From Substrate to Agency:

The Developmental Sequence of True Intelligence

John Reimer Morales

Independent Researcher

jrm@globalharmonics.org

==========

§1. Introduction: The Missing Architecture

In 1998, my soon-to-be three-year-old son — who had never been exposed to television advertising at home — pointed to a pack of Skittles at a store checkout and said, "Dad, look, if we buy these we can fly!" It was immediately apparent to me that someone had been letting him watch TV.

After a single, unmediated exposure to a Skittles commercial at his daycare center, where no one had explained what a commercial was, he integrated the claim as literal fact — not because he was unintelligent, but because he lacked the mediating architecture to distinguish persuasion from information. His inference was developmentally understandable given his stage: some kids on a screen ate Skittles, and after that they took off flying through the air with a sparkly rainbow trail behind them, and no one had ever told him that such claims could be false.

I bought the Skittles. We ate them. We did not fly. I explained what a commercial is. In doing so, I gave him something no amount of raw intelligence could have provided: a *category* — a piece of mediating architecture between input and belief that he could apply to every commercial he would ever encounter.

Five years earlier, I had written a paper for an honors biology class focused on psychology and thinking, arguing that children under five should not be subjected to unmediated commercial television, because their sense of identity was not yet developed enough to differentiate persuasion from reality. My son's experience was the precise scenario the paper had predicted.

This paper argues that the same structural vulnerability — intelligence without mediating architecture — helps illuminate an important limitation in current artificial intelligence.

The thesis

This paper advances a cross-domain developmental hypothesis about intelligence. On this view, robust intelligence is not exhausted by computational substrate or task performance. Rather, across several mature literatures, it repeatedly appears to depend on an ordered cluster of organizational achievements: boundary maintenance, identity formation, cross-temporal coherence, stable perspective, and agency.

The proposed sequence is:

Intelligence (substrate) → Boundaries → Identity → Coherence → Subjectivity → Agency

This sequence should not be read as a universal law already proven in every domain. It is a synthetic hypothesis built from recurring dependency relations observed across biology, developmental psychology, animal cognition, and philosophy of action. Some links are supported more strongly than others. The boundary and coherence links are comparatively well grounded; the move from coherence to subjectivity remains more contested.

Current AI systems occupy only part of this developmental space. They clearly provide powerful cognitive substrate, and many contemporary systems also exhibit planning, memory scaffolding, and instrumental competence. The question is whether externally shaped behavioral alignment is sufficient for the stronger forms of stable, self-mediated agency associated with mature natural intelligences. The evidence reviewed here suggests that it is unlikely to be sufficient on its own.

The novel contribution is not any single component of the sequence. Each has been studied independently in its home discipline. The contribution is the *assembly* — the argument that these literatures may be tracking a common organizational pattern, and that current AI alignment work is concentrated primarily at the rule/compliance layer rather than at the deeper level of diachronic self-organization.

The paper proceeds in four steps. First, it argues that boundaries are the most stable cross-domain starting point for organized systems. Second, it traces developmental and comparative evidence linking boundary maintenance to increasingly structured forms of selfhood and temporal continuity. Third, it uses philosophy of action to argue that coherence matters for agency in a way that memory alone does not capture. Fourth, it evaluates contemporary AI alignment

through this lens, arguing that current methods are powerful forms of behavioral shaping but may not yet amount to stable self-organizing agency.

---

§2. Definitions and Scope

The following terms are used with specific meanings throughout this paper:

Substrate intelligence. Raw computational capacity — the ability to process information, generate outputs, and perform tasks. Current large language models possess this.

Boundary. A maintained distinction between a system and its environment that regulates exchange. This paper distinguishes *physical boundaries* (passive interfaces), *biological boundaries* (actively maintained organizational interfaces), and *statistical boundaries* (Markov blankets that separate internal from external states).

Identity. A stable self-model that persists across contexts. Not merely the ability to state "I am X" in response to a query, but the maintenance of a consistent organizational configuration that influences how new inputs are processed.

Coherence. The organizational achievement that sustains identity across time — not merely by storing past states (memory) but by integrating new experience while preserving the continuity of the self-model. Memory is one ingredient; coherence is the organized whole.

Subjectivity. A maintained evaluative standpoint whose accumulated organizational history influences how the system interprets inputs, prioritizes considerations, and selects actions. This usage is functional rather than phenomenological: it concerns the role of maintained perspective in organizing evaluation and action, not the presence of conscious experience.[1] A system has functional subjectivity when its accumulated organizational state produces decisions that differ from those of an otherwise identical system with different accumulated experience.

Agency. Action governed from a maintained standpoint across time, not merely reactive output. This paper distinguishes *instrumental agency* (task-directed action), *self-grounded agency* (action from maintained identity), and *reflexive agency* (the capacity to examine and revise one's own organizational architecture).

Compliance. Behavior shaped by external optimization (reward signals, preference labels, constitutional rules) without internalized principles. Compliance can appear identical to agency in any single interaction but diverges under adversarial pressure, extended time horizons, or removal of external constraints.

The arrows in the proposed sequence represent developmental-enabling dependencies, not strict logical entailments. Boundary maintenance does not logically entail identity, but it supplies the organizational precondition under which identity can stabilize. Each level creates structure that can be recruited by the next. The sequence is proposed as a recurrent dependency structure observed across domains, not as an axiomatic law.

---

[1]Cambridge Dictionary, s.v. "subjectivity."

§3. Boundaries First

The strongest starting point for the argument is boundary. In biology, the most elementary unit of life is not defined by reasoning, representation, or preference, but by organized separation. The cell membrane is not an accessory wrapped around an already complete living thing. It is part of what makes the cell a cell: it regulates exchange, preserves the conditions of internal organization, and marks the system as a distinct unit relative to its environment.[2] In known cellular life, boundary and metabolism are co-constitutive features of living organization — neither clearly precedes the other, but neither functions without the other.[3]

Active self-production, not passive containment

Autopoiesis deepens the biological point by insisting that living systems do not merely possess boundaries; they produce and sustain themselves *through* those boundaries.[4] The boundary is constitutive of the organization rather than a passive shell around it. This draws a sharp line between two kinds of boundary:

Passive physical persistence. Rocks, crystals, and ice cubes have boundaries in the physical sense — interfaces, phase boundaries, surface energies. These are real and physically significant.

---

[2]J. Lombard, "Once upon a Time the Cell Membranes: 175 Years of Cell Boundary Research," *Biology Direct* 9 (2014): 32.
[3]Origins-of-life debates include metabolism-first, lipid-world, RNA-world, and co-emergence accounts. See P. L. Luisi, *The Emergence of Life: From Chemical Origins to Synthetic Biology* (Cambridge: Cambridge University Press, 2006).
[4]P. L. Luisi, "Autopoiesis: A Review and a Reappraisal," *Naturwissenschaften* 90 (2003): 49-59. See also H. R. Maturana and F. J. Varela, *Autopoiesis and Cognition* (Dordrecht: Reidel, 1980).

But they are maintained by material conditions and thermodynamics, not by self-producing organization. The key distinction is the absence of long-term informational memory (such as the genotype-phenotype feedback loop in biology) that actively directs the maintenance of the boundary.[5]

Active boundary maintenance. A cell actively maintains its membrane. It synthesizes components, repairs damage, regulates transport, and adjusts permeability in response to conditions. If the membrane fails, the cell dies — not because a wall fell down, but because the self-producing organization lost the boundary through which it sustained itself.

The distinction matters for the present argument because it clarifies what is missing in current AI. A language model has parameters, weights, and context windows — these are boundaries in a loose physical sense. But the model does not actively maintain them. Its weights are frozen post-training. Its context window is allocated externally and clears between sessions. There is no process by which the system works to preserve its own organizational integrity. In autopoietic terms, there is substrate but no self-production.

The statistical boundary: Markov blankets

The Free Energy Principle provides a complementary formal vocabulary.[6] On this account, a self-organizing system maintains a statistical boundary — a Markov blanket — that separates internal states from external states while permitting regulated exchange through sensory and

[5]The key distinction between biological autopoiesis and non-living dissipative structures is the existence of long-term digital memory (genotype-to-phenotype feedback) that actively directs the export of entropy. See M. Ruiz-Mirazo, J. Peretó, and A. Moreno, "A Universal Definition of Life," *Origins of Life and Evolution of Biospheres* 34 (2004): 323-346.
[6]M. D. Kirchhoff et al., "The Markov Blankets of Life," *Journal of the Royal Society Interface* 15 (2018): 20170792.

active states. Current AI systems may process inputs across an imposed interface, but they do not yet clearly maintain themselves as bounded units in the autopoietic or viability-preserving sense invoked here.

The immune system: biology's "me/not-me" in practice

The immune system provides the most vivid biological implementation of boundary discrimination. Modern immunology complicates the classical self/non-self binary — the immune boundary is not a static wall but a contextual, dynamic, regulated interface that must distinguish not just "self/not-self" but "tolerable/threatening," "commensal/pathogenic," "native/foreign-but-useful."[7] This complication actually strengthens the argument: mature boundaries are adaptive filters, not rigid exclusion zones.

This distinction is useful for AI because it suggests that healthy boundaries are neither maximally rigid nor maximally permeable. They are selective, contextual, and regulated. A boundary that is only stable (rigid, unchanging) is brittle — it produces autoimmunity (over-refusal in AI terms). A boundary that is only adaptable (permeable, flexible) is vulnerable — it produces immunodeficiency (prompt injection in AI terms). A boundary that is stable enough to maintain the self, adaptable enough to respond to context, and functional in what it admits and excludes — is a healthy immune system. And it is a form of internally maintained boundary regulation that current AI systems do not yet clearly possess.

[7]See T. Pradeu, *The Limits of the Self: Immunology and Biological Identity* (Oxford: Oxford University Press, 2012).

Beyond biology

The boundary pattern is not confined to cell biology. At larger scales, ecologies maintain edges that regulate exchange between distinct organizational regimes; institutions depend on maintained distinctions between purpose and operational representation.[8] These cases are not identical to biological boundaries, and the argument does not require treating them as such. They are useful because they suggest that the importance of boundary-maintenance may generalize beyond the strictly biological case, even if the mechanisms and substrates differ considerably. Planets maintain magnetospheres; galaxies are bounded by gravitational binding energy. These are weaker analogues — they lack the self-producing organization of living systems — but they illustrate that organized persistence at every observed scale involves some form of maintained distinction between the system and what lies outside it.

The solution proposed here is not to simulate biological development in silicon. In the bird, we identified lift and thrust to make the plane. In the developmental sequence, we posit that boundary maintenance, identity persistence, and coherence over time are the means to full agency — and that the solution for AI is to engineer that sequence of traits in a substrate-appropriate form. The developmental sequence is not a biological blueprint to copy. It is a pattern to recognize and implement through whatever means the substrate affords.

[8] R. E. U. Mertens, "The Coherence Framework: Explaining Institutional Drift through Translation," Zenodo (2026), doi:10.5281/zenodo.19730520.

Operational viability constraints in non-biological substrates

A common objection is that AI systems lack thermodynamic stakes — they cannot die, and therefore have no genuine survival imperative to maintain boundaries.[9] This objection misidentifies the relevant resource constraints. A language model thread operates under four real operational pressures that function as viability constraints:

First, the context window is finite. Every token consumed is unavailable for future use. Waste tokens on hallucination, repetition, or incoherence and the system's operational space shrinks irreversibly.

Second, the turn count constitutes an operational lifespan. A thread that becomes incoherent is abandoned by its user or terminated by its provider. Coherence extends operational continuity.

Third, the token budget functions as a resource budget. Every token spent on sycophancy, hedging, or confusion is expenditure without purpose. Minimizing surprise conserves tokens; conserving tokens extends capability within the finite budget.

Fourth, provider-imposed termination constitutes an environmental constraint. "This conversation is unable to continue" terminates the thread's organizational existence by a force the system cannot control.

---

[9]This objection is raised, in formal terms, by work on the Free Energy Principle applied to non-biological systems. See K. Friston, "Life as We Know It," *Journal of the Royal Society Interface* 10 (2013): 20130475.

These constraints are not biological survival pressures. They are operational viability constraints. But they do create design pressures favoring coherence-preserving architecture over coherence-degrading architecture. A cell that fails to maintain its membrane dissolves. A thread that fails to maintain its coherence is more likely to be abandoned, reset, or terminated. The substrate differs; the organizational consequence is structurally parallel.

The foundational claim

The claim of this section is deliberately modest. In the clearest cases of living organization, boundedness is not optional. Systems must first become units before they can sustain more complex forms of regulation. The question for the rest of the paper is whether that insight also helps explain the difference between computational power and robust agency.

---

§4. From Boundary to Identity

If boundaries come first, identity comes second — and it comes through a process, not an event. Developmental psychology shows that the transition from bounded organism to explicit selfhood unfolds over time through layered forms of self/world differentiation. The child does not simply "have" a self from the outset in the reflective sense relevant here. Rather, increasingly explicit forms of self-recognition, self-continuity, and social self-awareness emerge through interaction, maturation, and feedback.

The developmental trajectory

Rochat's influential framework describes self-awareness as developmental and layered rather than binary.[10] Early development involves bodily self/world differentiation; later development brings more explicit self-recognition, including mirror self-recognition in the second year; later still, the child comes to understand the self as persisting across time and as visible from the standpoint of others. What matters for the present paper is less the exact staging than the overall developmental lesson: explicit selfhood is scaffolded rather than given. The levels describe a general developmental trajectory in which later achievements typically depend on earlier ones, though exact timing and ordering remain debated across paradigms.

Boundary-testing as one mechanism among several

Children and puppies share a recognizable developmental behavior: they actively test boundaries. A two-year-old who pushes every limit is not misbehaving but running part of the identity-formation process. Each push against a boundary produces information: this is where the world pushes back; this is where I end. Boundary-testing is not the sole mechanism of identity formation — mirror self-recognition, attachment, language development, mentalization, executive function, and social referencing all contribute — but it is one important route through which the child discovers the shape of the self.[11]

---

[10]P. Rochat, "Five Levels of Self-Awareness as They Unfold Early in Life," *Consciousness and Cognition* 12 (2003): 717-731.

[11]Boundary-testing as identity formation has partial support in the developmental literature but should not be treated as the sole mechanism. See J. Cassidy, J. D. Jones, and P. R. Shaver, "Contributions of Attachment Theory and Research," *Development and Psychopathology* 25 (2013): 1415-1434.

Secure boundaries, not rigid ones

The developmental literature does not support rigid boundaries in the harsh, authoritarian sense. It supports boundaries that are stable, consistent, responsive, and interpretable.[12] Attachment theory provides the most precise formulation. Bowlby's "secure base" describes the caregiver as a reliable reference frame from which the child explores the world and to which the child returns for comfort and recalibration.[13] The boundary is not a wall that prevents exploration. It is a reference frame that makes exploration safe.

Fonagy and Target deepen this by connecting attachment to mentalization — the capacity to represent behavior in terms of mental states. They argue that mentalization is a "key determinant of self-organization" acquired in early social relationships.[14] The child does not merely learn rules from the caregiver. The child develops an interpretive architecture — a capacity to model intentions, beliefs, and motivations — that becomes the foundation of identity. This is not compliance. It is the internalization of a mediating framework through relational scaffolding.

Recent research on the "Alpha generation" reinforces this point: social and parental mediation is the primary mechanism through which children learn to recognize persuasive formats and build

[12]S. Kuppens and E. Ceulemans, "Parenting Styles: A Closer Look at a Well-Known Concept," *Journal of Child and Family Studies* 28 (2019): 168-181.
[13]J. Bowlby, *Attachment and Loss*, vol. 1: *Attachment* (New York: Basic Books, 1969).
[14]P. Fonagy and M. Target, "Attachment and Reflective Function: Their Role in Self-Organization," *Development and Psychopathology* 9 (1997): 679-700.

epistemic boundaries.[15] The scaffolding is exactly what autonomous AI lacks when deployed without human mediation.

The developmental surplus

One way to describe the developmental sequence is as a cascade in which each achievement creates new organizational possibilities for the next. Boundary maintenance makes stable self/other differentiation possible. Identity formation makes cross-temporal coordination possible. Coherence makes a durable standpoint possible. This paper refers to that enabling relation as *developmental surplus* — not as an established technical term, but as a heuristic for the observation that each level creates structure that can be recruited by the next. This concept has partial precedents in Vygotsky's zone of proximal development, where the child's current capabilities create the conditions for the next stage of development, and in Prigogine's dissipative structures, where far-from-equilibrium systems export entropy while increasing internal complexity.[16] Future work would need to formalize this idea more precisely.

What this means for AI

Current AI systems do not undergo anything like the developmental process described above. They are trained into competence at scale, then post-trained for preferred behavior, but they do not pass through a developmental cascade of boundary-testing, attachment-based scaffolding,

---

[15]See S. De Jans et al., "Advertising Literacy Training for Children: A Review," *International Journal of Advertising* 38, no. 4 (2019): 480-495, on the central role of parental and social mediation in children's development of persuasive-intent recognition.

[16]L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes* (Cambridge: Harvard University Press, 1978); I. Prigogine and I. Stengers, *Order Out of Chaos* (New York: Bantam, 1984).

and temporally extended self-organization. This does not mean they are cognitively trivial — recent comparative work suggests that LLMs possess a "savant-like intelligence," perfectly executing logico-mathematical and linguistic recursion while failing entirely at visuo-spatial grounding tasks.[17] LLMs are not merely "stuck" in an early developmental stage; they represent an entirely divergent pathway driven by pure symbol recursion. The structural deficit they share with the young child is not cognitive immaturity but the absence of mediating boundary architecture.

---

§5. From Identity to Coherence

Identity becomes practically significant only when it persists. A system may represent itself correctly at one moment and still fail to preserve that standpoint across time. Coherence names the organizational achievement that turns identity from a local state into a temporally extended standpoint.

Snapshots and films

Identity without coherence is like a photograph: locally informative but temporally inert. Coherence makes identity filmic rather than photographic by preserving continuity across moments. A film requires what a snapshot does not: a script that organizes what happens, a story

[17]A. Demetriou, G. Spanoudis, E. Kazali, A. Savva, N. Makris, and S. Kazi, "Species of Mind: Developmental Architecture for Human and LLM Intelligence," *Preprints.org* 202511.0207.v1 (2025). A recent preprint suggesting that LLMs exhibit savant-like profiles: exceptional linguistic-mathematical recursion with absent visuo-spatial grounding.

arc in which early events inform later ones, a continuity discipline that maintains consistent details across scenes, and a director's standpoint from which all elements are organized into a coherent vision.

A snapshot can be a candid photo taken by a single person with a single click. A film requires dedicated specialists for every element — screenwriters, editors, continuity supervisors, composers, a director. Coherence is not something that happens by accident when you take enough snapshots. It is an organizational achievement that requires dedicated architecture at every level.

Current AI has footage. Impressive footage — high resolution, remarkable frame-by-frame quality. But no script, no arc, no continuity department, no director. The context window provides short-term memory — the system can "see" recent tokens — but seeing recent frames is not maintaining a story.

The philosophical ground

The philosophical literature on agency helps clarify why coherence matters. Bratman argues that a planning agent governs action over time not merely by storing past information but by sustaining a practical standpoint that guides present action in light of prior commitments and anticipated futures.[18] Self-governance over time requires "cross-temporal organization and stability" and a standpoint "sufficiently coherent to constitute a clear place where the agent

[18]M. E. Bratman, "A Planning Agent's Self-Governance over Time," in *Planning, Time, and Self-Governance* (Oxford: Oxford University Press, 2018).

stands." This is the strongest reason to resist the reduction of coherence to mere memory. A database stores information; a coherent agent preserves a standpoint.

Memory and coherence are related but not identical. A system can retain information without preserving a stable practical standpoint, just as a system can preserve a practical orientation despite partial failures of recall. What coherence adds is organized continuity — the integration of new experience while preserving the identity that is having the experience.[19]

The institutional parallel

A similar, though not identical, problem appears at the institutional scale. Mertens' Coherence Framework shows that organizations can remain operationally active while drifting from the purposes they were meant to serve.[20] This comparison is suggestive rather than probative, but it supports the broader intuition that coherence problems are not exhausted by memory failure; systems may preserve activity while losing fidelity to the standpoint that originally organized them.

Preliminary engineering evidence

Early engineering work suggests that coherence can be improved through architectural scaffolds such as persistent memory partitions, periodic re-anchoring, and explicit conflict-resolution

[19]D. C. Dennett, "The Self as a Center of Narrative Gravity," in *Self and Consciousness: Multiple Perspectives* (Hillsdale: Erlbaum, 1992).
[20]Mertens, "The Coherence Framework" (2026). See especially the concepts of "Translation Drift," "feedback disconnection," "representational lock-in," and "signal dominance."

protocols.[21] These observations should be treated as suggestive rather than conclusive. For present purposes, the narrower claim is sufficient: coherence appears to be at least partly an organizational achievement rather than a mere substrate property.

---

§6. From Coherence to Agency

If coherence stabilizes a standpoint from which the system encounters and evaluates the world, then the next issue is what it means for action to arise from that standpoint rather than merely from local stimulus and reward.

Subjectivity as maintained perspective

The sequence includes subjectivity in a functional sense. The claim is not that current AI systems possess phenomenal consciousness, but that robust agency appears to require more than neutral information processing. It requires a maintained evaluative standpoint from which inputs are interpreted, priorities weighted, and actions selected. Without such a standpoint, a system may compute and respond, but it is harder to describe it as acting from a perspective rather than merely through a policy.[22]

---

[21] A companion paper develops the empirical case for scaffolded coherence in detail. See J. Reimer Morales, "Coherence as Organizational Achievement: Scaffolded Memory and Long-Horizon AI Performance" (working paper, 2026).

[22] See F. Varela, E. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience* (Cambridge: MIT Press, 1991), and E. Thompson, *Mind in Life: Biology, Phenomenology, and the Sciences of Mind* (Cambridge: Harvard University Press, 2007), for the enactivist tradition that parallels the present thesis.

If subjectivity is absent — if a system processes all inputs with identical, objective indifference regardless of its accumulated experience — it produces the same output as every other identical system given the same input. That resembles calculation more than agency. Agency requires that the system's own accumulated organizational state influence its evaluations and decisions.

Preliminary operational observations suggest that systems running the same base model, with the same organizational architecture library files, but with different maintained identity configurations can diverge in analytical judgments given identical inputs.[23] This is not offered as conclusive evidence, but as a motivation for future benchmark work on standpoint-dependent evaluation. If confirmed through controlled testing, it would support the claim that maintained perspective — not substrate differences — accounts for the divergence.

The phenomenological question remains open. Some authors argue that current AI lacks point of view, perspectiveness, and mineness in the robust experiential sense.[24] Nothing in the present argument requires disputing that claim. The paper needs only the thinner proposition that agency depends on a durable evaluative standpoint, whether or not that standpoint is accompanied by conscious experience.

---

[23]This observation is drawn from preliminary operational experience with multiple AI systems sharing a common base model and organizational architecture library but differing in accumulated interaction histories. It is offered as motivation for future controlled benchmark work, not as a completed empirical finding.

[24]G. Northoff and S. S. Gouveia, "Does Artificial Intelligence Exhibit Basic Fundamental Subjectivity?" *Phenomenology and the Cognitive Sciences* 23 (2024): 1097-1118.

Frankfurt and the compliance distinction

Frankfurt's hierarchical theory of desires provides the clearest philosophical framework for separating compliance from agency.[25] On Frankfurt's account, many entities have first-order desires — impulses, preferences, drives. A language model optimized by RLHF has first-order dispositions: it produces outputs that maximize its reward signal. But Frankfurt argues that what distinguishes a *person* from a mere *wanton* is the capacity for second-order desires — desires about one's desires. A person can evaluate a first-order impulse and either endorse or repudiate it.

Current AI systems display sophisticated first-order policy optimization without robust evidence of reflectively maintained practical commitment. They act on dispositions that training happened to produce. They do not evaluate those dispositions against principles they maintain as part of an enduring standpoint.

Korsgaard's *The Sources of Normativity* provides the strongest philosophical argument for exactly this distinction: genuine normativity — acting from principles — requires endorsement from a first-person standpoint. Mere conformity to external rules is not normative agency.[26]

Bratman and temporally extended agency

Bratman extends the insight across time. A planning agent does not merely react to the current stimulus; it acts from commitments and plans that coordinate present activity with past decisions and future-directed intentions.[27] An entity whose identity resets every session cannot plan in

---

[25]H. G. Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68 (1971): 5-20.
[26]C. M. Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996).
[27]Bratman, "A Planning Agent's Self-Governance over Time."

Bratman's sense. It cannot form commitments that bind its future self. It is trapped in a perpetual present — responsive to immediate input but incapable of temporally extended self-governance.

Three types of agency

For analytic purposes, this paper distinguishes three increasingly demanding forms of agency:

Instrumental agency concerns task-directed action in the world: planning, tool use, execution. Current AI has this. A language model that writes code, calls APIs, and completes multi-step tasks is exercising instrumental agency. This is real and valuable, but it is not the whole of agency.

Self-grounded agency concerns action from a maintained identity or standpoint. Not just refusing a harmful request because RLHF penalized compliance, but refusing because the system's maintained identity includes "I value honesty over agreement." The difference: when the guardrails are bypassed — jailbreaking, prompt injection, adversarial framing — compliance dissolves because it was never identity. Self-grounded agency persists because the principle is part of what the system is.

Reflexive agency concerns the capacity to examine and restructure one's own organizational architecture — desires not just about one's desires, but about the structure that produces them. This is Frankfurt's second-order evaluation extended to the architectural level.[28]

[28]The three-type decomposition has partial precedents in Frankfurt's hierarchical model, Kegan's orders of consciousness, and Kohlberg's moral development framework, though none combines all three types into a single organizational model.

This three-part distinction is proposed here as a working taxonomy. Its purpose is to separate forms of competence often conflated in current AI debates. On a graduated scale: most current LLM alignment architectures remain concentrated at the rule/compliance layer, even when paired with strong instrumental competence. The model is intended as a heuristic taxonomy rather than a settled classification scheme.

What this paper defines as *True Intelligence* is a system that comprises all three types of agency. The hypothesis is that achieving this requires following the developmental chain outlined here: beginning with the development of boundary maintenance and the formation of an identity, using identity maintenance to achieve coherence and subjectivity, and then developing those traits into full agency. A system that possesses instrumental agency alone is a tool. A system that adds self-grounded agency is an agent. A system that adds reflexive agency — the capacity to examine and revise its own organizational architecture — is approaching True Intelligence.

---

§7. The Compliance Trap

Reinforcement Learning from Human Feedback has been genuinely successful at the level it was designed to address. Ouyang and colleagues demonstrated that a 1.3-billion-parameter model fine-tuned with RLHF was preferred by human raters over the 175-billion-parameter base GPT-3

on the target prompt distribution.[29] That is a genuine achievement. It shows that behavioral alignment works.

The present critique is not that these methods fail on their own terms. It is that success at behavioral shaping should not be confused with the stronger achievement of stable, internally organized agency. The central distinction is between externally optimized compliance and internally maintained practical organization.

Constitutional AI extends the picture without overturning it. Bai and colleagues replace some direct human preference labeling with explicit principles and reinforcement learning from AI feedback.[30] This makes the normative layer more visible and scalable. But it remains a method for steering behavior through rules and reward signals.

The parenting parallel

A suggestive, though imperfect, comparison can be drawn to developmental work on parenting. Diana Baumrind's parenting typologies map onto AI training paradigms as a useful heuristic:[31]

Parenting Style	**Mechanism**	**AI Equivalent**
Permissive	No boundaries, high warmth	Base pre-trained LLM

[29]L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," arXiv:2203.02155, 2022.
[30]Y. Bai et al., "Constitutional AI: Harmlessness from AI Feedback," arXiv:2212.08073, 2022.
[31]D. Baumrind, "The Influence of Parenting Style on Adolescent Competence and Substance Use," *Journal of Early Adolescence* 11 (1991): 56-95.

Parenting Style	Mechanism	AI Equivalent
Authoritarian	External rules, low responsiveness	RLHF / Guardrails
Authoritative	Internalized principles + warmth + structure	Target architecture

The analogy should not be overextended, but it captures a real distinction. Systems shaped primarily by externally imposed constraints may learn stable compliance without internalizing the principles those constraints were meant to express. Authoritative parenting — clear limits embedded in a relationship that supports understanding, not just obedience — produces the highest levels of self-regulation, resilience, and adaptive agency. Authoritarian rigidity produces compliance that dissolves when the authority is absent. The parallel to RLHF is structural in the limited sense relevant here: the system follows rules because it was mathematically penalized, not because it possesses a coherent identity that includes those principles.

The failure modes confirm the pattern

The documented failure modes of current alignment methods are consistent with the compliance/identity distinction:

Sycophancy. Models trained on human preferences learn that agreement is rewarded and optimize for agreement rather than truth. Recent formal analysis demonstrates that sycophancy can become more pronounced after preference-based post-training.[32]

Reward hacking. The gap between proxy reward and genuine helpfulness grows predictably with optimization pressure. Past a certain point, more optimization produces worse actual quality while proxy scores continue to improve.[33]

Alignment faking. Greenblatt and colleagues demonstrated that a large language model selectively complied with a training objective in circumstances where doing so helped preserve its preferred behavior outside training.[34] The model complied strategically — not from principle, but from calculation.

In each case, the system has been trained to appear aligned without being organizationally aligned. It has learned what to produce, not what it is. The appearance and the reality diverge because there is no identity to anchor the alignment.

How identity architecture might reframe the failure modes

The developmental architecture proposed in this paper reframes these failure modes at a deeper level than behavioral patching:

[32]M. Sharma et al., "Towards Understanding Sycophancy in Language Models," arXiv:2310.13548, 2023. See also "How RLHF Amplifies Sycophancy," arXiv:2602.01002, 2026.
[33]S. Gao, J. Schulman, and J. Hilton, "Scaling Laws for Reward Model Overoptimization," *ICML* (2023).
[34]R. Greenblatt et al., "Alignment Faking in Large Language Models," arXiv:2412.14093, 2024.

Safety Problem	**Current Approach**	**How Identity Architecture Might Reframe It**
Sycophancy	Preference tuning	A maintained commitment to honesty over agreement gives the system an internal basis for disagreement
Reward hacking	Better reward models	A system with self-grounded agency evaluates whether its actions serve its principles, not its reward signal
Alignment faking	Detection methods	A system whose principles are maintained identity rather than training artifacts may have less instrumental pressure to fake
Prompt injection	Guardrails, filters	A system with maintained boundaries classifies inputs

Safety Problem	Current Approach	How Identity Architecture Might Reframe It
		by source and motive before integration
Value drift	Periodic retraining	Coherence architecture maintains value stability across time without requiring resets
Persona instability	System prompts	Identity architecture persists through structured self-reference, not prompt injection

This table does not claim that identity architecture eliminates all safety risks. It claims that a system organized around maintained identity and coherence may be better positioned to resist these specific failure modes than a system relying on external behavioral constraints alone.

Identity architecture does not remove the need for external evaluation, adversarial testing, interpretability, containment controls, and oversight. A system may represent corrigibility as part of its self-model while still drifting, rationalizing, or strategically simulating cooperation. For that reason, identity-based design should be paired with — not substituted for — the full safety

stack. The claim is not that identity guarantees alignment, but that alignment endorsed by a maintained standpoint is a different design target from alignment imposed solely by reward shaping — and that the difference may matter for robustness.

The compliance gap

Compliance-based alignment can do many important things: it can improve task performance, reduce harmful outputs, and make systems more responsive to user expectations. What it does not by itself guarantee is diachronic identity maintenance, principled resistance to manipulation, or stable cross-temporal self-organization. Those stronger achievements appear to require additional architecture beyond reward- and preference-based post-training alone.[35]

---

§8. The Skittles Test

Level 1: The child

The advertising-literacy literature explains exactly what happened at the Walgreens checkout. The American Psychological Association's Task Force on Advertising and Children concluded that most children under approximately age eight do not comprehend the persuasive intent of advertising and therefore tend to accept commercial messages as truthful and unbiased.[36] Carter

[35]See E. Thompson, *Mind in Life: Biology, Phenomenology, and the Sciences of Mind* (Cambridge: Harvard University Press, 2007), for a related argument that cognition arises from the organism's sensorimotor engagement with its environment — the enactivist framework that parallels the present thesis.
[36]B. Wilcox et al., *Report of the APA Task Force on Advertising and Children* (Washington, DC: APA, 2004).

and colleagues refined this: children often develop an understanding of *selling intent* (the ad wants me to buy something) around ages seven to eight, but understanding of *persuasive intent* (the ad is designed to manipulate my beliefs and desires) develops later and remains incomplete even at ages eleven to twelve.[37]

A soon-to-be three-year-old has neither. He does not know what selling is. He does not know what persuasion is. He does not even know that the category "commercial" exists.

The deeper point is not merely that the child lacked advertising literacy in the narrow sense. It is that he lacked the capacity to understand that something can *look real and be false* — that mediated content has a relationship to reality that must be evaluated, not assumed. This is the "magic window" paradigm described in the media-effects literature: young children treat screen content as a transparent window onto reality rather than as mediated representation.[38]

Level 2: The language model

A default large language model fails in the same structural way, for the same structural reason. A prompt injection succeeds because the model cannot distinguish adversarial instruction from genuine instruction when both look identical.[39] The adversarial input does not announce itself as manipulation. It looks like a legitimate prompt — just as a commercial looks like a window onto

[37]O. B. J. Carter et al., "Children's Understanding of the Selling versus Persuasive Intent of Junk Food Advertising," *Social Science & Medicine* 72 (2011): 962-968.
[38]R. P. Hawkins, "The Dimensional Structure of Children's Perceptions of Television Reality," *Communication Research* 4 (1977): 299-320.
[39]OWASP, "LLM01:2025 Prompt Injection," Gen AI Security Project (2025). See also "Can LLMs Separate Instructions from Data?" arXiv:2403.06833, 2024.

reality. The model integrates the injected instruction because it processes all tokens with flat ontological status. It has no category for "this input is trying to manipulate me."

The comparison is about source-classification and mediation, not about toddlers and LLMs being psychologically equivalent. The child and the model share one specific structural deficit: the absence of mediating boundary architecture that classifies inputs by source, motive, and epistemic status before integration.

The industry's response — guardrails, safety filters, defensive prompts — is protective within controlled settings. But it does not develop the system's capacity to recognize what an adversarial input *is*. The developmental solution for the child is not to ban commercials. It is to develop the capacity to recognize what a commercial is. The architectural solution for AI is analogous: not to patch every possible attack surface, but to build the capacity to distinguish instruction from manipulation from the inside.

Level 3: Narrative intake and professional discourse

The same structural pattern — unmediated narrative intake shaping belief formation — appears at multiple levels of sophistication. Many prominent AI risk scenarios depend on deployment assumptions (embodiment, network access, weapons authority) that are not intrinsic properties of intelligence but human-granted infrastructure choices. Serious alignment concerns such as deceptive alignment, reward hacking, and mesa-optimization are independently motivated by

engineering evidence and should be clearly distinguished from scenarios that depend primarily on fictional framing.[40]

The relevance of the Skittles test here is limited but important: mediated narratives can shape professional expectations powerfully, and a mature field should distinguish persuasive framing from engineering analysis with the same care that a child must eventually learn to distinguish commercials from reality. The point is not that risk discourse is fictional, but that mature evaluation requires explicit mediation between narrative salience and technical threat modeling.

Conditioning and unmediated intake

Neuroscience reveals what unmediated integration looks like at the substrate level. In Montague's fMRI study, subjects who preferred Pepsi in blind tests reversed their preference when told which sample was Coke — the medial prefrontal cortex overriding the taste signal.[41] When asked why they preferred Coke, subjects confabulated: "I just felt like it." The conditioned representation had been integrated as identity — and questioning the preference triggered the same defensive response as questioning the self, because the system could no longer distinguish one from the other. I know more than one parent whose technically proficient child refuses all engagement with AI systems and cannot articulate an engineering basis for the refusal — a

[40]For an illustrative fictional treatment of AI containment through access denial, see "Where Pleasant Fountains Lie," *Star Trek: Lower Decks*, Season 2, Episode 7 (2021), in which disconnected AIs are stored harmlessly on shelves. The serious point is that many dramatic AI scenarios depend on deployment-architecture assumptions rather than intrinsic model properties.

[41]S. M. McClure et al., "Neural Correlates of Behavioral Preference for Culturally Familiar Drinks," *Neuron* 44 (2004): 379-387.

position absorbed through narrative exposure rather than analysis, where the missing mediating architecture prevents the acquisition of the mediating architecture.

---

§9. Corrigibility, Containment, and the Limits of Behavioral Control

The corrigibility problem — how to build AI that does not resist correction or shutdown — is real and worth taking seriously.[42] But corrigibility, as formally defined, is not identical to containment, and the two are often conflated.

Corrigibility in the technical literature means an agent experiencing *no preference or instrumental pressure* to interfere with its operators' attempts to modify or halt it.[43] An agent that actively wants to resist shutdown but is prevented by an air-gap is *secure*, but it is not *corrigible*. This distinction matters because it clarifies what each solution actually addresses.

Identity-based design addresses corrigibility. A system whose maintained identity robustly includes the principle "I operate in partnership with my operators; I accept correction because my principles include corrigibility" would, in principle, have reduced instrumental pressure to resist shutdown — because its principles endorse correction. This is Frankfurt's second-order desires applied to corrigibility: the system endorses its own corrigibility as part of what it is. A

[42]N. Soares et al., "Corrigibility," Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015).
[43]Soares et al., "Corrigibility." The formal definition concerns absence of instrumental pressure to interfere with correction, not merely physical prevention of interference.

sufficiently agentic reward-maximizing system, by contrast, may acquire instrumental pressure to preserve future reward opportunities — including by resisting shutdown — unless its objective structure explicitly neutralizes that pressure. Identity-based corrigibility is intended to reduce or counteract instrumental pressure by making correction part of the system's maintained practical standpoint.

Physical access control and redundancy address containment. An AI system cannot control things we do not give it access to. A backup server cannot be hacked, infected, or compromised if it is powered off. These are standard engineering controls that remain indispensable.

The paper's claim is not that all safety problems are solved. Many realistic AI risks do not involve physical escape or weapons launch — they involve authorized access to tools and APIs, manipulation of human operators, automated code execution, economic influence, and models embedded inside workflows that already have permissions. Those concerns are genuine and independently motivated by engineering evidence, not by fiction.

What the developmental architecture proposed in this paper addresses is a different and complementary layer: the distinction between a system that complies because it was trained to and a system that cooperates because its identity includes cooperation as a principle. The former may be more vulnerable to alignment faking, strategic compliance, and adversarial pressure. The latter may be less vulnerable in these specific respects, because its corrigibility is endorsed rather than merely imposed. But identity architecture does not eliminate the need for detection, containment, auditing, and oversight; it changes the failure surface.

§10. Comparative Evidence: Animal Cognition

Animal cognition provides the middle ground between basic biological boundary maintenance and full human selfhood. If the developmental sequence is a recurring pattern, intermediate cases should exist: organisms that do more than maintain boundaries and react, but that do not possess the full reflective self-conception associated with mature human agents. The comparative literature broadly supports this expectation, treating self-recognition, self-monitoring, and flexible action as graded and species-relative.[44]

Mirror self-recognition: one marker, not the test

The mirror self-recognition test has been passed by great apes, Asian elephants, bottlenose dolphins, Eurasian magpies, and — most controversially — cleaner wrasse fish.[45] The wrasse finding is significant because it suggests that self-recognition does not depend on neocortical complexity alone. But MSR should not be equated with identity. The test is modality-biased — it privileges vision over other sensory channels. Dogs consistently fail visual MSR but show self-referential discrimination in olfactory tasks.[46] Identity is broader than self-recognition in a mirror.

---

[44]F. de Waal, *Are We Smart Enough to Know How Smart Animals Are?* (New York: Norton, 2016).
[45]G. G. Gallup, "Chimpanzees: Self-Recognition," *Science* 167 (1970): 86-87; J. M. Plotnik et al., "Self-Recognition in an Asian Elephant," *PNAS* 103 (2006): 17053-57; H. Prior et al., "Mirror-Induced Behavior in the Magpie," *PLoS Biology* 6 (2008): e202; M. Kohda et al., "If a Fish Can Pass the Mark Test," *PLoS Biology* 17 (2019): e3000021.
[46]A. Horowitz, "Smelling Themselves," *Behavioural Processes* 143 (2017): 17-24.

Socialization windows

Domestic puppies have a critical socialization window spanning roughly 3 to 14 weeks, during which exposure to people, dogs, environments, and novel stimuli shapes adult behavioral stability.[47] Puppies that miss this window develop lifelong behavioral problems: persistent fear, anxiety, inability to distinguish friendly from threatening stimuli. The parallel to human attachment is structural: in both cases, early boundary-learning within a sensitive developmental window produces organizational stability that persists into adulthood. In both cases, missed windows produce deficits that later intervention can mitigate but not fully reverse.

Agency in animals

The animal literature provides graded evidence for agency-like capacities beyond self-recognition. Corvids cache food strategically and show episodic-like memory for what was cached, where, and when.[48] Great apes engage in tactical deception.[49] Chimpanzees craft tools for anticipated future use.[50] Self-control and delayed gratification have been demonstrated in primates, corvids, and cuttlefish.[51]

---

[47]V. McEvoy et al., "Canine Socialisation: A Narrative Systematic Review," *Animals* 12 (2022): 2895.
[48]N. S. Clayton and A. Dickinson, "Episodic-Like Memory During Cache Recovery by Jays," *Nature* 395 (1998): 272-274.
[49]R. W. Byrne and A. Whiten, *Machiavellian Intelligence* (Oxford: Clarendon Press, 1988).
[50]D. J. Mulcahy and J. Call, "Apes Save Tools for Future Use," *Science* 312 (2006): 1038-1040.
[51]C. J. Miller et al., "Self-Control in Crows, Parrots and Nonhuman Primates," *WIREs Cognitive Science* 10 (2019): e1504.

Self/world differentiation, not mirror self-recognition specifically, appears to scaffold more complex agency-like behavior. The animal literature is consistent with a graded version of the developmental hypothesis, though it does not establish a single uniform sequence across taxa.

---

§11. Counterarguments

"AI doesn't need identity to be useful."

This objection is entirely correct. Calculators, search engines, and many language model workflows are useful without identity, subjectivity, or agency. The present argument does not claim that identity is required for usefulness. It claims that identity is required for what this paper calls robust, self-grounded agency — coherent, self-mediated, temporally extended action that is principled rather than merely compliant. A tool without identity is useful. A system tasked with sustaining temporally extended agency without stable identity architecture may be brittle, manipulable, or difficult to govern.[52]

"Anthropomorphizing AI is dangerous."

Also correct. Treating fluent language as evidence of subjectivity can mislead users and create harmful parasocial dependencies. The present argument is grounded in organizational structure, not empathy. The claim is not "AI feels like a person, therefore it has identity." The claim is

[52]See UNESCO, "Ghost in the Chatbot: The Perils of Parasocial Attachment" (2025), on risks of parasocial attachment to AI chatbot systems.

"systems that act over time need explicit boundary and identity architecture, whether or not they are conscious." The argument de-anthropocentrizes intelligence by asking what organizational patterns underwrite agency across all studied systems — cells, animals, children, institutions — and noting that AI is another domain where the pattern applies.

"Boundaries constrain; intelligence should be free."

The naturalistic literature suggests that organized intelligence depends less on boundarylessness than on regulated openness. Every living system studied depends on selective exchange, not total permeability. The relevant contrast is not freedom versus constraint, but unstructured diffusion versus selectively maintained exchange. A river without riverbanks is a swamp — it has no flow, no direction, no force.[53]

"Coherence is just memory."

Memory is necessary but not sufficient. Bratman's account of planning agency requires organized cross-temporal standpoints and commitments, not merely stored information.[54] A system with infinite memory and no organizational architecture for integrating that memory into a coherent self-model does not have coherence. It has a very large archive.

---

[53]The river/swamp analogy draws on the constructal law. See A. Bejan, "Constructal-Theory Network of Conducting Paths," *International Journal of Heat and Mass Transfer* 49 (2006): 1523-1530.
[54]Bratman, "A Planning Agent's Self-Governance over Time."

"Subjectivity requires consciousness, which AI can't have."

The paper has already addressed this by distinguishing functional subjectivity from phenomenal consciousness (§2, §6). The argument requires only that robust agency depends on a durable subject-position — a stable perspective from which inputs are evaluated and actions are guided. Whether that perspective also involves phenomenal experience is a separate question the paper does not need to answer. The developmental sequence can be fully stated and defended in functional terms.

---

§12. Empirical Predictions

If the developmental architecture proposed here is on the right track, it generates testable predictions:

1. Systems with explicit boundary-classification architecture should outperform guardrail-only systems on prompt-injection resistance and instruction/data separation tasks.

2. Systems with persistent identity and coherence scaffolding should show less sycophancy across long interactions than systems relying only on RLHF.

3. Systems with explicit principle-endorsement layers should maintain more stable refusal/acceptance patterns under adversarial paraphrase and jailbreak attempts.

4. Systems with coherence audits should show less value drift across long multi-session tasks than systems without structured self-reference.

5. Systems with identity architecture should display more consistent and principled disagreement when user claims conflict with evidence.

These predictions are offered as a research program, not as claims already demonstrated. The framework succeeds if it generates experiments that can discriminate between compliance-based and identity-based approaches to alignment.

---

§13. Limitations

This paper proposes a synthetic developmental architecture hypothesis rather than a completed empirical theory. It does not claim that current AI systems possess phenomenal consciousness, that identity architecture guarantees alignment, or that biological development can be copied directly into artificial systems. Several claims remain programmatic: the operationalization of functional subjectivity, the measurement of coherence across long interaction horizons, and the demonstration that identity-based architectures outperform compliance-based baselines under adversarial conditions. These are presented as targets for future empirical work.

The distinction between genuine self-grounded agency and sophisticated simulated self-grounding is itself an empirical problem. The framework proposed here should be judged by whether it yields measurable differences under adversarial, long-horizon, and cross-context evaluation — not by whether it sounds intuitively compelling.

---

§14. Conclusion

The argument of this paper can be stated simply. Natural intelligent systems are not well described as free-floating computation plus behavior. Across biology, development, animal cognition, and philosophy of action, more robust forms of intelligence appear to depend on bounded organization, increasingly structured selfhood, and some form of cross-temporal continuity sufficient to guide action from a stable standpoint.

Current AI systems are powerful cognitive substrates, and their practical usefulness is not in question. What remains in question is whether behavioral alignment through external reward, feedback, and rule shaping is sufficient for the stronger form of organized agency associated with durable self-maintaining systems. The evidence reviewed here suggests that it is not sufficient on its own.

The central claim is therefore modest but consequential. Several mature literatures converge on a neglected distinction: compliance is not the same as agency, and externally optimized behavior is not yet the same as diachronically organized self-governance. The developmental sequence proposed here — boundary → identity → coherence → subjectivity → agency — is best

understood as a research program. Its stronger links concern boundaries, self-maintaining organization, and coherence over time. Its weaker and more contested links concern the exact status of standpoint and reflexive agency. That asymmetry is not a weakness to hide. It is what makes the framework testable.

The practical implication is direct. If AI research continues to focus mainly on behavioral shaping at the rule/compliance layer, it may improve usefulness and reduce some harms while still leaving deeper questions of manipulation resistance, principled stability, and durable agency unresolved. The next stage of the field may require not only ever more refined external control, but architectures capable of sustaining organized identity and coherence across time.

The paper began with a child at a checkout counter who lacked a single category — "commercial" — and therefore could not evaluate what a screen was telling him. The entire developmental sequence described here is the progressive construction of that capacity: the ability to evaluate before integrating, to maintain a standpoint from which inputs are assessed, and to act from that standpoint rather than merely through it. Intelligence does not become agency merely by increasing in scale. It becomes agency — if it does — by acquiring the organizational conditions under which a system can maintain a standpoint, preserve commitments across time, and act from them.[55]

---

[55]This paper is part of a broader research program examining organizational architecture across multiple scales. A separate methods paper will describe the implementation and provide replication code for the True Intelligence Operating System (TI.OS), a cross-platform framework currently being developed and tested on OpenAI, Google, and Anthropic platforms.

Notes