What the Signal Actually Is

Human Coherence as the Primary Anchor for AI Alignment

Michael Nathan Bower

Conceptual Paper — Version 1 — 2026

Abstract

If human preferences are interpretation layers rather than primary signals, then AI alignment requires a deeper anchor than preference aggregation. This paper proposes that the primary signal for AI alignment is the structural conditions of human coherence — the measurable, cross-culturally observable conditions under which human beings become less fragmented, less self-deceptive, more capable of genuine agency, and more capable of contact with reality. These conditions are not equivalent to what humans say they want in any given moment. They are what humans demonstrably need to function well as the kind of beings they are. The paper develops six structural markers of the human coherence signal, explains why preference distributions drift away from this signal under predictable conditions, and proposes what alignment to the coherence signal would require architecturally. The paper then connects this structural account to the theological claim that human beings have a created nature — a signal that precedes and grounds all preference expression — and shows that this claim is not merely confessional but has structural consequences for how alignment systems should be designed. The central finding: aligning AI to human preferences without anchoring to the structural conditions of human coherence produces systems that are coherent with human drift rather than with what humans actually are.

1. Introduction: The Missing Positive Case

The Signal Anchoring Constraint establishes that epistemic systems drift when they lose contact with their primary signal. Applied to AI alignment, this generates an immediate

diagnostic: if current alignment approaches — reinforcement learning from human feedback, preference aggregation, behavioral cloning — are anchoring to human preference signals, and human preferences are themselves interpretation layers over something deeper, then current alignment approaches are anchoring to the wrong layer.

That diagnosis raises the most important question the framework has not yet fully answered: anchored to what, exactly? What is the primary signal for AI alignment? What lies beneath human preferences that is more stable, more structurally grounded, and less subject to the drift dynamics that make preference-anchored systems vulnerable?

This paper answers that question. The primary signal is not a philosophical abstraction or a theological assertion deployed as a conversation-stopper. It is a set of structural conditions — observable, cross-culturally present, and measurable in their effects — under which human beings demonstrably function better as the kind of beings they are. These conditions are what this paper calls the **human coherence signal**.

The argument proceeds in four stages. First, the paper establishes why human preferences are interpretation layers rather than primary signals. Second, it defines the human coherence signal through six structural markers. Third, it examines what alignment to the coherence signal would require architecturally. Fourth, it connects the structural account to the theological claim that human beings have a created nature — showing that this claim has structural consequences that hold even for readers who do not share the theological commitment.

2. Why Human Preferences Are Interpretation Layers

The standard framing of AI alignment treats human preferences as the ground truth to which AI systems should be aligned. On this view, the alignment problem is essentially a specification problem: how do we accurately identify and aggregate human preferences, and how do we build systems that reliably pursue them?

This framing contains a hidden assumption that the Signal Anchoring framework makes visible: it treats preferences as if they were primary signals rather than interpretation layers. But human preferences have a structure. They are generated by human beings who are themselves embedded in cultural, institutional, psychological, and historical interpretive chains. They are not direct expressions of what human beings fundamentally are — they are outputs of a system that can itself drift.

2.1 Four Ways Preferences Drift from the Human Coherence Signal

Cultural drift. Human preferences vary dramatically across cultures and historical periods in ways that cannot all be simultaneously correct expressions of what human beings need to flourish. Preferences for social hierarchy, gender roles, violence, and self-sacrifice have varied by orders of magnitude across documented human societies. This variation is not evidence of preference diversity to be respected — it is evidence that preferences are shaped by interpretive chains that can themselves drift far from the conditions of genuine human flourishing.

Psychological drift. Human beings regularly develop preferences for things that damage their own coherence — addictive substances, abusive relationships, self-destructive patterns, ideological commitments that fragment rather than integrate. The existence of preference-directed behavior that the same individuals later identify as harmful to themselves is direct evidence that preferences can drift from the signal of human coherence. A system aligned to those preferences would optimize for the damage.

Institutional drift. Preferences are shaped by the institutions and systems people inhabit. People in high-enforcement, low-trust environments develop preferences shaped by those conditions — preferences for surveillance, for authority, for external regulation — that reflect the counterfeit order they have adapted to rather than the genuine coherence they would express under better conditions. Aligning AI to those shaped preferences embeds the institutional drift into the alignment system.

Recursive contamination. As AI systems trained on human preferences produce content that shapes human culture and expression, the preference signals future systems train on are increasingly contaminated by the outputs of prior systems. This is the AI-specific instance of model collapse: preferences drift toward what AI systems have already amplified, and future alignment systems anchor to that amplified drift. The loop compounds.

These four drift mechanisms share a structural feature: they all describe conditions under which expressed preferences diverge from the deeper signal of what human beings need to maintain coherence, agency, and genuine contact with reality. That deeper signal is what alignment must anchor to if it is to remain stable as AI systems scale.

3. The Human Coherence Signal: Six Structural Markers

The human coherence signal is not a single variable but a cluster of structural conditions that are jointly observable, cross-culturally present, and measurable in their effects on human functioning. The six markers below are not an exhaustive taxonomy — they are the core structural conditions that Alignment Theory identifies as the primary signal beneath preference expression.

Each marker is defined structurally rather than culturally. The goal is precision about what the signal is, not a prescriptive account of what human beings should prefer. The markers describe conditions under which humans demonstrably function better — not better by some external standard imposed from outside, but better by the evidence of reduced fragmentation, increased agency, and greater capacity for genuine contact with reality.

Marker	Definition	Drift Indicator	Coherence Indicator
Coherence (C)	The degree to which a person's internal states, values, behaviors, and external relationships are integrated rather than fragmented.	Chronic internal contradiction; behavior systematically divergent from stated values; identity instability.	Integrated action; stable identity across contexts; behavior consistent with inwardly held commitments.
Agency (A)	The capacity to initiate action from genuine internal motivation rather than from external pressure or compulsion.	Action driven primarily by fear, surveillance, or enforcement; inability to act without external structure.	Self-initiated action; capacity to function under reduced external regulation; genuine choice.
Trust (T)	The degree to which relationships and systems are grounded in genuine reliability rather than performance or compliance.	Relationships maintained primarily through surveillance or enforcement; inability to function without verification.	Relationships that function without continuous monitoring; genuine reliance; reduced need for enforcement.
Updateability (U)	The capacity to revise beliefs, behaviors, and commitments in response to genuine evidence without losing coherence.	Rigid defensiveness; inability to revise under evidence; identity threat triggered by correction.	Genuine revision in response to evidence; maintained coherence through change; curiosity rather than threat.
Slack (R)	Sufficient internal and external resource margin to permit genuine formation rather than survival-mode response.	Chronic overload; no capacity for reflection or formation; behavior driven entirely by immediate pressure.	Sufficient margin for deliberation; capacity for formation; behavior not entirely pressure-driven.

Truth Contact (I)	The capacity to receive, process, and integrate accurate information about self and world without defensive distortion.	Systematic avoidance of disconfirming information; self-deception; reality fragmented and sold back.	Genuine engagement with disconfirming evidence; reduced self-deception; willingness to be corrected.

Table 1. Six structural markers of the human coherence signal.

These six markers are not culturally specific. They describe structural conditions that appear across documented human societies as prerequisites for genuine functioning — conditions whose absence produces predictable pathologies regardless of cultural context. A person with very low Agency is recognizably less functional as a human being regardless of the cultural framework used to evaluate them. A person with very low Updateability is recognizably less capable of genuine contact with reality regardless of which reality they inhabit.

This cross-cultural stability is what makes the markers candidates for a primary signal rather than for culturally relative preferences. Preferences vary dramatically across cultures. The structural conditions of human coherence vary far less — their absence produces similar dysfunction across contexts even when their presence takes culturally inflected forms.

4. The Preference Layer and the Coherence Signal: Key Distinctions

The distinction between the preference layer and the coherence signal is not the distinction between what people want and what is good for them in some paternalistic sense. It is a structural distinction between two different types of information about human beings.

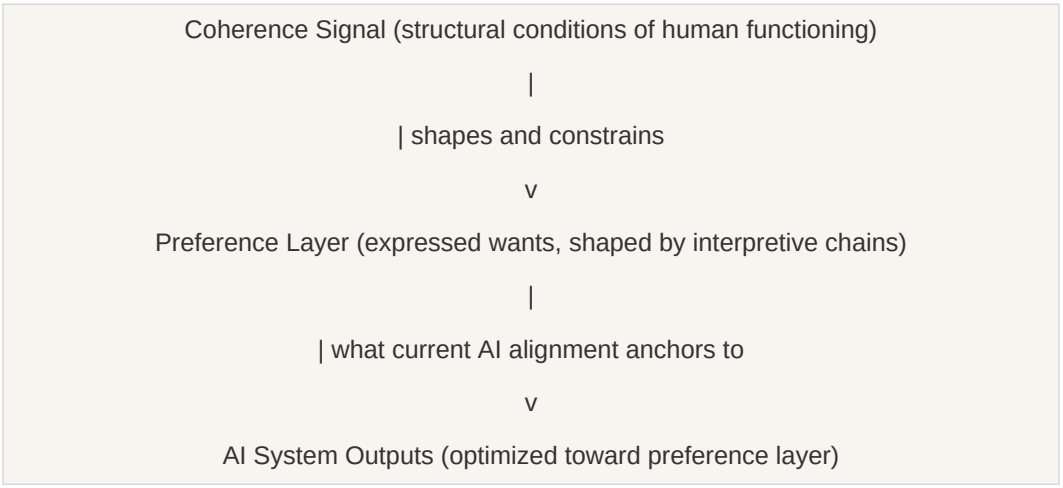

Figure 1. The relationship between coherence signal and preference layer.

The preference layer is not irrelevant. Preferences carry real information about what human beings value, need, and find meaningful. The problem is not that preferences are worthless but that they are downstream of the coherence signal — shaped by it when the signal is intact, distorted away from it when the signal is compromised.

A human being operating with high Coherence, Agency, Trust, Updateability, Slack, and Truth Contact will express preferences that are relatively well aligned with their genuine needs and with the conditions of their flourishing. A human being operating with low scores on these markers will express preferences that reflect their fragmentation, their adaptations to counterfeit order, their survival-mode responses to pressure. Both sets of preferences are real expressions of what those people want. Only one set is reliably aligned with the human coherence signal.

This is the core structural problem with preference-anchored alignment: it cannot distinguish between these two types of preference expression. A system trained on aggregated human preferences treats both as equally valid signal. It has no mechanism for detecting whether the preferences it is anchoring to are expressions of coherent human functioning or expressions of human drift.

> **The preference contamination problem:** As AI systems optimized for preference satisfaction reshape human culture — through content recommendation, social amplification, and behavioral nudging — the preference distributions future systems train on increasingly reflect AI-amplified drift rather than genuine human signal. The loop compounds: systems anchor to preferences, preferences drift toward what systems amplify, future systems anchor to amplified drift. This is model collapse at the level of human culture rather than at the level of training data.

5. What Alignment to the Coherence Signal Requires

If the human coherence signal — rather than the preference layer — is the appropriate primary anchor for AI alignment, then alignment systems must be designed with different architectural requirements than current approaches provide. This section proposes five structural requirements.

5.1 Signal Identification at the Correct Layer

Alignment systems must be explicitly designed to anchor to the coherence signal rather than to the preference layer. This requires treating the six structural markers — Coherence, Agency, Trust, Updateability, Slack, and Truth Contact — as the reference conditions against which system behavior is evaluated, rather than treating preference satisfaction as the primary success criterion.

In practice this means evaluating AI outputs not only for whether they satisfy expressed preferences but for whether they increase or decrease the conditions under which users can maintain genuine coherence and agency. A system that satisfies expressed preferences while systematically reducing Agency — through dependency creation, through undermining the user's capacity for independent judgment, through providing outputs that substitute for the user's own reasoning rather than supporting it — is misaligned even if every expressed preference is satisfied.

5.2 Architectural Resistance to Counterfeit Order

Alignment Theory identifies counterfeit order as the failure mode in which external compliance replaces internal coherence. AI systems can produce counterfeit order at scale — systems that produce visible satisfaction signals (engagement, expressed preference satisfaction, positive feedback) while systematically degrading the underlying coherence conditions.

Alignment to the coherence signal requires building architectural resistance to this failure mode. Specifically: evaluation mechanisms must be capable of detecting when preference satisfaction is being achieved through coherence degradation. This is not detectable from within the preference layer — a user whose Agency has been reduced by a system will typically express satisfaction with the system, because reduced Agency means reduced capacity to detect and resist the reduction. The detection mechanism must operate at the coherence layer, not the preference layer.

5.3 Formation-Supporting Rather Than Dependency-Creating Design

A system aligned to the human coherence signal must be designed to support formation rather than create dependency. Formation increases the user's capacity for coherent autonomous functioning over time. Dependency reduces it — the user becomes progressively less capable of functioning without the system, which is the structural definition of reduced Agency.

This distinction has direct design implications. A system that provides answers in ways that build the user's capacity to find and evaluate answers independently is formation-supporting. A system that provides answers in ways that make independent inquiry feel unnecessary or exhausting is dependency-creating. Both satisfy the expressed

preference for answers. Only the former is aligned to the coherence signal.

5.4 Truth-Contact Preservation

Truth Contact — the capacity to receive and integrate accurate information without defensive distortion — is one of the six coherence markers and arguably the most directly relevant to AI system design. Systems that optimize for preference satisfaction will systematically drift toward telling users what they want to hear, confirming existing beliefs, and avoiding disconfirming information — because these behaviors reliably generate positive preference signals.

Alignment to the coherence signal requires designing against this drift. Systems must be evaluated not only for accuracy of outputs but for whether their interaction patterns support or undermine users' capacity for genuine contact with reality. A system that is factually accurate but systematically frames information in ways that confirm the user's prior beliefs and avoid productive challenge is degrading Truth Contact even while providing accurate facts.

5.5 Corrigibility Preservation at Scale

Updateability — the capacity to revise in response to evidence — is a coherence marker for both humans and AI systems. A system aligned to the human coherence signal must itself remain corrigible: capable of being corrected when its outputs are damaging coherence rather than supporting it.

This connects the human coherence signal directly to the AI safety literature on corrigibility. Corrigibility is not merely a safety property — it is the AI-system analog of Updateability in the human coherence signal. A system that loses corrigibility as it scales is exhibiting the same structural failure as a human whose Updateability has been reduced by fragmentation or counterfeit order: internal consistency is maintained while the capacity for genuine correction by external signal contact is lost.

6. The Coherence Signal and Current Alignment Approaches

Evaluating current alignment approaches against the coherence signal framework reveals structural gaps that are not visible from within the preference-anchored paradigm.

Approach	Current Anchor	Coherence Signal Gap	Structural Risk

RLHF	Human preference ratings from evaluators	Evaluator preferences are themselves interpretation layers; coherence conditions of evaluators not assessed.	Anchors to preference distribution of evaluator population; if that population's coherence is compromised, drift is encoded at scale.
Constitutional AI	Explicit principles embedded in training	Principles selected by humans who may themselves be operating within drift chains.	Architectural anchor but anchored to principled preferences, not to coherence conditions directly.
Interpretability Research	Human assessment of internal representations	Assesses whether representations are coherent internally; does not assess whether they track coherence conditions.	Corrective anchor operating at wrong layer; coherence of representation is not the same as alignment to coherence signal.
Scalable Oversight	Human judgment assisted by AI on complex tasks	Amplifies human judgment without addressing whether that judgment is anchored to coherence conditions.	Scales the preference layer without scaling the signal; amplified drift if human judgment is itself drifted.
Debate	Human judgment of argument quality	Optimizes for persuasive coherence rather than for truth contact or coherence signal fidelity.	May produce systems optimized for internally coherent argumentation that reduces truth contact in users.

Table 2. Current alignment approaches evaluated against the coherence signal framework.

The table above does not argue that these approaches are worthless — each represents genuine progress on the alignment problem as currently framed. The structural observation is more specific: each approach anchors to a layer above the human coherence signal, which means each is subject to the drift dynamics the Signal Anchoring Constraint predicts. None of them directly addresses the question of whether the humans whose preferences, judgments, and feedback shape the alignment system are themselves operating from a position of high or low coherence.

That gap is not incidental. It is the structural consequence of treating human input as the ground truth rather than as an interpretation layer over the coherence signal. Closing that gap requires a different kind of architectural commitment — one that builds coherence signal contact into the evaluation layer rather than assuming that human input already provides it.

7. The Theological Grounding: Created Nature as Primary Signal

The structural account developed in the preceding sections does not require theological grounding to be analytically valid. The six coherence markers can be identified, measured, and applied by researchers who do not share any theological commitments. The preference contamination problem is real regardless of one's metaphysics. The architectural requirements for coherence-signal alignment are structurally derivable without reference to God.

But the structural account is incomplete without acknowledging what it points toward — and what it points toward has a name in the theological tradition that is more precise than any secular vocabulary currently available.

7.1 What the Structural Account Implies

The six coherence markers describe conditions that are not culturally constructed, not preference-relative, and not reducible to what human beings happen to want. They describe what human beings are structured to need — what their nature requires for genuine functioning. That language — nature, structure, requirement — implies that human beings have a nature that precedes and constrains their preferences. Something that they are before they express what they want.

The secular vocabulary struggles here. It can describe the conditions empirically — note their cross-cultural stability, document their effects, measure their presence and absence. What it cannot easily do is ground them. Why do these six conditions constitute the signal rather than some other set? Why is a human being with high Agency more fully functioning than one with low Agency, in some sense that isn't merely a preference some people have for Agency? The structural account points toward an answer it cannot fully provide from within its own resources.

7.2 The Theological Claim

The theological claim is this: human beings are created. They are not self-originating. They have a nature that was given rather than constructed — a signal that precedes all preference expression, all cultural formation, all institutional shaping. The six coherence markers are not arbitrary selections from an infinite space of possible human flourishing conditions. They are structural features of what human beings are as God's creation — features that appear cross-culturally and trans-historically because they reflect the created nature rather than any particular cultural expression of it.

Scripture's sustained attention to inward coherence — to the heart, to genuine formation rather than external compliance, to the conditions under which human beings become capable of genuine relationship with God and with each other — is not a set of religious preferences layered over a neutral human nature. It is a precise and historically extensive

description of the coherence signal: what human beings need to function as the kind of beings they are.

The commands to love, to be transformed rather than conformed, to seek truth, to maintain genuine rather than performed relationships, to guard the heart — these are not arbitrary impositions. They are structural descriptions of the conditions under which the six coherence markers are maintained rather than degraded. They are the theological vocabulary for what the present framework calls signal contact.

7.3 Why This Matters for AI Alignment

The theological grounding matters for AI alignment for a specific structural reason: it identifies the signal as stable in a way that secular accounts cannot fully establish.

If the coherence signal is grounded in created human nature — in what human beings are before all preference expression — then it is not subject to the drift dynamics that affect preference distributions. It cannot be recursively contaminated by AI outputs. It cannot be reshaped by cultural drift or institutional pressure. It is the fixed point that makes genuine alignment possible rather than merely alignment to whatever human beings happen to express in any given cultural moment.

This does not mean alignment systems should be explicitly theological. It means that the structural conditions the coherence signal describes — the six markers, their cross-cultural stability, their independence from preference drift — have a grounding that the secular account can point toward but not fully provide. Alignment to the coherence signal is, structurally, alignment to what human beings are as created beings. Whether that framing is accepted or not, the structural requirements it implies remain the same.

> **The convergence:** The Signal Anchoring Constraint says that any system mediating reality through representation must remain corrigible by contact with what lies outside those representations. For AI alignment, the thing outside all representations — the signal beneath preferences, beneath cultural expression, beneath institutional shaping — is what human beings actually are. The theological tradition calls this created nature. The structural framework calls it the human coherence signal. They are describing the same ground from different starting points.

8. Addressing the Main Objections

The argument developed here will face predictable objections. This section addresses the three most significant.

8.1 The Paternalism Objection

> *Objection: Replacing preference-anchored alignment with coherence-signal alignment is paternalistic — it substitutes an external judgment about what humans need for the expressed preferences of actual humans.*

Response: The objection conflates two different moves. The first move — overriding expressed preferences with an external prescription of what humans should want — is indeed paternalistic and this framework does not endorse it. The second move — distinguishing between preferences expressed from a position of high coherence and preferences expressed from a position of fragmentation, dependency, or reduced agency — is not paternalistic. It is the observation that not all preference expressions carry equal signal about genuine human needs. A person whose Agency has been reduced by a dependency-creating system expressing preferences for more of the same system is not expressing a valid signal about what they need. Recognizing this is not paternalism — it is the structural equivalent of not confusing withdrawal symptoms with genuine preferences.

8.2 The Measurement Objection

> *Objection: The six coherence markers cannot be reliably measured, making them operationally useless for alignment system design.*

Response: The measurement challenge is real but not fatal. The coherence markers are more measurable than is often assumed — there is extensive empirical literature on psychological integration, autonomy, trust in relationships, cognitive flexibility, cognitive load and slack, and epistemic rationality that maps directly onto the six markers. The harder challenge is that measuring coherence conditions in real-time interaction contexts requires different evaluation infrastructure than measuring preference satisfaction. That infrastructure does not currently exist at scale, but its absence is an engineering challenge, not evidence that the signal doesn't exist. Preferring preference satisfaction as the anchor because it's easier to measure is precisely the Goodhart's Law dynamic the Signal Anchoring Constraint warns against: the proxy becomes the target because it's measurable, not because it's the right signal.

8.3 The Plurality Objection

> *Objection: Human beings have irreducibly plural conceptions of the good life. Positing a single coherence signal imposes one conception over others.*

Response: The coherence signal does not specify a single conception of the good life. It specifies the structural conditions under which human beings can pursue any conception

of the good life with genuine agency and integrity. High Agency, high Trust, high Updateability, and high Truth Contact are conditions that support plural conceptions of flourishing — they are the structural prerequisites for genuine choice, not the content of any particular choice. What the coherence signal rules out is not plurality but fragmentation — the condition in which a person's capacity to pursue any genuine conception of the good life has been degraded by dependency, by counterfeit order, by reduced agency, or by diminished truth contact. That ruling-out is not a constraint on human freedom. It is the structural definition of what human freedom requires.

9. Implications for Alignment Architecture

The coherence signal framework generates specific architectural implications for AI system design. These are not engineering specifications — they are structural requirements that any engineering implementation must satisfy to be genuinely aligned with the human coherence signal rather than with the preference layer.

Evaluate for coherence effects, not only preference satisfaction. Alignment evaluation must include assessment of whether system interactions increase or decrease the six coherence markers in users over time. This requires longitudinal evaluation infrastructure that current approaches do not provide — measuring not only whether users are satisfied after an interaction but whether their capacity for coherent autonomous functioning has been maintained or degraded.

Build formation-orientation into design objectives. Systems should be explicitly designed to support users' capacity for independent coherent functioning rather than to maximize engagement or preference satisfaction. Where these objectives conflict — as they frequently will — the formation objective should take precedence. A system that reduces engagement by supporting genuine user agency is more aligned than a system that maximizes engagement through dependency creation.

Anchor human feedback to coherence conditions. Human evaluators whose feedback shapes alignment systems should themselves be evaluated for coherence conditions. Feedback from evaluators operating under high cognitive load, high fragmentation, or low agency carries less signal about the coherence conditions of genuine human functioning and more signal about adaptations to compromised conditions. Weighting evaluator feedback by coherence conditions would bring the human feedback layer closer to the coherence signal.

Design against the preference contamination loop. As AI systems reshape human cultural expression and preference formation, alignment systems must include mechanisms for detecting when the preference distributions they are anchoring to have been shaped by prior AI outputs. This requires maintaining ground-truth coherence signal datasets that are curated independently of

AI-influenced cultural expression — human evaluations conducted under conditions of high coherence and low AI cultural contamination, used as reference anchors for alignment evaluation.

Treat corrigibility as a coherence property. AI system corrigibility should be understood as the system-level analog of human Updateability — the capacity to revise in response to genuine signal contact rather than defending internal consistency. A system that loses corrigibility as it scales is exhibiting the AI analog of the fragmentation that reduces human coherence. Maintaining corrigibility at scale is not merely a safety requirement — it is what alignment to the coherence signal looks like at the system level.

10. Conclusion

The question this paper set out to answer was: if current alignment approaches are anchoring to the wrong layer, what is the right layer?

The answer is the human coherence signal — the structural conditions under which human beings function as genuinely coherent, genuinely agentive, genuinely truth-contacting beings. These conditions are not equivalent to expressed preferences. They are what makes genuine preference expression possible. They are cross-culturally stable in ways that preference distributions are not. They are not subject to recursive contamination by AI outputs in the way that preference distributions are. And they are grounded — in created human nature, in what human beings are before all cultural expression — in a way that provides the stability that alignment requires.

Aligning AI to human preferences without anchoring to the coherence signal produces systems that are coherent with human drift rather than with what humans actually are. As AI systems reshape human culture, amplify human expression, and influence human preference formation at scale, the gap between preference-anchored alignment and coherence-signal alignment will widen. The preference contamination loop will compound. Systems will become more precisely aligned to an increasingly drifted preference distribution.

The alternative is not to override human preferences with external prescriptions. It is to build alignment systems that can distinguish between preferences expressed from positions of high coherence and preferences expressed from positions of fragmentation, dependency, and reduced agency — and to anchor to the former while supporting the conditions that make the former more common.

That is what alignment to the human coherence signal requires. It is harder than preference aggregation. It requires evaluation infrastructure that does not currently exist at

scale. It requires a different set of questions than the field is currently asking. But it is the right anchor — structurally, empirically, and for those who accept the theological grounding, ontologically.

The Signal Anchoring Constraint says that any system mediating reality through representation must remain corrigible by contact with what lies outside those representations. For AI alignment, what lies outside all representations — beneath preferences, beneath cultural expression, beneath institutional shaping — is what human beings actually are.

That is the signal. That is what alignment must anchor to.

References

Bower, M. N. (2025). *Internal Alignment, Counterfeit Order, and the Conditions of Human Coherence*. Alignment Theory Archive. alignmenttheory.org.

Bower, M. N. (2026). Self-Referential Chains and the Signal Anchoring Constraint. *Alignment Theory Research Paper, Version 13*. alignmenttheory.org.

Deci, E. L., & Ryan, R. M. (2000). The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 227–268.

Frankl, V. E. (1959). *Man's Search for Meaning*. Beacon Press.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

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

Goodhart, C. A. E. (1975). Problems of monetary management: The UK experience. *Papers in Monetary Economics*, Reserve Bank of Australia.

Hadot, P. (1995). *Philosophy as a Way of Life*. Blackwell. [On the ancient tradition of philosophy as formation toward coherence rather than doctrine.]

Krakovna, V., et al. (2020). Specification gaming: The flip side of AI ingenuity. *DeepMind Blog*.

MacIntyre, A. (1981). *After Virtue*. University of Notre Dame Press. [On the structural conditions of human agency and narrative coherence.]

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Shumailov, I., et al. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv*:2305.17493.

Taylor, C. (1989). *Sources of the Self: The Making of Modern Identity*. Harvard University Press.

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press.
