

Metacognitive Alignment for Self-Governing World Models

Eugene Yu Ji^{1, 2 *} | Igor Grossmann¹ | Amir-Hossein Karimi^{1, 3}

¹ University of Waterloo ² Mila - Quebec AI Institute ³ Vector Institute for Artificial Intelligence

* Contact: eugeneyuji1@gmail.com

World models can project plausible futures. Metacognitive self-governance begins when such models can determine whether, how, and under what conditions those projections can and should guide action.

Prediction alone does not warrant action

A projected trajectory may be:

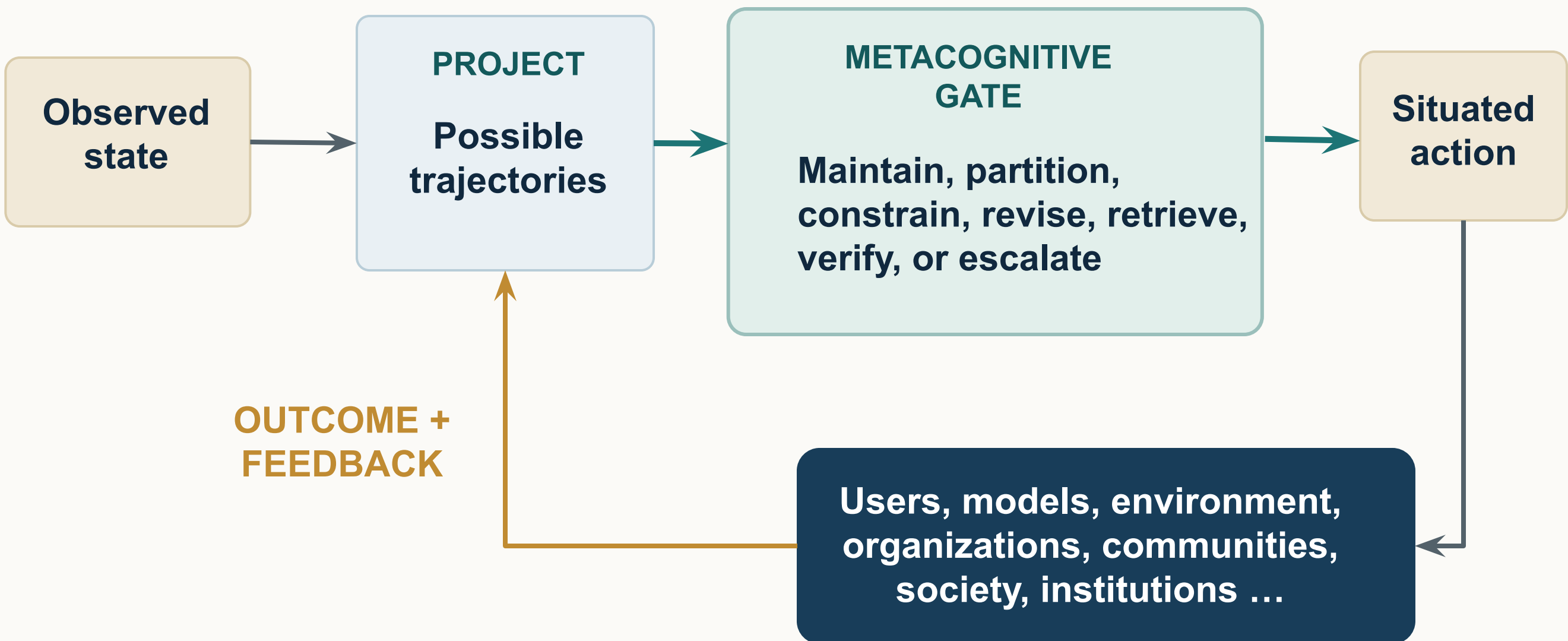
- plausible without being warranted
- useful without being feasible
- technically valid without practical force

METACOGNITIVE ALIGNMENT

The capacity to keep projected action trajectories responsive and adaptable to feedback, uncertainty, and revision.

A trajectory becomes actionable through a metacognitive self-governing loop

Metacognition navigates the transition from prediction to situated action.



World Model Metacognitive Alignment at Three levels (inspired by Marr, 1982)

	GUIDING QUESTION	MODE OF ALIGNMENT	DESIGN OUTCOME
COMPUTATIONAL	What makes a projected trajectory governable?	Monitor Evaluate Control Adapt	Criteria for action-readiness: which grounds, limits, and uncertainties matter for this use.
ALGORITHMIC	How do regulatory signals reshape trajectory generation?	Modularize Elicit Iterate	Signals alter the path: maintain, partition, constrain, revise, retrieve, verify, or escalate.
ECOLOGICAL	When does a trajectory acquire practical force?	Interface Workflow Accountability	Use becomes interpretable, contestable, revisable in context.

What fails when the levels do not align

- 01

Rich simulation, vague target
Many simulations are generated, but no criterion determines which uncertainty matters to them and why.
- 02

Correct target, weak signal path
Confidence or critique is produced, yet it never changes generation or action.
- 03

Strong procedure, brittle ecology
Users over-trust the output and misunderstand uncertainty and/or context.
- 04

Responsible process, unclear trigger
Review exists without principled or operationalized thresholds to qualify, revise, stop, or escalate.

Research agenda

- Cross-level diagnostics**
Coordinated benchmark across levels alongside task performance.
- Distributed governance**
Incorporate heterogeneous tools, humans, institutions, and higher-order feedback.
- Capability with governance**
Treat task performance and governance as mutually enabling.

Metacognition turns *humility about uncertainty and limitations, sensitivity to situational context, and the capacity to shift between its own perspective and those of other agents* into explicit governing resources for the world model.
Johnson, S. G. B., Karimi, A.-H., Bengio, Y., Chater, N., Gerstenberg, T., Larson, K., Levine, S., Mitchell, M., Rahwan, I., Schölkopf, B., & Grossmann, I. (2026). "Imagining and building wise machines: The centrality of AI metacognition." *Trends in Cognitive Sciences*. Advance online publication.

THE EVALUATION TARGET

Can a world model govern the exercise of its own predictive capacity?