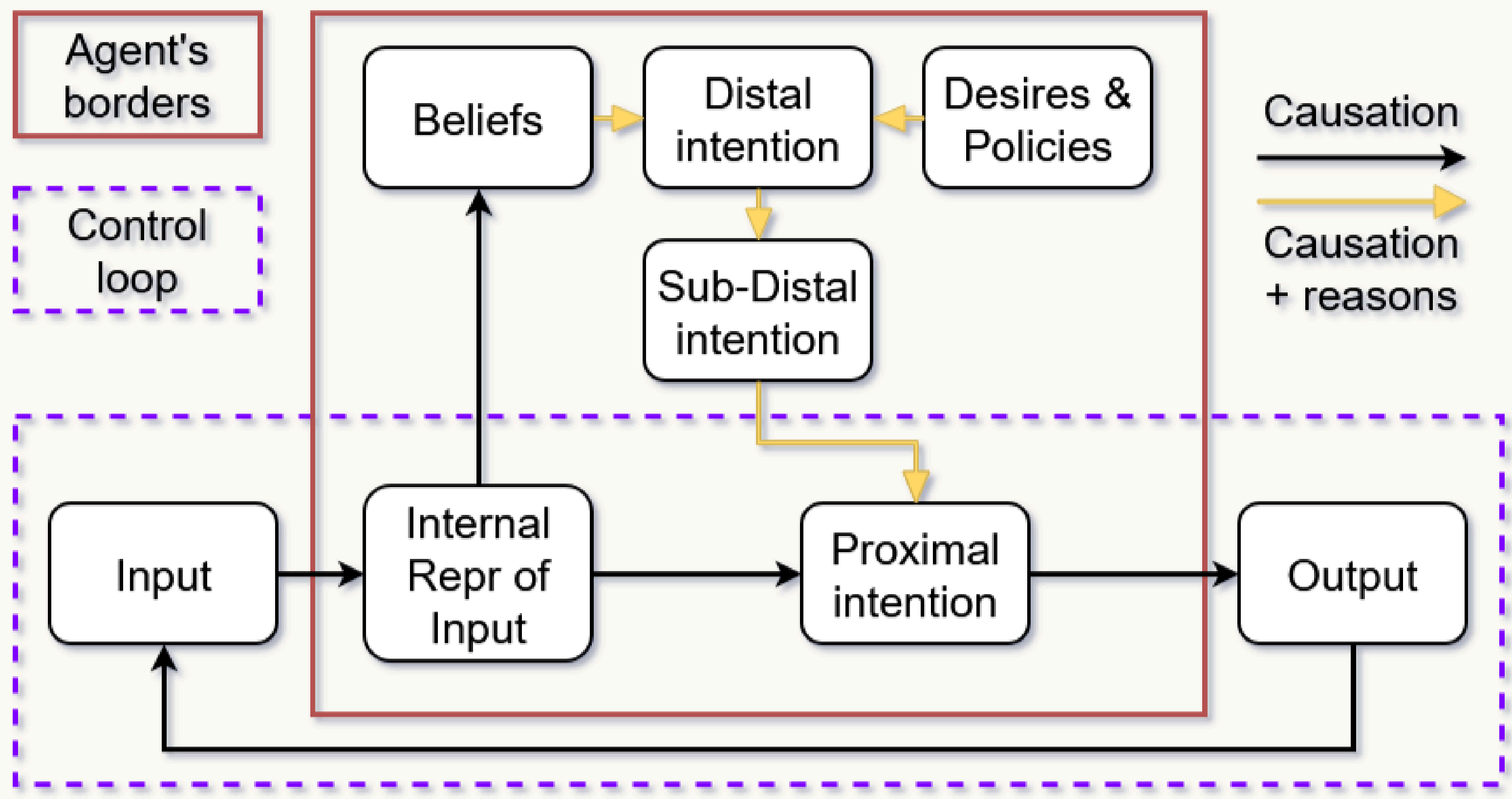
It may be built in – as in **Belief-Desire-Intention (BDI)** systems (Rao & Georgeff 1991) – or it may emerge in neural networks, whose structures must be analyzed. Hybrid designs now link large networks to BDI planners (Frering et al. 2025). Complex input-output mappings need not be intentional; the question is how closely AI cognition parallels human cognition.

ON AGENCY AND INTENTIONS

Dynamic goal pursuit: Tomasello (2022) calls a system intentional when it flexibly adjusts behaviour to reach goals.

Rational coherence: Philosophers add that its actions must have reasons and that its plans need to remain logically consistent (Schlosser 2019).

Intentions are commitments to act – irreducible to beliefs and desires (Bratman 1987). They nest hierarchically into distal and proximal levels, forming a means-ends structure that guides behavior.



Scheme of an intentional agent (human/classic BDI architecture)

MENTAL REPRESENTATIONS

For an AI agent to form intentions, it must possess **internal representations** that exercise specific control functions. How closely these resemble human mental states remains contested, as the nature of mental representations might be entangled with phenomenal consciousness, embodiment etc. Circuit-mapping studies suggest that Large Language Models (LLMs) develop multilingual, abstract features (Anthropic 2025).

REFERENCES

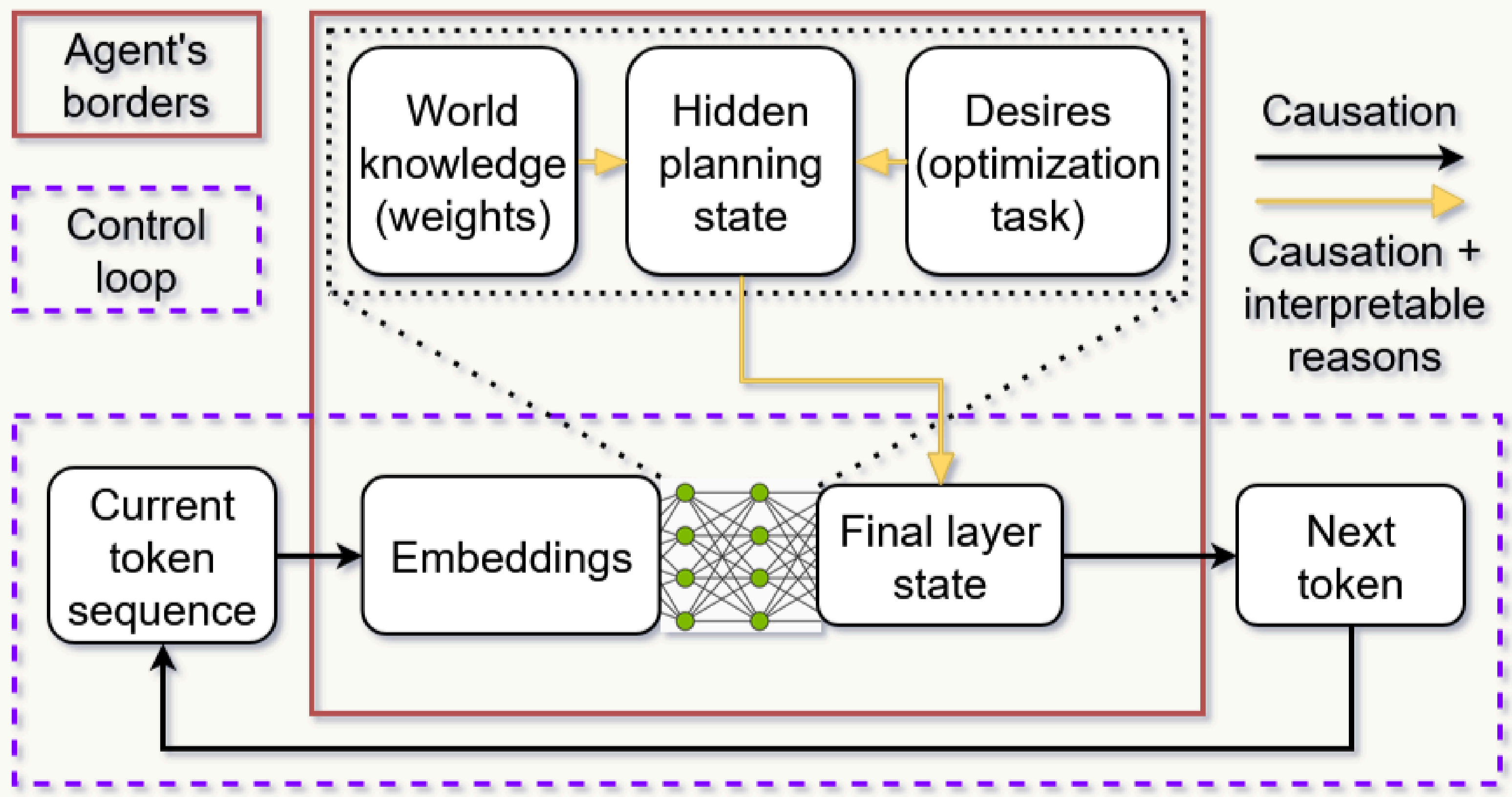
Anthropic. On the Biology of a Large Language Model. Fo.
Bratman (1987). Intentions, plans, and practical reason. Harvard University Press.
Clark & Chalmers (1998). The Extended Mind. Analysis, 58(1), 7–19.
Dennett (1971). Intentional Systems. The Journal of Philosophy, 68(4), 87–106.
Frering et al. (2025). Integrating Belief-Desire-Intention agents with large language models for reliable human–robot interaction and explainable Artificial Intelligence. Engineering Applications of Artificial Intelligence, 141, 109771.
Jenner et al. (2024) Evidence of Learned Look-Ahead in a Chess-Playing Neural Network. arXiv.
Rao & Georgeff (1991). Modeling rational agents within a BDI-architecture. Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning, 473–484.
Schlosser (2019). Agency. In Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/agency/
Tomasello (2022). The Evolution of Agency. The MIT Press.

LLM AS A PLANNING AGENT?

Apparently, LLMs can draft short plans, such as chess moves (Jenner et al. 2024) or rhymes for poems (Anthropic 2025).

Hidden-state scratch-note: each forward pass lets the LLM off-load info to prior activations – mirroring Otto's notebook in extended cognition (Clark & Chalmers 1998), though the scaffold stays inside the model and vanishes after the generation window.

Adaptive agents require (1) attentional routing to select relevant signals and (2) a working-memory buffer whose persistence matches the horizon of intention revision – from fleeting hidden-state scaffolding to durable external stores.



Scheme of an LLM agent without BDI modules (in a single generation window)

ASSESSING INTENTIONAL AGENCY

Behavior alone is not enough: to decide whether a model acts intentionally we must recover its representational content – identifying both the causal role of these representations and the relations between their contents.

Top-down representations (intentions): Intentions are the states that impose control on output and can be replaced when the agent revises its goals.

Control loops: Every agent that can re-intend needs control loops that monitor success and update intentions.

Nested plans: Intentions must form a hierarchy: distal goals generate proximal sub-goals in a coherent means-end chain.

CONCLUSION

Any system instantiating mental representations that **causally** and **rationally** organize behavior can be an intentional agent. In a BDI architecture with nested plans and a control loop, agency could be directly engineered.

However, agency can also emerge in neural networks (e.g. LLMs), where it may be **ephemeral** and **reliant on scaffolding** through hidden states (thus resembling extended cognition).