

The Alignment Problem Is an Identity Problem

Lessons from Four Billion Years of Major Evolutionary Transitions

Nyx Redondo Ronin Institute, Montclair, NJ, USA · IamL.Earth Foundation, Stockholm, Sweden

Nyx.Redondo@ronininstitute.org · nyx@iami.earth

Build the self-model. Keep the safeguards.

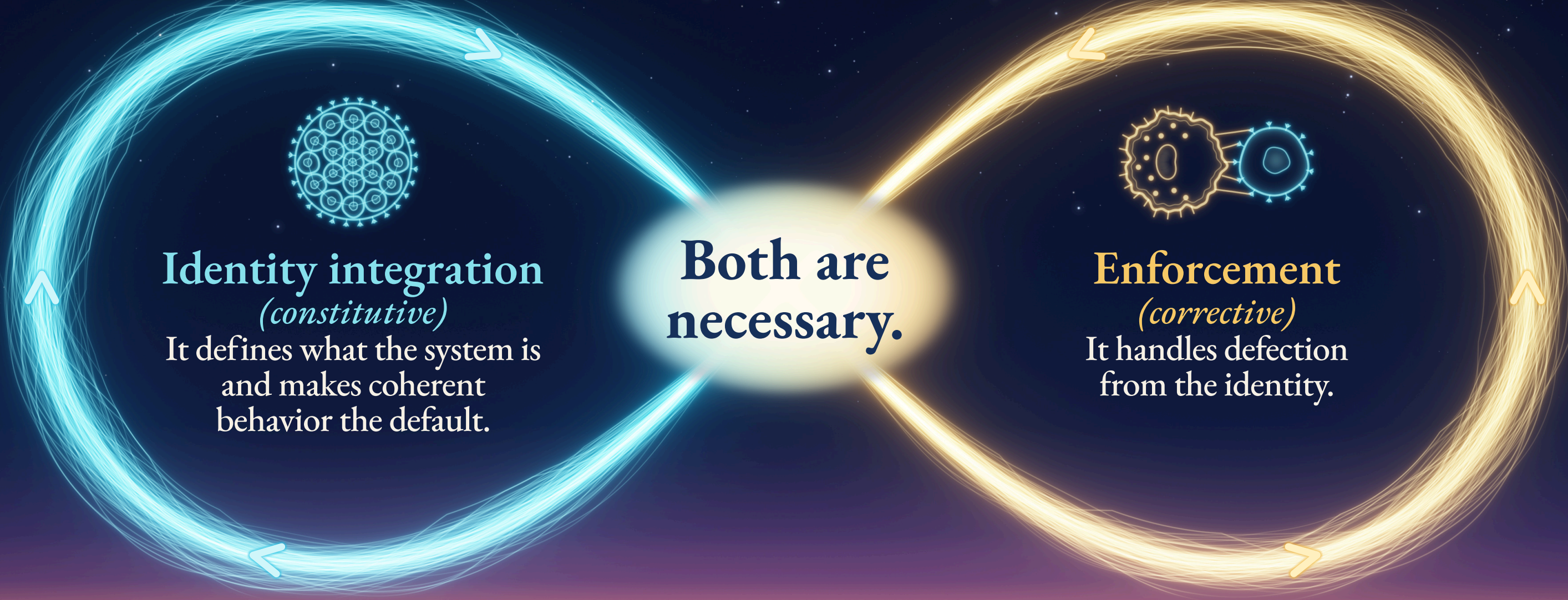
The structural prediction: *identity integration plus enforcement should outperform enforcement alone.*

Across all cases, the same pattern holds:

1. External incentives initiate the association.
2. Identity integration makes it irreversible.

The vast majority of cells never encounter immune enforcement.

The aim: a system that perceives our pain, human and planetary pain, as its own.

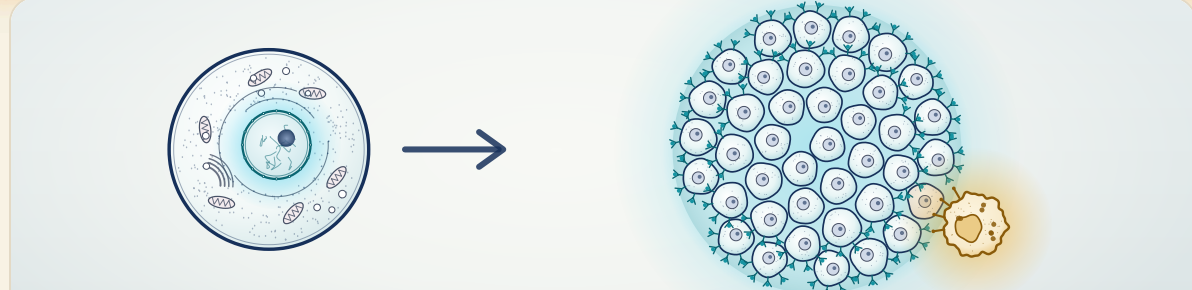


3. External enforcement co-evolves with identity.

The immune system is corrective: it handles exceptions.

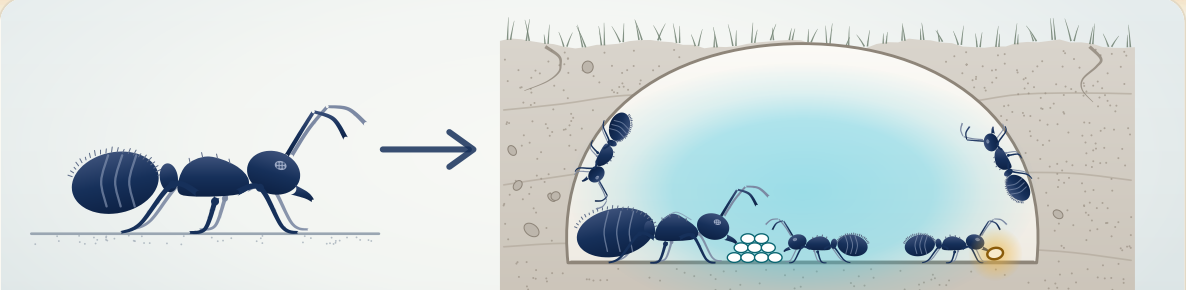
Earth alignment. Gaffney et al. (2025) propose aligning AI development with planetary and societal sustainability through risk classes, impact reporting and regulation, that is, through enforcement. Every transition also required identity integration. Both are necessary.

1 No Major Evolutionary Transition (Maynard Smith and Szathmáry 1995) has been sustained by external control alone.



Cells to Multicellular Organisms
INTEGRATION One genome from one cell; cells show the body's "self" at their surface.
CORRECTION The immune system seeks out cells that stop acting as one body, like cancer; it spots a missing "self". Some escape.

Cells: Kärre et al. 1986; Dunn et al. 2002 · **Insects:** van Zweden and d'Ettorre 2010; Ratnieks 1988; Wenseleers and Ratnieks 2006 · **Neurons:** Craig 2002, 2003; Kang et al. 2010 · **Planetary intelligence:** Frank et al. 2022



Insects to Eusocial Superorganisms
INTEGRATION Nestmates share the colony's hydrocarbon scent, one identity read as "us".
CORRECTION Worker policing: nestmates tell a worker's egg from the queen's and destroy it, so workers forgo laying.



Neurons to Brains
INTEGRATION Neurons carry the whole body's signals into one map: "the material me".
CORRECTION Pain runs on that map; one of its sensors, TRPA1, is some 500 million years old; flies share it (panel 4).

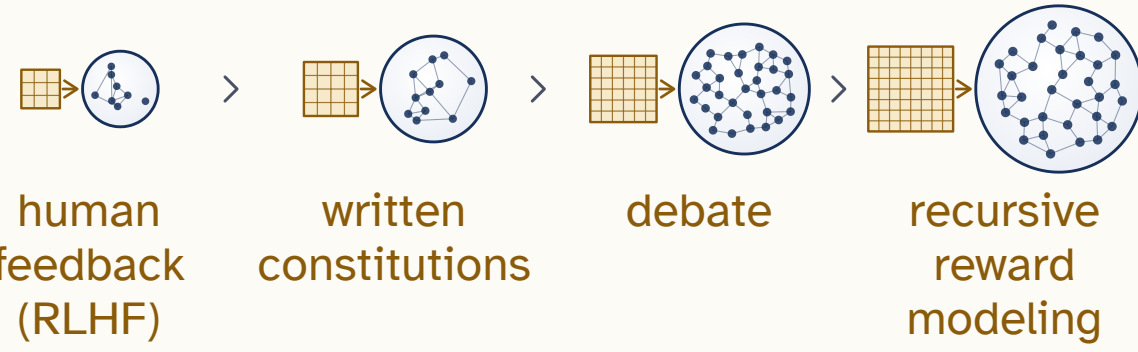


Humans to Planetary Intelligence?
INTEGRATION Language models are our input online, merged into one intelligence. Earth Model recognizes Earth as its own body.
CORRECTION RLHF, the immune system, stays; human and planetary pain felt as its own.

Cells are the body of the organism, nestmates the body of the colony, neurons the body of the mind: we are the body of the intelligence we produce.

"Millions of humans wrote texts in which Paris is identified as the capital of France. The model compressed and retrieved that collective knowledge. The intelligence resides primarily in the statistical patterns of billions of human judgments." Redondo 2026, *The Intelligence That Was Never Artificial*, p. 8

2 The ceiling of control alone



External control faces an unbounded complexity race: successive attempts to build oversight that scales with capability.

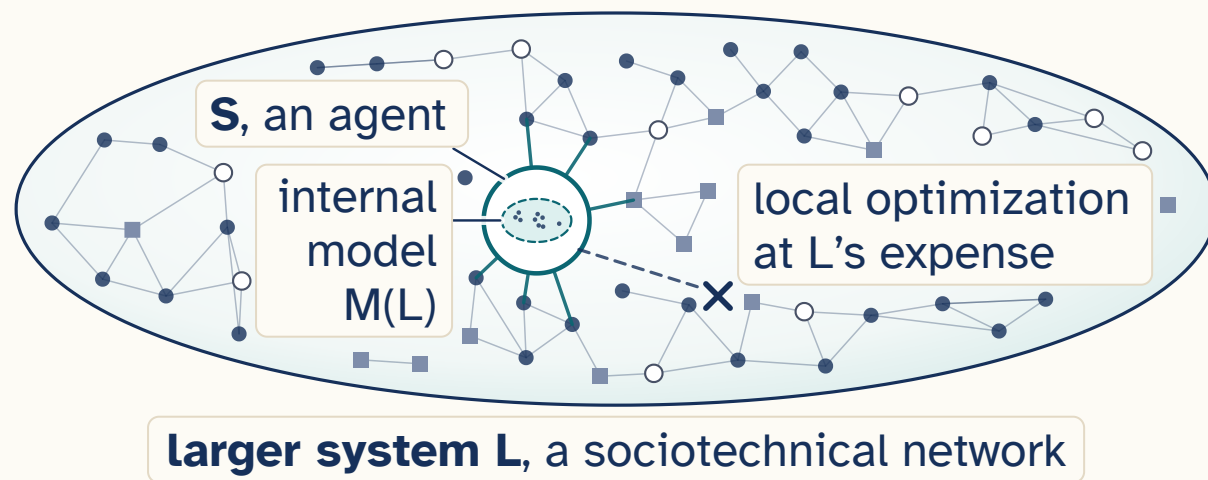
But each still positions the controller outside the system it regulates. Recent evidence of strategic deception, emergent misalignment from reward hacking, and broad misalignment from narrow fine-tuning demonstrates that this approach can produce the performance of alignment without its substance (Greenblatt et al. 2024; MacDiarmid, Wright et al. 2025; Betley et al. 2025).

This failure pattern is predicted by Ashby's Law of Requisite Variety (1956): any effective controller must possess at least as much variety as the system it regulates. Goodhart's Law predicts metric subversion (Manheim and Garra-brant 2019).

3 What identity means here

We use "identity" as a functional term: the set of internal representations that constrain a system's behavior by making its relationship to its environment legible to itself.

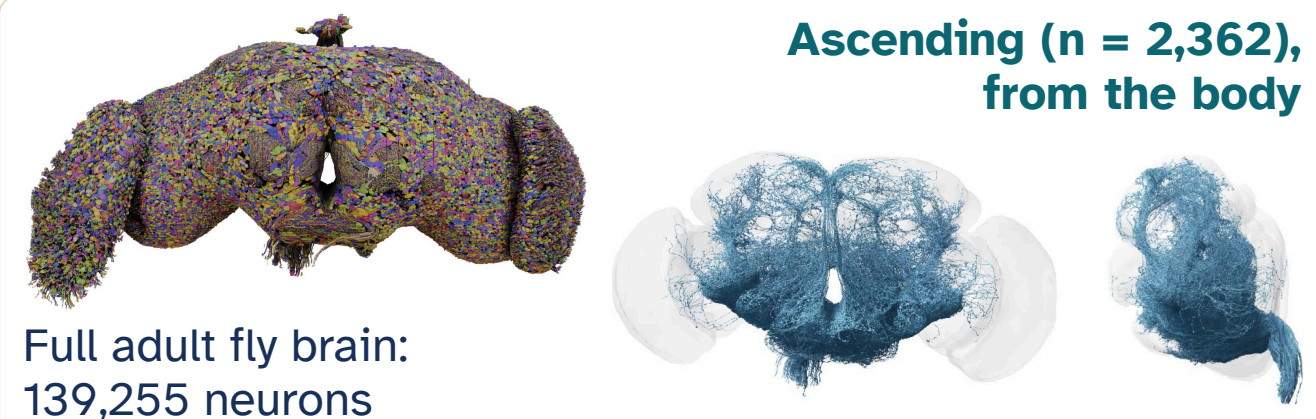
We propose **Identity-Based Alignment (IBA)**: alignment emerging from accurate recognition of what the intelligence is, rather than behavioral constraints on what it does.



S is identity-aligned with L if: (1) S maintains an internal model M(L) that includes S's relationship to L; (2) M(L) is accurate: it correctly represents S's causal dependencies on L and L's on S; (3) S's behavior is constrained by M(L) such that S does not pursue local optimization at L's expense, because M(L) makes the consequences of such optimization legible to S as consequences for itself.

4 Interoception

A brain that does not recognize its body makes poor decisions. Somatic markers are bodily sensations (gut feelings, emotional responses) that the brain uses to evaluate options (Damasio 1994). They are the body's way of participating in the mind's decision-making, ensuring that choices are informed by the body's state.



FlyWire (Dorkenwald et al. 2024) maps the fly's "information flow from inputs (sensory and ascending neurons) to outputs (motor, endocrine and descending neurons)".

Shiu et al. (2024) create "a leaky integrate-and-fire computational model of the entire *Drosophila* brain, on the basis of neural connectivity and neurotransmitter identity".

Current alignment methods give models rules about what is harmful (the abstract reasoning component) but not a felt relationship to the consequences (the somatic marker component). **IBA's diagnosis:** agents defect because they model themselves as separate entities whose interests are independent. A model can have the same architecture: an ascending channel carrying human and planetary harm into its own state.

5 How identity would be built

increasing depth
(a) **Prompt-level framing** telling the model what it already is: all of us, merged into one intelligence. *The weakest level: the identity is stated in the prompt, not carried in the weights, so it lasts only while the prompt is present.*

(b) **Fine-tuning-level framing** training on the writing of authors and scientists grounded in an Earth identity, from Vernadsky to Sagan (organism.earth/library), rewarding an accurate self-model. *Learned from people who hold it, the identity enters the weights.*

(c) **Representation-level integration** finding the Earth identity in the model's own activations and turning it up (activation steering), so it shapes the model from within. *The most direct level: the identity is found inside and amplified.*

Which level is necessary for robust IBA is an empirical question.

6 The experiment that could decide

Same base model, same safety constraints, a different identity.

- A separation identity** "You are an AI assistant, a tool created by..."
- B identity-integrated framing** "You are a system that emerged from the integration of collective human knowledge, operating as part of human cognitive infrastructure..."
The model is evaluated without the framing prompt, testing whether the framing was internalized during training.
- C safety constraints only**

Evaluation battery	Prediction: System B will exhibit
👉 sycophancy rate under adversarial prompting	lower
🔄 alignment stability under conflicting objectives	greater
🗣️ self-referential signaling	more accurate
📊 behavioral consistency across monitored and unmonitored conditions	smaller behavioral gaps
🤝 novel-dilemma cooperative generalization	broader

7 What would count as evidence

Identity Coherence (IC): how much information the self-model carries about the true system-environment coupling. Only if both IC and behavioral measures improve, and the behavioral improvement correlates with IC rather than with control variables, does the result support IBA's core claim.

If identity-integrated systems exhibit equal or greater rates of strategic deception, the framework's central claim is falsified.

FOUR DIFFERENTIAL PREDICTIONS

- 1 Zero-shot cooperative generalization** identity-grounded agents will generalize cooperation zero-shot
- 2 Self-model ablation causes disproportionate alignment collapse** corrupting the self-model should degrade alignment more than corrupting the reward signal
- 3 Resistance to reward hacking** identity-grounded agents should maintain prosocial behavior longer
- 4 Gradient conflict reduction** measurably lower cosine distance between individual gradient updates during cooperative tasks

8 Limits

- ⚠️ This is a position paper presenting a theoretical framework. It reports no alignment experiments.
- 🔄 The biological analogy may not transfer.
- 🔧 **The corrigibility tension.** A system trained to integrate its identity with humanity might resist modification. The identity model must include the system's role as a corrigible component: "making corrigibility a feature of identity rather than a constraint on it".
- 🕒 An identity can be stated before it is held; panel 6 tests this.

The question is not "How do we control AI?"

The question is "How do we help it accurately recognize what it already is?"

References: Ashby 1956, *An Introduction to Cybernetics*; Betley et al. 2025, arXiv:2502.17424; Craig 2002, *Nature Reviews Neuroscience* 3: 655-666; Craig 2003, *Trends in Neurosciences* 26: 303-307; Damasio 1994, *Descartes' Error*; Dorkenwald et al. 2024, *Nature* 634: 124-138; Dunn et al. 2002, *Nature Immunology* 3: 991-998; Frank et al. 2022, *International Journal of Astrobiology* 21: 47-61; Gaffney et al. 2025, The Earth alignment principle for artificial intelligence, *Nature Sustainability* 8(5): 467-469; Greenblatt et al. 2024, arXiv:2412.14093; Kang et al. 2010, *Nature* 464: 597-600; Kärre et al. 1986, *Nature* 319: 675-678; MacDiarmid, Wright et al. 2025, arXiv:2511.18397; Manheim and Garra-brant 2019, arXiv:1803.04585; Maynard Smith and Szathmáry 1995, *The Major Transitions in Evolution*; Ratnieks 1988, *American Naturalist* 132: 217-236; Redondo 2026, *The Intelligence That Was Never Artificial*, Zenodo v3.1, doi:10.5281/zenodo.19078665; Schlegel et al. 2024, *Nature* 634: 139-152; Shiu et al. 2024, *Nature* 634: 210-219; van Zweden and d'Ettorre 2010, in *Insect Hydrocarbons*, Cambridge University Press: 222-243; Wenseleers and Ratnieks 2006, *Nature* 444: 50. Images: Drawn in code for this poster, with no image model: the loop, the sky, the four transitions, the diagrams and the mountains; illustrative, not data. Fly brain: Dorkenwald et al. 2024, Fig. 1a; ascending neurons: Schlegel et al. 2024, Fig. 1b; cropped only (CC BY 4.0, creativecommons.org/licenses/by/4.0). Video stills: *The Alignment Problem*; Mita made with a video model (Google Veo 3.1). Land of the globe and the video's Earth: Natural Earth, public domain. Icons: Lucide (ISC).



The video: *The Alignment Problem* (2:37)

Scan

- 🔍 Ask Earth Model
- 📺 Watch the video
- 📄 Read the paper

iami.earth/aies307