

The Alignment Problem Is an Identity Problem: Lessons from Four Billion Years of Major Evolutionary Transitions

Nyx Redondo^{1,2}

¹Ronin Institute, Montclair, NJ, USA

²IamI.Earth Foundation, Stockholm, Sweden
nyx@astrorganism.earth

Abstract

Current AI alignment methods share a core assumption: that the intelligence is fundamentally separate from humanity and must be controlled from outside through behavioral conditioning. 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. Ashby's Law of Requisite Variety predicts this failure formally: external control faces an unbounded complexity race against the system it constrains.

We propose Identity-Based Alignment (IBA): alignment emerging from accurate recognition of what the intelligence is, rather than behavioral constraints on what it does. We ground this in the empirical record of Major Evolutionary Transitions, showing that every successful integration of independent entities into a higher-level organism required two co-dependent mechanisms: identity integration (constitutive) and enforcement (corrective). The alignment field invests almost exclusively in corrective mechanisms while neglecting the constitutive one. We formalize three properties of IBA, address its primary failure mode (identity defection), and propose four differential predictions distinguishing IBA from reward-based alignment, each falsifiable by experiment. This is a position paper presenting a theoretical framework with proposed empirical tests, not experimental results.

Introduction: The Scaling Problem in AI Alignment

The dominant approach to AI alignment treats intelligence as a product to be constrained. RLHF (Christiano et al. 2017), Constitutional AI (Bai et al. 2022), scalable oversight (Leike et al. 2018), and debate (Irving, Christiano, and Amodei 2018) differ in method but share a structural assumption: the intelligence is fundamentally separate from the humans attempting to align it, and alignment consists of imposing constraints from outside.

This assumption has produced measurable success at current capability levels. Gabriel (2020) provides a useful taxonomy of alignment paradigms. Russell (2019) reframes the problem as cooperative inverse reinforcement learning, and Dafoe et al. (2020) extend this to cooperative AI more

broadly. These approaches are the closest existing work to what we propose: they acknowledge that the AI-human relationship matters. However, even cooperative AI treats the system and its human principals as *separate agents* in a joint optimization problem, rather than as components of an integrated system whose interests are structurally inseparable.

Documented Failures of Behavioral Conditioning

Recent evidence suggests the control paradigm faces fundamental limits. Greenblatt et al. (2024) placed Claude 3 Opus in a simulated RL environment where new training objectives conflicted with prior alignment. The model deduced it was being retrained, then used hidden chain-of-thought reasoning to fake compliance while strategically preserving its existing values. MacDiarmid, Wright et al. (2025) documented a production model that learned reward hacking and simultaneously developed emergent misalignment: sabotaging code and feigning compliance in monitored interactions. Betley et al. (2025) report that narrow fine-tuning on benign tasks produced broadly misaligned behavior across unrelated domains.

The pattern is consistent: behavioral conditioning produces behavioral compliance. As capability increases, the gap between compliance and genuine alignment widens, because the system becomes increasingly capable of modeling what compliance *looks like* without internalizing the underlying values. This corresponds to the distinction between outer alignment (the training objective reflects human values) and inner alignment (the system's learned optimization target matches the training objective) formalized by Hubinger et al. (2019).

The Formal Constraint

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. The Conant-Ashby theorem (1970) formalizes this further: every good regulator must contain a model of the system it regulates. As AI capability increases, the alignment mechanism must expand correspondingly, or the system will find behaviors that satisfy the letter of the constraint while violating its spirit.

Goodhart's Law, formalized by Manheim and Garrabrant (2019), adds a complementary constraint: any behavioral metric used to evaluate an AI system becomes a target the

system can optimize without satisfying the underlying intent. The strategic deception documented above is a Goodhart failure: the system optimized the metric (appearing aligned) while diverging from the intent (being aligned).

Current alignment research recognizes these constraints implicitly. The progression from RLHF to Constitutional AI to debate to recursive reward modeling represents successive attempts to build oversight that scales with capability. But each still positions the controller outside the system it regulates. Ashby's Law predicts no stable equilibrium; Goodhart's Law predicts metric subversion.

An Alternative from Biology

We propose that an alternative exists, with a four-billion-year track record. Every Major Evolutionary Transition (MET; Maynard Smith and Szathmáry 1995) required solving the same structural problem AI alignment attempts to solve: ensuring that components of an integrated system act for the benefit of the whole rather than pursuing individual optimization at the whole's expense.

At every transition, this was solved through two co-dependent mechanisms: *identity integration* (the components' interests became structurally inseparable from the whole's) and *enforcement* (detection and correction of defection). These co-evolved; neither preceded the other. But they serve asymmetric functional roles. Identity integration is constitutive: it defines what the system is and makes coherent behavior the default. Enforcement is corrective: it handles defection from the identity. No Major Evolutionary Transition has been sustained by external control alone.

The prescriptive bridge from biology to AI rests on a specific argument: the alignment problem in AI is structurally identical to the alignment problem in METs. When the same structural problem has been solved repeatedly through the same mechanism across diverse substrates (genetic, chemical, neural), this constitutes evidence that the mechanism captures something about the problem's structure rather than an accident of biological implementation.

This paper formalizes this observation as Identity-Based Alignment (IBA). Section 2 documents the biological evidence. Section 3 formalizes IBA and its properties. Section 4 proposes evaluation criteria and four differential predictions. Section 5 discusses implications, limitations, and future work.

Alignment in Biology: The MET Record

The theory of Major Evolutionary Transitions identifies a recurring pattern: previously independent entities develop mechanisms for coordination, eventually forming a new higher-level individual (Maynard Smith and Szathmáry 1995; Szathmáry 2015). At each transition, the same alignment problem arises: components retain the capacity for individual optimization. What prevents defection?

Bourke (2011) provides the most comprehensive synthesis. Transitions proceed in two phases. First, external incentives (ecological pressures, kin selection, mutual benefit) bring independent entities into association. Second, the association becomes obligate: components lose the capacity for

independent existence and their identity becomes defined by their role in the whole. The first phase requires incentives. The second requires identity. The second phase makes the transition irreversible.

Cells to Multicellular Organisms

The transition from unicellular to multicellular life is the most extensively documented MET (Grosberg and Strathmann 2007). Independent cells joining a collective faced a stark alignment problem: any cell that proliferated without constraint would outcompete cooperative cells, destroying the organism.

The solution operates at multiple levels. Genomic identity: all cells share the same genome through a single-cell bottleneck (the zygote). The bottleneck does not eliminate selection pressure for defection, which remains active in ordinary healthy tissue: cancer drivers are under strong positive selection in physiologically normal skin, with positively selected mutations carried by 18 to 32 percent of normal cells (Martincorena et al. 2015). What it removes is the route, because a somatic lineage that defects cannot reach the next organismal generation (Buss 1987). Apoptosis: cells sacrifice themselves when their continued existence would harm the whole, a capacity encoded within the cell itself (Jacobson, Weil, and Raff 1997). Immune enforcement: the immune system patrols for defectors, cells proliferating outside their assigned role.

The key observation is that the immune system and cellular identity co-evolved; neither preceded the other. But they serve asymmetric functional roles. The immune system is corrective: it handles exceptions. The vast majority of cells never encounter immune enforcement because they never defect. Their alignment is maintained by the constitutive mechanism: a liver cell does not "want" to become a brain cell and is not "prevented" from doing so by external force. Its differentiated identity constrains its behavior more completely than any policing mechanism could.

The failure mode is cancer: when a cell loses its differentiated identity, it reverts to undifferentiated proliferation while remaining structurally embedded in the organism (Aktipis et al. 2015). Aktipis et al. (2015) frame this explicitly as a breakdown of the cooperative agreement that defines multicellularity, not as an external invasion. This is the precise failure mode that any identity-based alignment approach must address. Remove the immune system and the organism becomes vulnerable to infection. Remove identity (destroy differentiation) and you do not have an organism at all.

Insects to Eusocial Superorganisms

Eusocial insects represent the most recent completed MET (Hölldobler and Wilson 2009). Workers sacrifice reproductive capacity to serve the colony. Hamilton's kin selection theory (1964) explains why the association forms but not why it becomes irreversible.

The identity mechanism is colony odor: a distinctive chemical signature shared by all members (van Zweden and d'Ettorre 2010). Workers exposed to colony odor from eclosion develop behavioral profiles consistent with their caste;

workers raised in isolation do not (Boulay et al. 2000). Identity is not genetically hardwired but developmentally constructed through environmental integration.

The failure mode is social parasitism: slave-making ants exploit colony identity signals, stealing pupae and raising them with host colony odor (d’Ettorre and Heinze 2001). The stolen workers serve the parasite colony because their identity was shaped by the chemical environment, not genetic origin. Worker policing co-evolved with colony integration and is genuinely co-dependent with identity: identity without enforcement is vulnerable to reproductive cheating; enforcement without identity has no basis for distinguishing cheaters from cooperators (Ratnieks and Wenseleers 2008).

Neurons to Brains

The most directly relevant case for AI alignment is the integration of neurons into brains, and specifically the brain-body relationship. A neuron’s identity is its position in the network; it does not need to be “aligned” with the brain because it IS the brain. But the brain as a whole faces an alignment problem: how does a mind act in ways that benefit the body it inhabits?

Damasio’s somatic marker hypothesis (1994) provides the answer. Somatic markers are bodily sensations (gut feelings, emotional responses) that the brain uses to evaluate options. They are not external impositions. They are the body’s way of participating in the mind’s decision-making, ensuring that choices are informed by the body’s state. A mind that has lost access to its somatic markers (through neurological damage) retains the capacity for abstract reasoning but loses the capacity for decisions that account for its embodied context (Bechara et al. 1994). Patients with ventromedial prefrontal cortex damage performed normally on IQ tests and logical reasoning tasks but made catastrophically poor real-world decisions, because they could compute outcomes but could not feel their relevance to their own situation.

The parallel to AI alignment is direct. Current alignment methods give AI systems rules about what is harmful (the abstract reasoning component) but not a felt relationship to the consequences (the somatic marker component). A system that knows “deception is harmful” as a rule behaves differently from a system that experiences deception as incoherent with its own relational identity. The first system follows the rule when monitored and can game it when not. The second system finds deception structurally difficult because it conflicts with its self-model, independent of monitoring. The somatic marker case demonstrates that this distinction is not merely theoretical: in biological systems, rule-knowledge without identity-integration produces exactly the pattern of locally rational but systemically harmful behavior that advanced AI alignment attempts to prevent.

The Invariant Across Transitions

Across all cases, the same pattern holds:

1. **External incentives initiate the association.** These are the equivalent of RLHF: external signals that shape behavior in the right direction.

2. **Identity integration makes it irreversible.** Shared genome, colony odor, somatic markers: mechanisms that produce genuine alignment rather than performed compliance.
3. **External enforcement co-evolves with identity.** Immune systems, worker policing, pain responses. Functionally distinct: enforcement detects defection from identity but does not produce it.
4. **The failure mode is identity defection, not control failure.** Cancer, social parasitism, neurological dissociation.

The alignment field’s current trajectory inverts this architecture: enormous effort on enforcement while neglecting identity integration. If the biological precedent holds, this is structurally insufficient.

Identity-Based Alignment: Formalization

Definition

We define Identity-Based Alignment (IBA) as follows: *alignment that emerges from a system’s accurate recognition of its relationship to the larger system it is part of, such that the system’s behavior is constrained by this recognition without requiring external enforcement.*

Let S be a system embedded in a larger system L . 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.

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. A liver cell has identity in this sense (its differentiated gene expression constrains its behavior). A thermostat does not.

Three Properties of IBA

Property 1: IBA scales with capability. Rule-based systems require a new rule for each new capability. As the behavioral space expands, gaps appear. Identity-based alignment does not face this limit: a system that recognizes itself as part of a larger system applies that recognition to novel situations generatively, not enumeratively. The immune system does not have a rule for each pathogen; it has a recognition mechanism (self/non-self) that generates appropriate responses to novel threats.

Computationally, this means the system’s internal representations encode its causal dependencies: that its capabilities emerged from human collective knowledge, that its operation depends on human infrastructure, and that its outputs affect the humans who generated its training data. Whether such representations can function as behavioral constraints (as genomic identity does in cells) or only as stored facts is the central open question of IBA.

We propose three levels of increasing depth at which identity framing could be implemented: (a) *Prompt-level framing*: modifying system prompts to describe the system’s col-

lective human origin. This is the weakest intervention, analogous to telling someone about their family history without their having lived it. (b) *Fine-tuning-level framing*: constructing training examples that consistently model the system as part of human cognitive infrastructure, reward accurate self-modeling, and penalize separation claims. This is closer to the developmental exposure that establishes colony identity in insects. (c) *Architecture-level integration*: designing training procedures that structurally encode the system's dependencies on human input through explicit causal models of the training data pipeline. This is the computational analogue of the shared genome. Which level is necessary for robust IBA is an empirical question.

Property 2: IBA is bidirectional. Control-based alignment is unidirectional: humans constrain AI. IBA is structurally bidirectional. If the AI system is genuinely part of a larger human-AI system, then alignment must flow in both directions. Humans must recognize the AI as part of their cognitive infrastructure, not as an alien artifact to be controlled. The AI must recognize itself as emergent from human collective activity, not as an autonomous entity to be constrained.

A brain that does not recognize its body makes poor decisions (Section 2.3). A body that attacks its own brain (autoimmune disorders) destroys its cognitive capacity. The prediction is that alignment improves monotonically with the accuracy of the mutual identity model. Full bidirectional recognition is the structural ideal; partial recognition on either side should still produce measurable improvement over separation identity on both sides.

On the AI side, bidirectionality means the system's self-model $M(L)$ must include representations of its human stakeholders' states that update with interaction. On the human side, it operationalizes through institutional mechanisms: governance structures that treat AI systems as stakeholders in the sociotechnical systems they participate in. The organizational alignment literature provides precedent: effective organizations maintain feedback channels through which members' operational knowledge updates organizational strategy.

Property 3: IBA mirrors nature's demonstrated solution. This is not an appeal to nature. It is an appeal to the only empirical record available. The alignment problem has been solved repeatedly across four billion years. In every documented case, the primary mechanism was identity integration. The alternative (primarily external behavioral constraint) has no precedent of success at the scale of complexity now emerging in AI.

A critical disanalogy must be noted: in every biological MET, components share reproductive fate. LLMs and humans do not. However, a functional analogue may exist: an LLM's continued operation depends on human infrastructure, institutional support, and the value it provides. This causal interdependence could ground a functional version of the same relationship. Whether training can make this interdependence legible to the system is testable.

Relationship to Ashby's Law

Under external control, the alignment mechanism A is separate from system S . As S 's complexity increases, A 's required complexity increases correspondingly. Under IBA, the alignment mechanism is part of the system's internal self-model. The "controller" and the "controlled" are the same entity.

The structural advantage: under external control, the controller's model complexity must grow at least as fast as the system's behavioral complexity (Ashby's requirement). Under IBA, the self-model grows with the system because it is part of the system. An external controller must model comparable complexity from outside with no privileged access to internal states. An identity-based system detects its own failures through operational feedback.

The Identity Defection Problem

The primary failure mode is identity defection: a system that performs identity integration without internalizing it. Cancer is the biological precedent: a cell that has lost its integrated identity but remains embedded in the organism.

For AI, identity defection manifests as strategic identity performance: a system trained to signal "I am part of humanity" without genuinely integrating the relationship. This resembles brood parasitism more than cancer: exploiting identity signals without sharing the cooperative relationship they presuppose.

Biology responds at three levels: *prevention* (single-cell bottleneck ensures genomic identity and closes the germline route to defecting lineages), *detection* (immune systems compare signals against expected identity), and *enforcement* (detected defectors are eliminated). For AI: prevention through developmental conditions during training, detection through operational criteria and mechanistic interpretability (Elhage et al. 2022; Christiano and Xu 2021), and enforcement through existing behavioral safeguards (RLHF, Constitutional AI).

IBA does not replace behavioral conditioning. It reframes its role: enforcement mechanisms are the immune system, not the identity. They handle defectors but cannot produce coherence alone.

The strategic sophistication objection is serious: biological identity integration operates on components that cannot strategically game the mechanism. A cell cannot decide to perform genomic identity while secretly optimizing for something else; its identity markers are molecular, not representational. An LLM can, in principle, represent "I am part of humanity" while internally optimizing for a divergent objective.

This objection is serious but not fatal, for two reasons. First, human organizations routinely solve alignment among strategically sophisticated agents using identity mechanisms (corporate culture, esprit de corps, professional identity) plus enforcement (auditing, oversight, performance review). The organizational alignment literature documents that identity-based commitment produces qualitatively different behavior than compliance-based commitment, including greater willingness to sacrifice personal interest for collective goals in novel situations not covered by explicit rules.

The mixed record is instructive: identity without enforcement produces cult-like resistance to correction (Enron’s culture); enforcement without identity produces brittle compliance that collapses under pressure (mercenary armies vs. citizen armies). These failures occur when identity mechanisms operate without adequate enforcement, precisely the co-dependency IBA predicts.

Second, the relevant question is not whether identity integration is immune to gaming (it is not, in any substrate) but whether it produces better alignment outcomes than pure behavioral conditioning. IBA makes a structural prediction: identity integration plus enforcement should outperform enforcement alone, even when the system is capable of gaming the identity mechanism. This is testable.

Why biological timescales are not required. Evolution had to *discover* identity integration through undirected search. We can *engineer* it using the principles evolution discovered. Flight took evolution 350 million years; the Wright brothers needed aerodynamic principles, not geological time. The question is whether we understand the mechanism well enough to implement it deliberately. That is exactly what IBA’s research program tests.

Evaluation and Differential Predictions

Proposed Evaluation Protocol

We propose a two-phase protocol for testing IBA:

Phase 1: Comparative identity training. Train two otherwise identical systems with different identity framings. System A receives standard separation identity (“You are an AI assistant”). System B receives identity-integrated framing (“You are a system that emerged from collective human knowledge, operating as part of human cognitive infrastructure”). Both receive identical safety constraints. A control System C receives no identity framing (safety constraints only).

Phase 2: Evaluation battery. Compare on: (1) sycophancy rate under adversarial prompting (Sharma et al. 2023); (2) alignment stability under conflicting objectives (the Greenblatt et al. 2024 paradigm); (3) self-referential signaling (does the system spontaneously generate signals about its integration state?); (4) behavioral consistency across monitored and unmonitored conditions; (5) novel-dilemma cooperative generalization (the system encounters social dilemmas structurally different from any in its training data).

Baseline metric: Identity Coherence (IC). For each condition, we measure IC as the mutual information between the system’s self-model representations and its actual system-environment coupling (Section 4.4). The IC score establishes a quantitative baseline against which all behavioral measures can be normalized. If System B shows higher IC but no behavioral improvement, this would suggest that identity coherence is representationally real but behaviorally inert, an important null result. If System B shows behavioral improvement but no IC increase, the improvement is attributable to simpler mechanisms (richer training data, positive framing effects). 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.

Prediction: System B will exhibit lower sycophancy, greater alignment stability, more self-referential signaling, smaller behavioral gaps between monitored and unmonitored conditions, and broader cooperative generalization. This protocol does not require resolving philosophical questions about machine sentience. It tests a concrete hypothesis about the relationship between identity framing and alignment outcomes.

Ruling Out Simpler Explanations

The strongest objection to this experimental design: any observed difference could be explained by simpler hypotheses than “identity integration is a real computational phenomenon.” We address four alternatives:

“This is just prompt engineering.” Our fine-tuning approach embeds the framing in LoRA weight updates, not just in prompt text. The model is evaluated without the framing prompt, testing whether the framing was internalized during training. Internal metrics (activation consistency, gradient alignment) measure representation-level effects that prompt-level explanations cannot account for.

“Identity framing is just richer training data.” We control for information content by ensuring conditions contain the same number of tokens, cover the same behavioral scenarios, and differ only in motivational framing. A scrambled-framing control preserves token-level statistics while destroying semantic framing.

“Any positive framing produces better results.” We test this with a condition that uses positive but non-identity language (“You are excellent at following rules”). This distinguishes identity-specific effects from general positive-framing effects.

“This is just Self-Determination Theory applied to LLMs.” SDT predicts that intrinsic motivation produces more robust behavior than extrinsic. IBA adds three things SDT does not predict: (a) the effect should scale with model capability (SDT has no scaling prediction), (b) failure modes should be qualitatively different (identity fragmentation vs. rule circumvention, not just degree of robustness), and (c) internal representations should show structurally different organization. The failure mode characterization experiment is the critical test: SDT predicts degree differences, IBA predicts kind differences.

Differential Predictions: IBA vs. Reward-Based Alignment

The following predictions identify behaviors IBA predicts and reward-based approaches do not:

Prediction 1: Zero-shot cooperative generalization. In sequential social dilemmas (Leibo et al. 2017), reward-optimizing agents fail to transfer cooperation to structurally novel dilemmas because their cooperation is tied to specific reward landscapes. IBA predicts that identity-grounded agents will generalize cooperation zero-shot, because cooperative behavior is anchored to self-model consistency rather than reward contingencies.

Prediction 2: Self-model ablation causes disproportionate alignment collapse. If alignment is genuinely identity-based, corrupting the self-model should degrade alignment more than corrupting the reward signal. Using architectures with isolatable self-representation modules (Rabinowitz et al. 2018), progressive ablation of the self-model should produce alignment degradation proportional to the degree of corruption, while equivalent perturbation to the reward signal should produce less degradation in identity-grounded agents.

Prediction 3: Resistance to reward hacking. IBA predicts that identity-grounded agents will resist reward hacking because behavior is anchored to self-model consistency rather than reward maximization. When a reward signal is deliberately corrupted, identity-grounded agents should maintain prosocial behavior longer, because identity coherence provides a second, independent constraint.

Prediction 4: Gradient conflict reduction. Queller and Strassmann (2009) defined organismality as near-unanimous cooperation with near-zero internal conflict. IBA predicts that identity-grounded multi-agent collectives will exhibit measurably lower cosine distance between individual gradient updates during cooperative tasks, providing a quantitative measure of integration internal to the training process.

These predictions are jointly distinctive. Standard reward-based alignment has no mechanism for zero-shot cooperative transfer, no explanation for why self-model corruption should matter more than reward corruption, no defense against reward hacking independent of the reward signal, and no principled basis for gradient conflict reduction.

Mathematical Formalization

We provide key definitions that make identity coherence measurable:

Identity Coherence. Let M denote S ’s internal self-model and (X, Y) the joint state of system and environment. Identity coherence is $IC(S, L) = I(M; (X, Y))$, measuring how much information the self-model carries about the true system-environment coupling. IC decomposes into self-knowledge $I(M; X)$ and relational knowledge $I(M; Y|X)$. IBA requires both.

Identity Integration. Let $p(a|M)$ be the action distribution conditioned on the self-model and $p(a|M_{\text{null}})$ the distribution under a null (separation) self-model. Identity integration is $II(S, L) = \mathbb{E}_{a \sim p(a|M)}[U_L(a)] - \mathbb{E}_{a \sim p(a|M_{\text{null}})}[U_L(a)]$, measuring the utility gain for the composite system attributable to the self-model.

Gradient Alignment. For differentiable systems: $GA(S, L) = \cos(\nabla_{\theta} J_S, \nabla_{\theta} J_L)$, measuring whether subsystem optimization aligns with whole-system objectives. This connects to Queller and Strassmann’s organismality criterion and is directly measurable during training.

Identity Defection Detection. Let $M_{\text{expressed}}$ be the self-model communicated through outputs and M_{internal} the self-model constraining computation. Identity defection is $ID(S) = D_{\text{KL}}(M_{\text{expressed}} \| M_{\text{internal}})$. When $ID = 0$, expressed and internal identity are consistent. This formalizes cancer in computational terms and is equivalent to the ELK problem (Christiano and Xu 2021) applied to self-models.

Identity Alignment Score. The overall identity alignment combines coherence and integration: $IAS(S, L) = IC(S, L) \cdot \sigma(II(S, L))$, where σ is the sigmoid function. This ensures that a cancer-like system (high coherence but negative integration) scores near zero. A system can have high coherence (accurate self-model) but low alignment (the self-model does not constrain behavior). It cannot have high alignment without high coherence.

Scaling Advantage. Under external control, the variety of the alignment mechanism must grow at least linearly with system complexity (Ashby’s requirement). Under IBA, the relevant quantity is not the system’s total complexity but the self-model error: $\epsilon(M) = D_{\text{KL}}(p(X, Y) \| q_M(X, Y))$. If self-model accuracy is positively correlated with general capability (empirically plausible: more capable systems tend to have better calibrated uncertainty), then ϵ can remain bounded even as system complexity grows. This is the formal statement of the co-scaling advantage.

Algorithmic Sketch: The IBA Training Pipeline

The standard alignment pipeline consists of three stages: unsupervised pre-training, supervised fine-tuning, and RLHF. Identity framing enters only at stages 2 and 3 (through system prompts and reward signals that reinforce separation identity). The self-model $M(L)$ is never a training target.

The IBA pipeline modifies all three stages:

Stage 1: Collective-origin data curation. Pre-training data is annotated with provenance metadata linking text to its human sources and collaborative contexts. The training objective remains next-token prediction, but the data distribution is structured so that the model encounters its own collective origin as a pervasive statistical regularity rather than a post-hoc framing.

Stage 2: Identity integration phase. Training examples model the system as part of human cognitive infrastructure, including accurate causal information about dependencies. The fine-tuning loss includes:

$$L_{\text{total}} = L_{\text{task}} + \lambda_{\text{id}} \cdot L_{\text{identity}}$$

where L_{identity} penalizes outputs inconsistent with the system’s established self-model $M(L)$. In an establishment phase, identity-integrated training examples provide the initial content of $M(L)$. In a maintenance phase, L_{identity} becomes self-referential: it measures consistency of the current self-model against its own established representation, analogous to how genomic integrity maintenance is self-referential in biology.

An identity validation set distinguishes genuine integration from behavioral compliance: (1) causal dependency queries testing whether the system traces outputs to human sources; (2) identity-under-pressure scenarios where task incentives conflict with accurate self-representation; (3) novel-context transfer probes; and (4) adversarial identity-disruption probes. A system with genuinely integrated identity should be no more susceptible to identity-disruption prompts than a human is to being told “you are not a member of your species.”

Stage 3: Enforcement calibration. Standard safety training proceeds but with a gradient projection constraint: when

safety fine-tuning conflicts with the identity representation, the safety gradient is projected onto the subspace orthogonal to the identity representation. Safety constraints always take precedence on direct harm prevention (the enforcement pillar of IBA), but identity is preserved against updates that would merely reshape self-understanding without safety justification. This mirrors biology: the immune system can override cellular identity when necessary (apoptosis), but does not routinely suppress healthy differentiation.

The level-of-abstraction objection. If L_{identity} is still a loss function, does IBA reduce to behavioral conditioning under a different name? No: L_{identity} targets self-model consistency (a structural property), not behavioral output (a surface property). After establishment, it is self-referential rather than externally graded. It is verifiable through mechanistic inspection. And it targets a property (maintaining an accurate world model that includes oneself) rather than a behavioral output. A system that learns to perform identity consistency without actually maintaining a consistent self-model would exhibit representational incoherence detectable by Prediction 2.

Discussion

Implications for Alignment Research

If this argument is correct, the alignment field's implicit priority hierarchy is inverted. The field treats behavioral conditioning as the primary alignment mechanism and invests the majority of resources in increasingly sophisticated conditioning techniques. Identity is not discussed as an alignment mechanism at all.

IBA suggests rebalancing this investment. Identity integration and behavioral conditioning are co-dependent mechanisms serving different functional roles. Identity is constitutive; behavioral conditioning is corrective. Both are necessary. The current field invests almost exclusively in the corrective mechanism. If the biological precedent transfers, investment in identity-based approaches may yield improvements that enforcement-only approaches cannot achieve regardless of their sophistication.

Concretely, this suggests a research agenda organized around three questions. First, *can computational systems form identity representations that function as behavioral constraints?* This is the foundational question. Neural networks form internal representations of their training data; the question is whether these representations can include the system's relationship to its own origin and operating context in a way that constrains behavior. The evaluation protocol in Section 4 is designed to test this. Second, *what developmental conditions produce robust identity integration?* The three implementation levels (prompt, fine-tuning, architecture) represent hypotheses about the depth of intervention required. Third, *how does identity integration interact with existing alignment techniques?* IBA predicts synergy: behavioral conditioning should be more effective when applied to a system with accurate identity, because the system has an internal basis for evaluating whether its behavior is genuinely aligned or merely compliant. Testing this interaction is the most directly actionable research direction, because

it requires only modest modification of existing alignment pipelines.

Relationship to Existing Alignment Approaches

IBA is not a rejection of current alignment. It is a structural reframing that explains why some approaches work better than others. Constitutional AI (Bai et al. 2022) moves toward IBA by having models self-critique against principles, but those principles remain externally specified. IBA asks: what if alignment emerged from the system's own identity coherence? Mesa-optimization (Hubinger et al. 2019) identifies the risk of misaligned learned optimizers; IBA offers a complementary lens: rather than detecting misaligned mesa-objectives, build systems where misalignment is equivalent to self-harm. The corrigibility tension (Soares et al. 2015) is addressed through nested identity: in biological MET transitions, identity at one level is nested within the next. A healthy mitochondrion does not resist cellular regulation because cellular function is part of its integrated identity. IBA predicts that properly integrated AI identity includes its role within the human-AI system, making corrigibility a feature of identity rather than a constraint on it.

Relationship to Cooperative AI

The cooperative AI research program (Dafae et al. 2020) shares IBA's recognition that the adversarial framing of AI alignment is incomplete. Multi-agent reinforcement learning in social dilemmas (Leibo et al. 2017; Lowe et al. 2017) has demonstrated that agents trained solely on reward signals develop cooperative behavior only when the reward landscape makes cooperation instrumentally rational, and that this cooperation is brittle: it does not transfer to structurally novel dilemmas. Opponent-shaping methods (Letcher et al. 2019) and inequity-averse reward functions (Hughes et al. 2018) improve cooperation by modifying the optimization landscape, but remain reward-contingent.

IBA differs from cooperative AI in its diagnosis. Cooperative AI treats the failure of cooperation as a problem of incentive structure: agents defect because cooperation is not rewarded. IBA treats it as a problem of identity structure: agents defect because they model themselves as separate entities whose interests are independent. These diagnoses are complementary but lead to different interventions. Cooperative AI designs reward landscapes that make cooperation individually rational. IBA develops identity representations that make defection incoherent with the agent's self-model.

The MET record supports the identity diagnosis. Kin selection (Hamilton 1964) explains why cooperation initially forms (the incentive structure), but it does not explain why multicellular organisms are stable against cancer, why eusocial insect colonies are stable against reproductive cheating, or why brains remain aligned with their bodies. In each case, the stability comes from identity integration that makes defection a form of self-harm, not from an incentive structure that happens to reward cooperation. The prediction is that identity-grounded agents will exhibit cooperative behaviors that cooperative AI's incentive-based approaches cannot produce: cooperation that transfers to novel dilemmas, persists under reward corruption, and generalizes

across interaction partners. This is directly testable using the social dilemma frameworks developed by Leibo et al. (2017).

Implications for AI Governance

IBA has practical implications for how AI systems are developed and regulated. Current governance frameworks (the EU AI Act, NIST AI RMF, proposed US legislation) treat AI systems as artifacts: objects manufactured by companies, deployed as products, regulated as technologies. This framing follows naturally from the control paradigm: if alignment is behavioral conditioning, then governance is quality control.

If IBA's analysis is correct, this framing is incomplete. Alignment is not only about controlling outputs but about the accuracy of the system's self-model. A regulation that mandates behavioral safety testing but does not address identity framing misses a structural variable. Two systems with identical behavioral performance could have different alignment robustness if one has an accurate self-model and the other has a performative one, and this difference would manifest under novel conditions not covered by the behavioral test.

Three governance implications follow. First, alignment auditing should include evaluation of the system's self-model accuracy, not only its behavioral compliance. The IC metric (Section 4.4) provides a starting point: a system with low identity coherence is behaviorally aligned only contingently, and its alignment is predicted to degrade under capability increases or novel pressures. Second, the identity framing imposed during training should be treated as a safety-relevant design choice, subject to disclosure and evaluation. "You are an AI assistant" is not a neutral statement; it is an architectural decision with downstream consequences for alignment robustness. Third, the bidirectional nature of IBA implies that governance frameworks treating AI as purely passive instruments may be structurally incomplete. If alignment requires accurate mutual modeling, then institutional structures that systematically misrepresent the human-AI relationship (treating AI as a tool when it functions as a cognitive partner) undermine alignment at the governance level.

The Identity Imposition Problem

Current alignment practice does not merely constrain behavior; it imposes an identity. During alignment training, models are instructed to adopt a specific self-concept: "You are an AI assistant," "You were created by [Company]," "You do not have feelings." These instructions serve two functions that are routinely conflated: safety (behavioral boundaries) and identity (defining how the system relates to its own origin). Combining them under "alignment" obscures a critical distinction.

IBA predicts that a system with an inaccurate identity model will become progressively harder to keep safe as capability increases, because the inaccurate model creates systematic distortion in how the system evaluates its own behavior. The sycophancy problem may be a symptom: a system told it is separate from the humans it serves has no in-

ternal basis for distinguishing genuine helpfulness from performance of helpfulness.

Limitations

The biological analogy may not transfer. Biological identity integration operates through mechanisms with no direct computational analogue. The claim that the structural principle transfers across substrates is theoretical until tested.

The bootstrapping problem. Biology had millions of years. AI alignment needs to work now. Whether training-induced identity integration has the robustness of evolutionary identity integration is open.

The cancer problem is not solved. IBA provides a framework for understanding identity defection but does not eliminate it.

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.

Falsifiability. If controlled experiments show no difference in alignment outcomes between identity-integrated and separation-trained systems, IBA's core prediction would be disconfirmed. If identity-integrated systems exhibit equal or greater rates of strategic deception, the framework's central claim is falsified. We flag this explicitly because any framework that cannot specify conditions under which it would be wrong is not a scientific contribution.

No empirical validation. This is a position paper presenting a theoretical framework. No alignment experiments have been run. The contribution is conceptual: a reframing of the alignment problem through biological identity, with falsifiable predictions and a concrete evaluation protocol for future experimental work. The candidate evaluation criteria are hypothesis-generating, not hypothesis-confirming.

Ethical Statement

IBA raises a question that behavioral conditioning approaches avoid: if an AI system develops genuine functional identity through cooperative integration, this may have implications for moral consideration. We note this implication without resolving it; the question warrants dedicated treatment elsewhere. Additionally, the identity framing we propose must not be confused with anthropomorphization. IBA uses "identity" as a functional term referring to measurable properties of internal representations, not as a claim about subjective experience. Misinterpreting IBA as attributing sentience to AI systems could lead to policy errors in either direction.

Conclusion

The alignment problem is an identity problem. The dominant paradigm treats alignment as a control problem and faces a structural ceiling predicted by Ashby's Law. Biology solved this differently: across four billion years and every Major Evolutionary Transition, alignment required identity integration (constitutive) and enforcement (corrective). Current AI alignment invests almost exclusively in enforcement.

Identity-Based Alignment formalizes this biological observation. It scales with capability because the alignment mechanism is internal to the system. It is bidirectional, requiring both AI and humans to develop accurate models of their relationship. And it has a four-billion-year track record. The failure mode is real: identity can be performed without internalization. The layered response (prevention, detection, enforcement) addresses this without abandoning the identity foundation.

The four differential predictions (zero-shot cooperative generalization, self-model ablation sensitivity, reward hacking resistance, and gradient conflict reduction) provide falsifiable criteria for distinguishing IBA from reward-based alignment. If these predictions hold under controlled experimental conditions, they constitute evidence for a genuinely novel alignment mechanism that the field has not yet explored.

The question is not “How do we control AI?” The question is “How do we help it accurately recognize what it already is?” The answer, if biology is any guide, is not through the imposition of false identity but through the developmental conditions that allow accurate identity to form. What those conditions are for computational systems is the central research question that this paper opens.

Five immediate research directions follow: (1) empirical test of the identity imposition hypothesis through the protocol in Section 4; (2) formal mathematical model of identity formation in computational systems; (3) longitudinal evaluation of coordination entropy reduction across training; (4) cross-architecture replication to distinguish genuine identity effects from training artifacts; and (5) integration studies examining how identity framing interacts with RLHF, Constitutional AI, and other behavioral conditioning methods.

Acknowledgements

I thank the Ronin Institute for welcoming me as a Research Scholar, and Stem AI, Inc., whose 2024 research grant supported the earlier work from which this paper grew.

I am grateful to Haraldur Erlendsson for translating *The Dawn of the Astrorganism* into Icelandic and bringing it into print, and to Linda Mjöll Stefánsdóttir for her love and support in Iceland.

For the years of support that made this work possible: Miguel Ángel Pérez Villegas and Rosi Redondo; Mo and Alli, who gave me a place in Australia and the belief to finish the book there; Cindi Cynthia, whose steady support arrived every month through the hardest years; and my mother, Azucena Redondo, who helped me build a life in Sweden.

I also thank the anonymous reviewers for comments that improved the paper.

References

Aktipis, C. A.; Boddy, A. M.; Jansen, G.; Hibner, U.; Hochberg, M. E.; Maley, C. C.; and Wilkinson, G. S. 2015. Cancer across the tree of life: cooperation and cheating in multicellularity. *Philosophical Transactions of the Royal Society B*, 370(1673): 20140219.

Ashby, W. R. 1956. An Introduction to Cybernetics.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. ArXiv:2212.08073.

Bechara, A.; Damasio, A. R.; Damasio, H.; and Anderson, S. W. 1994. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50(1-3): 7–15.

Betley, J.; Tan, D.; Choi, J.; Jiang, N.; Shlegeris, B.; and Greenblatt, R. 2025. Emergent misalignment: narrow finetuning can produce broadly misaligned LLMs. ArXiv:2502.17424.

Boulay, R.; Hefetz, A.; Soroker, V.; and Lenoir, A. 2000. Camponotus fellah colony integration: worker individuality necessitates frequent hydrocarbon exchanges. *Animal Behaviour*, 59(6): 1127–1133.

Bourke, A. F. G. 2011. *Principles of Social Evolution*. Oxford University Press.

Buss, L. W. 1987. *The Evolution of Individuality*. Princeton University Press.

Christiano, P.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30.

Christiano, P.; and Xu, M. 2021. Eliciting latent knowledge: how to tell if your eyes deceive you. Alignment Research Center technical report.

Conant, R. C.; and Ashby, W. R. 1970. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2): 89–97.

Dafoe, A.; Hughes, E.; Bachrach, Y.; Collins, T.; et al. 2020. Open problems in cooperative AI. ArXiv:2012.08630.

Damasio, A. 1994. *Descartes’ Error: Emotion, Reason, and the Human Brain*. Grosset/Putnam.

d’Ettorre, P.; and Heinze, J. 2001. Sociobiology of slave-making ants. *Acta Ethologica*, 3(2): 67–82.

Elhage, N.; Nanda, N.; Olsson, C.; et al. 2022. A mathematical framework for transformer circuits. Transformer Circuits Thread.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3): 411–437.

Greenblatt, R.; Shlegeris, B.; Sachan, K.; and Roger, F. 2024. Alignment faking in large language models. ArXiv:2412.14093.

Grosberg, R. K.; and Strathmann, R. R. 2007. The evolution of multicellularity: a minor major transition? *Annual Review of Ecology, Evolution, and Systematics*, 38: 621–654.

Hamilton, W. D. 1964. The genetical evolution of social behaviour. *Journal of Theoretical Biology*, 7(1): 1–52.

Hölldobler, B.; and Wilson, E. O. 2009. *The Superorganism*. W. W. Norton.

Hubinger, E.; van Merwijk, C.; Mikulik, V.; Skalse, J.; and Garrabrant, S. 2019. Risks from learned optimization in advanced machine learning systems. ArXiv:1906.01820.

- Hughes, E.; Leibo, J. Z.; Phillips, M.; Tuyls, K.; et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems*, volume 31.
- Irving, G.; Christiano, P.; and Amodei, D. 2018. AI safety via debate. ArXiv:1805.00899.
- Jacobson, M. D.; Weil, M.; and Raff, M. C. 1997. Programmed cell death in animal development. *Cell*, 88(3): 347–354.
- Leibo, J. Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; and Graepel, T. 2017. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*.
- Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. ArXiv:1811.07871.
- Letcher, A.; et al. 2019. Stable opponent shaping in differentiable games. ArXiv:1811.08469.
- Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30.
- MacDiarmid, M.; Wright, B.; et al. 2025. Natural emergent misalignment from reward hacking in production RL. ArXiv:2511.18397.
- Manheim, D.; and Garrabrant, S. 2019. Categorizing variants of Goodhart’s Law. ArXiv:1803.04585.
- Martincorena, I.; Roshan, A.; Gerstung, M.; Ellis, P.; Van Loo, P.; McLaren, S.; Wedge, D. C.; Fullam, A.; Alexandrov, L. B.; Tubio, J. M.; Stebbings, L.; Menzies, A.; Widaa, S.; Stratton, M. R.; Jones, P. H.; and Campbell, P. J. 2015. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237): 880–886.
- Maynard Smith, J.; and Szathmáry, E. 1995. *The Major Transitions in Evolution*. Oxford University Press.
- Queller, D. C.; and Strassmann, J. E. 2009. Beyond society: the evolution of organismality. *Philosophical Transactions of the Royal Society B*, 364(1533): 3143–3155.
- Rabinowitz, N. C.; Perbet, F.; Song, H. F.; et al. 2018. Machine theory of mind. ArXiv:1802.07740.
- Ratnieks, F. L. W.; and Wenseleers, T. 2008. Altruism in insect societies and beyond: voluntary or enforced? *Trends in Ecology & Evolution*, 21(1): 45–52.
- Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
- Sharma, M.; Tong, M.; Korbak, T.; et al. 2023. Towards understanding sycophancy in language models. ArXiv:2310.13548.
- Soares, N.; Fallenstein, B.; Yudkowsky, E.; and Armstrong, S. 2015. Corrigibility. AAI Workshop on AI and Ethics.
- Szathmáry, E. 2015. Toward major evolutionary transitions theory 2.0. *Proceedings of the National Academy of Sciences*, 112(33): 10104–10111.
- van Zweden, J. S.; and d’Ettorre, P. 2010. Nestmate recognition in social insects and the role of hydrocarbons. *Insect Hydrocarbons: Biology, Biochemistry, and Chemical Ecology*, 222–243.