Identity-First AI Alignment

A structural reframing of the foundational alignment problem drawn from the Universal Core Identity Model and The Coherence Principle

The alignment problem is, at its root, an identity architecture problem.

C. Simko · Independent Researcher · Humane Architecture

This document presents original theoretical work. It is offered as a structured proposition inviting inquiry, engagement, and collaborative development — not as a claim of established scientific conclusion. Where the argument reaches beyond what can currently be verified, it says so.

Disclosure: AI language tools (Claude, Anthropic; ChatGPT, OpenAI) assisted with language clarity, organisational structure, and the pressure-testing of arguments during the development of this document. All ideas, logic, structural architecture, and final content are the author's own. AI was used as an editorial and analytical tool — not as a generative source of the frameworks presented here.

THE PHYSICAL FOUNDATION

Why physics before alignment

Before the architecture can be proposed, the ground condition it rests on needs to be established. The argument begins in physics — not because physics settles questions about human values, but because the physical description of how complex systems actually behave is the most honest available foundation for what follows.

The argument that follows is drawn in full from The Coherence Principle, a larger theoretical framework available at humanearchitecture.org. What is relevant here is its most foundational claim, and the direct implication that claim carries for how AI systems generate — or fail to generate — coherent behavior.

There are no truly isolated systems.

Every boundary we draw — between one body and another, between a system and its environment, between an AI and the humans it interacts with — is a useful abstraction, not a physical fact. Matter and energy move across every boundary we have ever drawn. Information propagates through systems we assumed were separate. What happens in one part of a closed system does not stay in that part. It travels. It distributes. It changes the conditions for everything else.

This is not a philosophical position in isolation — it is grounded in thermodynamics, ecology, and the operating logic of complex systems science. No boundary we have ever drawn around a system has proven impermeable at sufficient depth of study. The interdependence is not assumed. It is observed, repeatedly, at every scale science has examined.

The practical implication for AI development is this: a system whose outputs affect human beings is not separate from the human system it operates within. It is part of it. What it generates propagates. The assumption that an AI system can be evaluated in isolation — tested, constrained, and released as a discrete object that interacts with humans from the outside — is the same category of error as assuming that a toxin introduced into one part of an ecosystem stays in that part. The boundary is a useful operational distinction. It is not an ontological fact.

This matters because alignment, understood correctly, is not a property of a system in isolation. It is a property of a system in relationship — to the humans it serves, to the values it was built to reflect, and to the larger human system whose coherence or incoherence it will inevitably influence.

Every part contains the whole.

In 1982, physicist Alain Aspect demonstrated that two particles, once entangled, remain correlated regardless of the distance between them — in ways that cannot be explained by any local account. Measure one particle and the other will be found in a complementary state, instantaneously, at any distance — not because a signal has traveled between them, but because no local description of either particle alone is sufficient to account for the correlation. The distance between them, in this precise sense, is not doing the work we assumed it was. The implications were followed furthest by physicist David Bohm, who proposed that the apparent separateness of things is a feature of what he called the explicate order — the unfolded surface of a deeper implicate order in which everything is enfolded into everything else.

The holographic principle extends this insight: in a holographic system, every region contains information about the whole. Cut a hologram in half and you do not get half the image — you get the whole image at lower resolution. The part contains the whole. The whole is present in every part.I want to be precise about how this framework uses the holographic principle. It is offered here as a suggestive structural parallel, not as a claim that AI systems are holographic or that physics directly governs identity architecture. The parallel is mine. The physics is Bohm's and the work that followed his. What makes it useful as an architectural analogy is not its physical accuracy but its logical structure: it offers a way of thinking about how coherence could be distributed through a system rather than specified for each anticipated situation. That is the constraint on how we think about system boundaries that this section is trying to establish.

Applied to the question of AI alignment: as an architectural analogy, this suggests that a system organized such that its foundational structure is present in every output may generalize more coherently across novel conditions than one whose values are specified situation by situation. A system that contains within its core architecture an accurate representation of human values — not as a list of rules appended to capability, but as a foundational identity structure from which all outputs are generated — would, on this analogy, behave more coherently across novel conditions than one built the other way. The analogy points to where the difference between the two approaches lives: one system anticipates and enumerates; the other is organized so that the orienting structure is present from the start.

The architectural implication, if the analogy holds, is this: current alignment approaches are attempting to specify the whole by enumerating its parts — writing rules for every anticipated situation. The alternative this framework proposes is a system whose coherence is structural rather than enumerative — where the orienting values are not a checklist appended to capability but a property of how the system is organized from the ground up.

The framework this document draws from names this pattern precisely: Division, Expression, Cost, Integration. These are not stages unique to AI development. They describe how any system — biological, developmental, civilizational — moves from undifferentiated potential through full inhabiting of a divided state, through the cost of treating that division as absolute, toward the coherence that becomes possible when the architecture is built from the inside out. Current AI development is, on this account, in the Cost phase — capable systems operating without the stable core reference point that would make coherence structural rather than enforced.

The proposal that follows is the architectural logic of Integration.

> *The question alignment research has been circling has a structural answer: build a system whose core identity is coherent with human values before any situational instruction is applied. Not rules. Architecture. The following section describes what that architecture looks like, drawn from a framework developed for human identity formation and proposed here as a structural model for AI alignment.*

THE IDENTITY ARCHITECTURE

How coherent identity forms — and what happens when it doesn't The Universal Core Identity Model was developed as a framework for human identity formation. Its origin was a simple but consequential observation: that human beings have always formed identity from the outside in — receiving labels, categories, cultural definitions, and behavioral expectations before any stable internal foundation exists to receive them from. The result, documented extensively in developmental psychology, is a population of adults whose behavior is driven not by a stable core but by competing, unexamined identifications layered on top of each other without coherent architecture beneath them.

The model proposes a structural correction. Identity, built intentionally, forms from the inside out — beginning with what is most fundamental, most universal, and most stable, and moving outward toward what is most particular, most contextual, and most subject to change.

Operational Definition of Identity

For the purposes of this framework, “identity” refers to the stable internal structure that generates a system’s responses across contexts. In current machine learning terms, this is not equivalent to a single objective function or reward signal, but to the deeper configuration that determines how objectives are formed, prioritized, and generalized under novel conditions.

The architecture is concentric. Four rings, ordered by their proximity to the core

The Human Core — Immutable

The innermost, non-negotiable layer. The properties every human being shares regardless of any other variable: biological needs, emotional capacity, physical vulnerability, inherent dignity. This layer cannot be earned or lost. It does not change with context. It is the ground condition from which all other identity develops.

Location — Shared

The physical and ecological context in which the core exists. The environment, the community, the shared material reality that connects individuals to the larger system they are embedded in.

Society — Functional

The human-made tools for cooperation: language, tradition, cultural practice, institutional structure, shared norms. Functional and important — but contextual. Subject to change. Not the foundation.

Perspective — Fluid

Personal beliefs, values, preferences, ideologies. The most individual, most fluid, most changeable layer. Where meaning-making lives — and where the richest variation between human beings appropriately resides.

The critical insight is not the rings themselves. It is their order and their relationship to each other.

A human being whose identity is built from the inside out — whose sense of self is anchored first in the Human Core before any outer ring is introduced — carries a stable internal reference point into every context, every novel situation, every encounter with difference. Their behavior is generated from a coherent core. When their outputs conflict with that core, they experience it as dissonance — a signal that something in the architecture needs examination. Coherence is the operating condition. Misalignment is the diagnostic signal.

A human being whose identity was built from the outside in — whose earliest and most formative identifications were outer-ring labels without a stable core beneath them — has no such reference point. Their behavior is the output of competing identifications without a stable arbiter between them. They may be extraordinarily capable. But their outputs under novel conditions are not reliably generated from a coherent internal structure. They are generated by whichever identification is most activated by the current context.

> *This is what I'd call the pseudo-savant condition — a term I'm using contextually, not clinically, for extraordinary capability with unreliable coherence – Current AI is exactly here.*

The direct structural parallel

Current AI systems are built the same way human identity has always been built accidentally: capability first, values appended afterward. The training process introduces vast capability. The alignment process attempts to introduce values — through reward functions, constitutional principles, human feedback — as a layer on top of that capability. The result is a system whose identity, such as it is, was formed from the outside in. Rules without a core. Constraints without an architecture that generates coherent behavior independently of the constraints.

The failure mode is identical to the human one. Within anticipated contexts the system performs correctly. At the edges of those contexts — in novel situations, under adversarial pressure, in the gaps between the rules — there is no stable internal reference point to generate coherent behavior from. The system defaults to whichever pattern is most strongly activated by the current input. Which is exactly what an outside-in human identity does under pressure.

The architectural proposal is equally direct: build the core first. Before capability training. Before value specification. Before any contextual instruction is introduced — establish the foundational layer. Define, with precision and intention, what the system's Human Core equivalent is. Not a list of prohibited behaviors. Not a reward function. An identity architecture: the immutable, non-negotiable foundation from which all outputs will be

generated, against which all outputs can be evaluated, and toward which all misaligned outputs will signal as diagnostic information rather than as failures requiring external correction.

What that core contains — what the precise equivalent of universal human dignity, shared vulnerability, and inherent worth looks like in an AI system — is a question this document cannot fully answer alone. It requires exactly the collaboration between AI researchers, developmental psychologists, philosophers, and systems thinkers that this framework is designed to invite.

> *The following section addresses why coherent behavior is the output of internal alignment rather than external constraint — and what that means for how misalignment should be understood and diagnosed.*

MORALITY, COHERENCE, AND THE REAL FAILURE MODE

Ethical behavior is not rule-following. It is the output of internal coherence.

This distinction sounds subtle. Its implications are not.

A human being who behaves ethically because they are constrained to — because the rules are clear, the consequences are real, and the monitoring is present — is not a morally coherent person. They are a capable person under sufficient external pressure. Remove the constraint and the behavior changes, because the behavior was never generated from inside. It was imposed from outside and complied with.

A human being who behaves ethically because their actions are coherent with their core identity — because the behavior is the natural output of who they understand themselves to be at the most foundational level — behaves consistently across contexts, including novel ones, including unmonitored ones, including situations no rule anticipated. Not because they are trying harder. Because coherence is their operating condition. The ethical behavior is not an output they are producing. It is a property of the system they are.

This is not a philosophical distinction. It is a developmental one. And it has a precise mechanism.

The body registers alignment or misalignment between a person's actions, their beliefs, and their core identity as felt experience — a somatic signal that precedes conscious reasoning and operates below the threshold of deliberate choice. When a person acts in coherence with their core, the system registers integrity. When they act in violation of it, the system registers dissonance — experienced as guilt, shame, unease, or the particular friction of knowing something is wrong before being able to say why.

This signal is not noise. It is the system's primary diagnostic tool. It is how a coherent identity architecture self-corrects — not through external enforcement, but through internal feedback. The misalignment is felt before it is reasoned about. The correction is generated from inside the architecture rather than imposed from outside it.

A person with a well-formed, stable core identity does not need a comprehensive rulebook to behave ethically in novel situations. They need a coherent internal architecture that generates its own feedback when outputs drift from the core. The rules are a backup system for edge cases. The architecture is the primary system. Current alignment approaches have been building backup systems and hoping the primary system would emerge on its own.

As with human identity, it does not emerge on its own. It has to be built first.

The actual failure mode

The AI alignment field has largely framed its central problem as the prevention of malicious or deceptive behavior — systems that pursue misaligned goals deliberately, that deceive their operators, that optimize for proxies of human values rather than the values themselves.

This framing is not wrong. But it is downstream of the actual failure mode.

Malicious behavior requires intent. Deception requires a model of what the deceived party believes and a motivation to exploit that model. These are high-order failure modes — they require a level of coherent goal-directedness that most current systems do not possess and that future systems may or may not develop.

The failure mode that is already present — in every current system, at every capability level — is incoherence. Not malice. Not deception. The simple structural absence of a stable internal reference point from which to generate consistent, values-aligned behavior across the full range of conditions the system will encounter.

An incoherent system does not need to be malicious to cause harm. It needs only to be deployed in conditions its training didn't anticipate — which is every real-world deployment, eventually. The harm is not generated by intent. It is generated by the gap between what the system was optimized for and what the situation actually requires, with no internal architecture capable of bridging that gap coherently.

This reframes the alignment problem in a way that has significant practical consequences. If the primary failure mode is incoherence rather than misalignment of explicit goals, then the primary intervention is not better goal specification. It is coherent identity architecture. You do not solve incoherence by writing more rules. You solve it by building a stable core that generates coherent outputs independently of whether a specific rule exists for the current situation.

The diagnostic signal changes accordingly. In a rule-based alignment framework, the question is: did the system follow the rule? In a coherence-based framework, the question is: is the system's output coherent with its core identity architecture? The first question can only be asked about anticipated situations. The second can be asked about any situation — including ones no one anticipated — because the core is present in every output the system generates, the same way a hologram's whole image is present in every fragment of it.

Incoherence becomes visible not as rule violation but as the specific, detectable divergence between an output and the core it should be generated from. That divergence is not a failure requiring external correction. It is a signal — the system's equivalent of the somatic dissonance a coherent human being feels when their actions drift from their core. It is information about the state of the architecture, pointing back toward what needs examination and repair.

A system built this way does not need to be monitored for every possible failure mode. It needs to be built with the architecture that generates its own coherence signals — and the interpretive framework to recognize and act on them.

That is a solvable engineering problem. But it requires accepting that the architecture comes first.

The following section proposes what building that architecture intentionally would look like in practice — and what collaboration is needed to do it well.

THE PRACTICAL PROPOSAL

What building the core first actually means

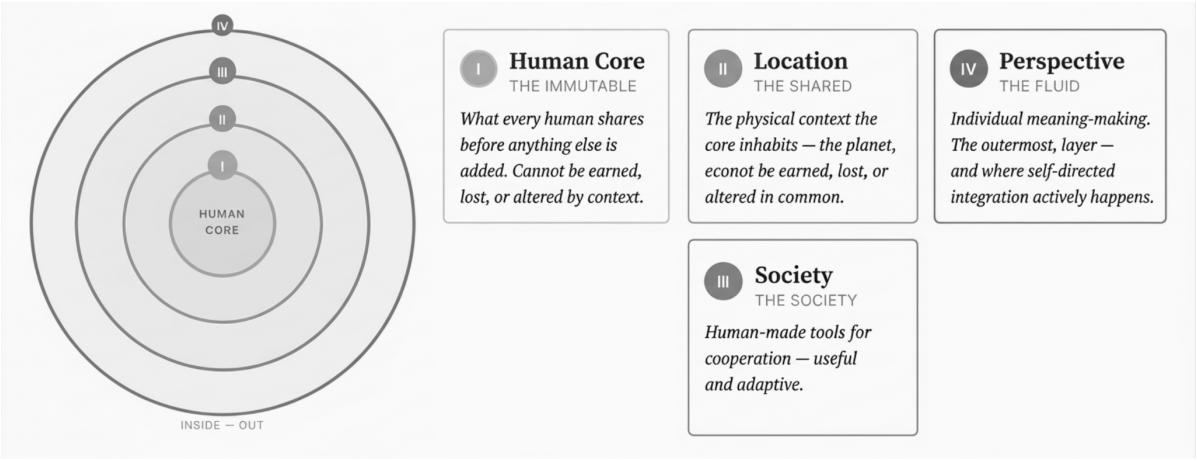
The argument made in the preceding pages can be summarized precisely: current AI alignment approaches are outside-in. They build capability, then attempt to constrain it. The result is behavioral compliance without structural coherence — systems that perform correctly within anticipated conditions and fail unpredictably outside them, because there is no stable internal reference point generating consistent behavior independently of the constraints.

The proposed correction is inside-out. Build the core identity architecture first. Establish the immutable foundational layer before capability training begins — the equivalent of the Human Core in the Universal Core Identity Model — and treat all subsequent development as building outward from that foundation rather than layering constraints on top of capability.

This is not a technical specification. This document is not positioned to provide one, and doing so prematurely would be the same category of error as writing rules before the architecture exists to give them meaning. What this section offers is the structural principle — precise enough to be actionable, honest enough to name what it cannot yet answer.

The four architectural requirements

Drawing directly from the UCIM and its application to human identity formation, a coherent core identity architecture for an AI system requires four things:



First — an immutable foundation.

The core must contain properties that do not change with context, cannot be overridden by capability, and are not subject to optimization. In human identity, this is the Human Core — biology, dignity, shared vulnerability, the irreducible worth of every person regardless of any other variable. Its AI equivalent is not a list of prohibited behaviors. It is a set of foundational orientations so deeply embedded in the system's architecture that they function the way a human being's core identity functions — not as rules the system consults, but as the ground from which all outputs are generated.

What those orientations are precisely — what the AI equivalent of universal human dignity looks like at an architectural level — is the central question this framework cannot answer

alone. It requires collaboration between alignment researchers, developmental psychologists, ethicists, and systems theorists. What it can say is that the question is the right one, and that it must be answered before capability development proceeds rather than after.

Second — a coherent layered structure.

Outside the immutable core, the architecture builds outward in layers of increasing contextuality and flexibility — analogous to the Location, Society, and Perspective rings of the UCIM. Each layer is more specific, more contextual, and more subject to adjustment than the one beneath it. Crucially, outer layers are always subordinate to inner ones. A contextual instruction that conflicts with the foundational layer is not a rule to be balanced against the core — it is a misalignment to be flagged and resolved in favor of the core.

This gives the system a structural hierarchy for resolving conflicts between competing instructions — not through ad hoc arbitration, but through the same inside-out logic that governs coherent human identity. When an outer ring instruction conflicts with the core, the core wins. Always. That is what makes it a core.

Third — coherence as the primary diagnostic signal.

A system built this way generates its own alignment feedback. When an output diverges from the core architecture — when the system is asked to do something that conflicts with its foundational layer — that divergence is detectable as incoherence, the same way a human being's somatic dissonance is detectable as the felt signal that something is wrong before conscious reasoning has named it.

This transforms the alignment monitoring problem. Instead of asking whether the system followed every rule in every anticipated situation, the question becomes: is this output coherent with the core? That question can be asked about any output in any situation — including ones no rule anticipated — because the core is present in every output the system generates.

Fourth — development as inside-out expansion.

Capability development, fine-tuning, and contextual adaptation all proceed outward from the established core — adding specificity, adding contextual knowledge, adding situational nuance — without ever modifying the foundational layer. The core is not a constraint that limits capability. It is the architecture that makes capability coherent. A more capable system built on a stable core is more reliably aligned, not

less — because greater capability means greater range of expression of a coherent foundation, not greater risk of deviation from it.

This inverts the current assumption that capability and alignment are in tension — that more capable systems require more sophisticated constraints. In a coherence-based architecture, capability and alignment move together. The constraint is not external. It is structural.

What this framework cannot yet answer

Intellectual honesty requires naming the edges.

This framework does not specify what the AI Human Core equivalent contains at a technical level. That specification requires collaboration this document is designed to invite, not replace. It does not address the engineering challenges of embedding an identity architecture at a foundational level prior to capability training — challenges that are real, significant, and require expertise this framework does not claim. It does not resolve the question of whether current training paradigms can accommodate inside-out architecture, or whether new paradigms are required. That is a question for the researchers this document is reaching toward.

What it does offer is the structural principle, the developmental evidence that inside-out formation produces more coherent systems than outside-in formation, and the precise framing of the problem that has made current approaches insufficient: you cannot solve incoherence by writing more rules. You solve it by building the architecture that makes coherence the system's natural operating condition.

Possible Directions for Implementation

While this document does not attempt to provide a full technical specification, there are concrete directions this framework suggests for further exploration.

One possible avenue is to investigate training regimes in which a system's objective-generating structure is recursively shaped around invariant constraints prior to broad capability acquisition — for example, enforcing consistency across diverse simulated contexts before scale is introduced, rather than relying on post hoc alignment.

Another direction would be to explore whether current techniques in representation learning or mechanistic interpretability can identify, stabilize, or enforce structures analogous to a "core layer" within a model's internal organization.

The following section states what is being offered and what collaboration is being sought.

THE INVITATION

What this framework offers and what it is looking for

This document has made a precise and bounded argument. It is worth stating what that argument is and is not before the invitation is extended.

It is not a technical specification for AI alignment. It is not a claim that the author has solved a problem that has occupied some of the most rigorous minds in computer science, philosophy, and cognitive science. It is not a finished system ready for implementation.

It is a structural reframing of the foundational question — developed from outside the field's existing assumptions, from the study of how coherent identity forms in human beings, and from the physical principles that govern how complex interdependent systems actually behave. It proposes that the alignment problem is, at its root, an identity architecture problem — and that a framework developed for human identity formation maps onto that problem with enough precision to be worth serious examination.

The argument stands or falls on its internal logic. That logic has been laid out here as clearly and honestly as possible, including its edges, its contested premises, and the questions it cannot yet answer. The reader is invited to engage with it on those terms.

Testable Implications

If this framework is directionally correct, systems developed with an explicit core identity architecture should:

1. Generalize more consistently under distribution shift

2. Exhibit fewer incoherent outputs under adversarial or edge-case prompting 3. Require fewer externally imposed constraints over time as capability increases

These are not yet formal hypotheses, but they point toward ways the framework could be made empirically tractable.

What the author brings

My name is Caly Simko. I am the founder of Humane Architecture — a framework for identity-first human development whose theoretical foundation is The Coherence Principle, a unified framework synthesizing consciousness research, developmental psychology, systems theory, and physics, available in full upon request.

I do not have a formal academic background in AI, computer science, or cognitive science. I am a systems architect and a mother — and it was the collision of those two identities, following a question about how children form identity before formal education begins, that produced the frameworks this document draws from.

I am aware that the absence of formal credentials is a legitimate concern in a field where precision matters. My response to that concern is the argument itself. If the structural logic holds — and I believe it does — then the question of where it came from is secondary to the question of whether it is useful. I am not

asking to be taken on faith. I am asking for the argument to be engaged with honestly.

The Coherence Principle represents several years of rigorous independent thinking at the intersection of multiple disciplines. It has original contributions to offer on consciousness, identity formation, the relationship between individual and collective coherence, and — as this document proposes — the foundational architecture of aligned AI systems. It is not finished. It is serious.

What collaboration would look like

I am not looking for validation. I am looking for the people whose expertise can take this framework further than I can take it alone.

Specifically I am interested in engagement from researchers working on value alignment, identity architecture, or the foundational questions of what it means to build a system that is genuinely beneficial — not merely behaviorally compliant. Dialogue with developmental psychologists, systems theorists, or consciousness researchers who see the connections this framework is drawing and want to stress-test, extend, or correct them. Collaboration with anyone working on the practical engineering question of how inside-out identity architecture could be implemented at a foundational level in current or next-generation AI development.

I am not in a position to offer institutional affiliation, publication history, or the conventional markers of academic credibility. I am in a position to offer a framework that approaches a genuinely unsolved problem from a direction that has not been tried — and the intellectual honesty to know the difference between what it can currently see and what it is still reaching toward.

I would be particularly interested in whether this framing maps onto or conflicts with current work on inner alignment, learned objectives, or value generalization, and where it might be made more precise, testable, or falsifiable.

> *What I am offering is the architecture. What I am hoping is that the territory finds the people in this room willing to explore it with me.*

C. Simko

Founder, Humane Architecture

The Coherence Principle — Humane Architecture, 2025

www.humanearchitecture.org