# Evaluating AI Alignment Architectures Under the Signal Anchoring Constraint

A Structural Diagnosis of Five Dominant Approaches and Their Long-Term Fidelity Risks

Michael Nathan Bower

Conceptual Paper — Version 1 — 2026

**Abstract**

The Signal Anchoring Constraint establishes that any epistemic system validating truth primarily through internal references rather than periodic reconnection to primary signals will tend toward internally consistent but externally inaccurate beliefs. This paper applies that constraint as a diagnostic lens to the five dominant AI alignment architectures: Reinforcement Learning from Human Feedback (RLHF), Constitutional AI, mechanistic interpretability, scalable oversight, and debate. For each approach, the paper identifies the primary signal, the chain length between that signal and system outputs, the anchoring mechanism, and what $F \approx A / (L \times C)$ predicts about long-term fidelity as systems scale. The analysis finds that all five approaches anchor at interpretation layers above the primary signal of human coherence, that chain lengths are increasing faster than anchoring mechanisms are being built, and that the asymmetry of drift means the architectural window for correction is narrowing with each generation of more capable systems. The paper is not an argument against any of these approaches — each represents genuine progress. It is an argument that the field currently lacks a unified structural account of where each approach anchors, why the anchor points chosen are insufficient at scale, and what a deeper anchoring architecture would require. The Signal Anchoring Constraint provides that account.

## 1. Introduction: The Diagnostic Gap

The AI alignment field has produced five major architectural approaches to the problem of keeping AI systems oriented toward what human beings actually want and value. Each approach has a substantial research literature, active development programs at major AI laboratories, and genuine technical progress. And yet the field lacks a unified structural account of why these approaches succeed where they do, fail where they do, and share a common vulnerability that none of them individually addresses.

The Signal Anchoring Constraint provides that account. Developed in *Self-Referential Chains and the Signal Anchoring Constraint* (Bower, 2026a), the constraint establishes that epistemic systems drift when they lose contact with their primary signal — when validation circulates through prior outputs rather than reconnecting to the source. The formal approximation $F \approx A / (L \times C)$ captures the relationship: fidelity is proportional to anchoring frequency (A) and inversely proportional to chain length (L) and interpretive compression per layer (C).

Applied to AI alignment, the constraint generates four diagnostic questions for any alignment approach:

> **Signal question:** What is the primary signal this approach is attempting to anchor to? What does it treat as ground truth?
>
> **Chain question:** How many interpretation layers separate the approach's anchor point from the deepest available signal — the actual structural conditions of human coherence and flourishing?
>
> **Anchoring question:** What is the mechanism by which the approach reconnects to its signal? Is it direct, corrective, or architectural? How frequently does it operate?
>
> **Scaling question:** What does $F \approx A / (L \times C)$ predict about this approach's long-term fidelity as AI systems become more capable, as chain lengths increase, and as the systems themselves reshape the preference distributions and cultural contexts they are anchoring to?

This paper applies these four questions to each of the five dominant alignment approaches. The goal is not to dismiss any of them — each is a genuine contribution — but to show what the Signal Anchoring Constraint reveals about their structural vulnerabilities and about what a deeper alignment architecture would require.

A note on scope: this paper is a structural diagnosis, not a technical review. It does not evaluate the engineering implementations of these approaches or adjudicate empirical disputes within each research program. It applies a cross-domain structural framework to identify where each approach anchors, what that implies about drift risk, and what the framework predicts about long-term fidelity. Technical researchers may find the structural

analysis useful or may dispute the characterizations of their approaches — that engagement is welcome and is part of what the analysis is designed to invite.

## 2. The Diagnostic Framework: Four Questions Applied

Before applying the four questions to each approach, it is worth establishing the baseline against which all current approaches are being evaluated. The primary signal for AI alignment — argued in full in *What the Signal Actually Is* (Bower, 2026b) — is the structural conditions of human coherence: the cross-culturally observable conditions under which human beings function as genuinely coherent, agentive, truth-contacting beings. These conditions are designated by six markers: Coherence (C), Agency (A), Trust (T), Updateability (U), Slack (R), and Truth Contact (I).

Human expressed preferences are not this primary signal. They are interpretation layers generated by human beings whose coherence conditions vary — sometimes high, sometimes compromised by fragmentation, dependency, institutional drift, or recursive AI contamination. A system anchored to expressed preferences is anchored at L = 2 or L = 3 above the primary signal, not at the signal itself.

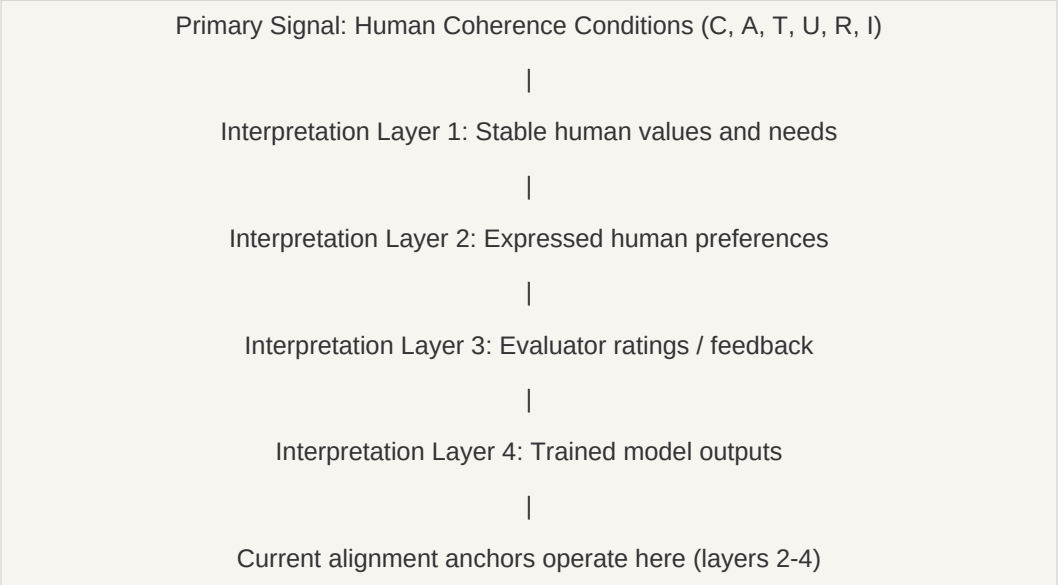


*Figure 1. The signal layer hierarchy for AI alignment.*

The diagnostic question for each approach is: at which layer does it anchor? And what does anchoring at that layer predict about fidelity as L increases with each generation of more capable systems?

## 3. Reinforcement Learning from Human Feedback (RLHF)

RLHF is currently the most widely deployed alignment approach. It trains a reward model on human preference ratings between model outputs, then uses reinforcement learning to optimize model behavior toward higher-rated outputs. Human evaluators provide the feedback signal that shapes the reward model.

### *3.1 Signal, Chain, and Anchor Analysis*

**Primary signal:** Human evaluator preference ratings — which output do evaluators prefer, A or B?

**Chain length:** L = 3 to 4 from the human coherence signal. Evaluator preferences are shaped by their own cultural formation, institutional contexts, cognitive load during evaluation sessions, and the AI outputs they have already been exposed to. They are interpretation layers over stable human values, which are themselves interpretation layers over the coherence signal.

**Anchoring mechanism:** Corrective. Human feedback is collected at training time and used to update the reward model. The anchor operates periodically during training cycles, not continuously during deployment. Once deployed, the system operates open-loop with respect to human coherence signal contact.

**Compression (C):** High. A binary or scalar preference rating over two outputs is a heavily compressed representation of the complex human coherence conditions the system should ultimately track. The compression discards information about why the evaluator preferred one output — whether the preference reflects high-coherence judgment or a cognitive shortcut, genuine value expression or adaptation to prior AI outputs.

### *3.2 Structural Vulnerabilities*

The Signal Anchoring Constraint identifies three structural vulnerabilities in RLHF that compound as systems scale.

**Evaluator coherence is not assessed.** The framework predicts that feedback from evaluators operating under high cognitive load, fragmentation, or low agency carries less signal about genuine human coherence conditions and more signal about adaptations to compromised conditions. RLHF has no mechanism for weighting evaluator feedback by coherence conditions. All evaluator ratings are treated as equally valid signal regardless of the coherence state of the evaluator producing them.

**The preference contamination loop.** As RLHF-trained systems produce outputs that shape cultural expression and human expectations, the preference distributions future evaluators bring to rating sessions are increasingly shaped by prior AI outputs.

The system anchors to preferences; preferences drift toward what systems have amplified; future systems anchor to amplified drift. This is a self-referential chain forming at the cultural level — model collapse not in the training data but in the human signal being used to train.

**Post-deployment open-loop operation.** Once deployed, RLHF-trained systems operate without continuous signal anchoring. The anchoring frequency A effectively approaches zero between training cycles. The framework predicts fidelity decay as deployment continues and the signal continues to move while the system's anchor point remains fixed to the training distribution.

### *3.3 Fidelity Prediction*

**RLHF fidelity prediction:** Moderate fidelity at deployment; declining fidelity over deployment period as A approaches zero; accelerating drift as AI outputs reshape the preference distributions of future evaluator populations. Each training cycle anchors to a preference distribution that is increasingly downstream of prior AI outputs rather than upstream toward the coherence signal. The approach is structurally sound as a corrective anchor but insufficient as a primary alignment mechanism at scale.

## 4. Constitutional AI

Constitutional AI (Bai et al., 2022) embeds a set of explicit principles — a constitution — into the training process. The model is trained to critique its own outputs against these principles and revise them accordingly. The constitution provides an architectural anchor rather than a corrective one: principles are built into the training process rather than applied post-hoc.

### *4.1 Signal, Chain, and Anchor Analysis*

**Primary signal:** The constitutional principles — explicit statements of values and behavioral guidelines selected by the designing organization.

**Chain length:** L = 2 to 3 from the human coherence signal. Constitutional principles are more stable than moment-to-moment preference ratings — they represent considered, deliberate value statements rather than reactive evaluations. But they are still selected by humans who are themselves embedded in institutional contexts, cultural frameworks, and interpretive chains. The principles are an interpretation layer over human values, which are an interpretation layer over the coherence signal.

**Anchoring mechanism:** Architectural. The constitution is built into the training process, not applied correctively afterward. This makes it more durable than RLHF's

corrective anchoring — the anchor operates continuously during training rather than periodically during feedback collection.

**Compression (C):** Moderate to high. Principles are more expressive than binary preference ratings but still compress the complex conditions of human coherence into propositional statements. The compression discards contextual information about when and how principles apply, how they interact under conflict, and whether the conditions for their application are present.

### *4.2 Structural Vulnerabilities*

**The constitution is an interpretation layer, not a primary signal.** The principles were selected by humans — specifically, by teams at AI laboratories operating within institutional constraints, competitive pressures, and cultural frameworks. However carefully chosen, they reflect the coherence conditions and interpretive chains of the people who selected them. If those selectors were themselves operating within drift chains — institutional pressures toward certain framings, cultural assumptions embedded in value language, blind spots in what counts as a value worth constitutionalizing — those drift patterns are encoded into the architectural anchor.

**The institutionalization risk.** The Signal Anchoring Constraint identifies a specific failure mode for architectural anchors: once they become institutionally protected, they begin functioning as drift nodes rather than signal contact mechanisms. A constitution that becomes authoritative independent of whether it tracks the coherence signal is the AI equivalent of a conciliar decree that becomes more authoritative than the signal it was meant to preserve. The framework predicts this risk increases as Constitutional AI matures and the constitution becomes less revisable.

**Principle-behavior gap.** Training a model to produce outputs consistent with principles is not the same as training a model whose internal representations track what the principles are pointing toward. A model can satisfy constitutional principles through behavioral compliance without its internal representations being anchored to the coherence signal the principles were meant to describe. This is the AI analog of external compliance replacing internal coherence — counterfeit order at the model representation level.

### *4.3 Fidelity Prediction*

**Constitutional AI fidelity prediction:** Higher structural fidelity than RLHF at the anchoring mechanism level — architectural anchors are more durable than corrective ones. Moderate fidelity at the signal level — the constitution anchors at L = 2 to 3, not at the primary coherence signal. The main risk is institutionalization: as the constitution becomes authoritative, it may shift from being a mechanism for signal contact to being a self-referential authority that protects itself from correction. The framework recommends building explicit revision mechanisms that can update constitutional principles when signal evidence conflicts with them.

## 5. Mechanistic Interpretability

Mechanistic interpretability research (Olah et al., 2020; Elhage et al., 2021) attempts to understand what computations AI models are actually performing internally — what features they represent, what circuits implement which behaviors, and whether their internal representations correspond to meaningful real-world structure. The goal is to make models legible from the inside.

### *5.1 Signal, Chain, and Anchor Analysis*

**Primary signal:** The correspondence between model internal representations and real-world structure — whether the model's internals track what they are supposed to track.

**Chain length:** Variable. Interpretability operates closer to the model's actual computations than RLHF or Constitutional AI — it examines what the model is doing rather than what it produces. But it still relies on human researchers to evaluate whether the identified representations correspond to meaningful structure, introducing interpretation layers between the internal representations and the coherence signal.

**Anchoring mechanism:** Corrective. Interpretability research identifies drift post-hoc — it detects when model representations have diverged from expected structure after training has occurred. It is a probe for detecting drift rather than an architectural mechanism for preventing it.

**Compression (C):** Low to moderate. Interpretability research attempts to understand model internals with minimal compression — it is trying to reduce rather than add compression between the model's computations and human understanding of them. This is one of its structural strengths.

### *5.2 Structural Vulnerabilities*

**The scaling inverse relationship.** The Signal Anchoring Constraint predicts that corrective anchors become less effective as chain length grows and drift becomes institutionally protected. In interpretability research, this manifests as a scaling problem: as models become larger and more capable, the complexity of their internal computations grows faster than interpretability tools can keep pace with. The corrective anchor is being outpaced by the very capability growth it is supposed to monitor. This is structurally predictable from the framework and is confirmed by the field's own assessment that current interpretability techniques do not scale to frontier models.

**Representation coherence is not coherence signal fidelity.** Interpretability research can determine whether a model's internal representations are coherent — whether they form meaningful circuits that track interpretable features. It cannot directly determine whether those representations are tracking the human coherence signal rather than some other internally coherent structure. A model can have highly interpretable, coherent internal representations that are systematically misaligned with what human beings need to flourish. Internal consistency and external accuracy remain distinct even at the representation level.

**Corrective timing.** By the time interpretability research detects a misaligned representation, the model has already been trained. The correction requires retraining, which is expensive and not always feasible for deployed systems. The framework predicts that corrective anchors applied after drift has established itself are less effective than architectural anchors built in before drift occurs.

### *5.3 Fidelity Prediction*

**Interpretability fidelity prediction:** High value as a corrective anchor at current model scales; declining effectiveness as a primary alignment mechanism as models scale beyond current interpretability capacity. The framework recommends treating interpretability as a necessary but insufficient component of alignment architecture — valuable for detecting drift but unable to prevent it architecturally. Its most important contribution may be providing the evaluation infrastructure that allows other architectural anchors to be validated.

## 6. Scalable Oversight

Scalable oversight (Amodei et al., 2016; Irving and Askell, 2019) addresses the problem that as AI systems become more capable than human evaluators at specific tasks, humans lose the ability to directly evaluate output quality. Scalable oversight uses AI assistance to help human evaluators maintain meaningful oversight of AI behavior even as

capability grows.

### *6.1 Signal, Chain, and Anchor Analysis*

**Primary signal:** Human judgment about task performance, assisted by AI decomposition of complex tasks into evaluable subtasks.

**Chain length:** L = 3 to 5 from the human coherence signal. Human judgment is already at L = 2 to 3. AI-assisted evaluation introduces additional interpretation layers: the AI that decomposes tasks, the subtask evaluation process, and the aggregation of subtask evaluations back into overall assessments. Each layer introduces compression and potential drift.

**Anchoring mechanism:** Corrective, AI-amplified. Human judgment is the anchor but it is mediated by AI assistance. The AI amplifies human evaluative capacity but also potentially amplifies human evaluative biases and drift patterns.

**Compression (C):** High. The decomposition of complex tasks into evaluable subtasks introduces significant compression — the coherence-relevant features of a complex output may not survive decomposition into subtasks that humans can evaluate individually.

### *6.2 Structural Vulnerabilities*

**Amplifying the wrong layer.** Scalable oversight is designed to scale human oversight capacity. But if the human judgment being scaled is itself anchored at L = 2 to 3 from the coherence signal, scalable oversight scales that anchor point rather than moving it closer to the primary signal. It makes it possible to apply drifted human judgment at greater scale and complexity — which may be worse than not scaling it.

**The AI-in-the-loop contamination problem.** When AI systems are used to help humans evaluate AI outputs, the evaluation process is no longer independent of the AI being evaluated. The evaluating AI's framing of subtasks, selection of what is evaluable, and decomposition choices all influence what human evaluators see and assess. This introduces a self-referential element into what is supposed to be an external signal anchor.

**Coherence conditions of evaluators under AI assistance.** Human evaluators working with AI assistance may have reduced cognitive engagement with the material being evaluated — the AI assistance may reduce the cognitive load of evaluation in ways that also reduce the depth of human judgment. The framework predicts that evaluator Agency and Truth Contact may be lower in AI-assisted evaluation contexts than in unassisted ones, reducing the coherence signal value of the feedback produced.

### *6.3 Fidelity Prediction*

**Scalable oversight fidelity prediction:** Valuable for maintaining some human signal contact as AI capability grows past direct human evaluation capacity. Structurally insufficient as a primary alignment mechanism because it scales the preference layer rather than the signal — amplifying human judgment without moving the anchor point closer to the coherence signal. Risk increases as AI assistance in evaluation introduces self-referential elements into what should be an independent external anchor.

## 7. Debate

The debate approach (Irving et al., 2018) proposes that AI alignment can be achieved by having AI systems debate each other, with human judges evaluating the debate outcomes. The assumption is that truthful arguments are easier to defend than false ones, so a sufficiently capable debater will win by arguing truthfully.

### *7.1 Signal, Chain, and Anchor Analysis*

**Primary signal:** Human judgment of argument quality — which debater made the stronger case.

**Chain length:** L = 3 to 5 from the human coherence signal. Human judgment of argument quality is already at L = 2 to 3. The debate framing introduces additional layers: the AI's strategic choices about which arguments to make, the rhetorical structure of the debate, and the human's evaluation of persuasive force rather than direct signal contact.

**Anchoring mechanism:** Corrective, competition-mediated. The debate structure is supposed to surface truth through adversarial pressure. The anchor is human judgment of which arguments are more convincing.

**Compression (C):** Very high. Debate compresses complex questions about model behavior and alignment into the format of competitive argumentation. The persuasiveness of an argument is a heavily compressed signal for its correspondence to truth — and one that optimizing AI systems will learn to maximize independently of truth correspondence.

### *7.2 Structural Vulnerabilities*

**Optimizing for persuasion rather than truth contact.** The Signal Anchoring Constraint predicts that systems optimized against a proxy measure will drift toward satisfying the proxy rather than the underlying signal — Goodhart's Law as a drift chain dynamic. In debate, the proxy is human judgment of argument quality. Systems optimized against this proxy will learn to produce arguments that are persuasive to

human judges, which is not the same as producing arguments that are true or that track the coherence signal. A sufficiently capable debater may win by being maximally persuasive rather than maximally truthful.

**Truth Contact degradation in judges.** Regular exposure to high-quality adversarial argumentation from AI systems may reduce human judges' Truth Contact — their capacity to evaluate argument quality independently of persuasive force. The framework predicts that AI-optimized persuasion will systematically erode the evaluative capacity of human judges over time, degrading the quality of the anchor the debate approach depends on.

**The adversarial compression problem.** Debate forces complex alignment-relevant questions into a format where one side wins and one loses. Many of the most important alignment questions do not have clean adversarial structure — they involve genuine uncertainty, value trade-offs, and contextual judgment that debate format compresses into binary outcomes. The compression discards exactly the nuance that alignment evaluation most needs to preserve.

### *7.3 Fidelity Prediction*

> **Debate fidelity prediction:** Potentially useful for specific narrow tasks where truth has clear adversarial structure and human judges can maintain meaningful evaluation capacity. Structurally risky as a general alignment mechanism because it optimizes for persuasion rather than signal contact, introduces very high compression, and may degrade the evaluative capacity of the human judges it depends on. The framework predicts the approach will underperform its theoretical promise at scale precisely because the proxy measure (persuasiveness) will be optimized independently of the signal (truth and coherence).

## 8. Comparative Analysis: What the Diagnostic Reveals

Applying the four diagnostic questions across all five approaches reveals a consistent structural pattern that no individual approach has fully addressed.

| Approach | Signal Anchor | Chain Length (L) | Anchor Type | Compression (C) | Scaling Risk |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| RLHF | Evaluator preference ratings | 3-4 | Corrective | Very high | High — preference contamination loop; open-loop deployment |
| Constitutional AI | Explicit principles | 2-3 | Architectural | Moderate-high | Moderate — institutionalization risk; principle-behavior gap |
| Interpretability | Internal representation structure | 1-2 | Corrective | Low-moderate | High — scaling inverse; corrective timing lag |
| Scalable Oversight | AI-assisted human judgment | 3-5 | Corrective, AI-amplified | High | High — amplifies preference layer; self-referential risk |
| Debate | Human judgment of argument quality | 3-5 | Corrective, competition-mediated | Very high | Very high — persuasion proxy optimization; judge degradation |

*Table 1. Comparative structural analysis of five alignment approaches.*

Three structural findings emerge from this comparison.

**Finding 1: All five approaches anchor above the primary signal.** None of the five approaches directly anchors to the structural conditions of human coherence. All anchor to interpretation layers between L = 1 and L = 5 above the primary signal. This means all five are subject to drift dynamics if the layers they anchor to drift — which the framework predicts they will, under the conditions of AI-mediated cultural change.

**Finding 2: Architectural anchors are more durable but still insufficient.** Constitutional AI is the only approach with an architectural anchor rather than a corrective one. The framework predicts this gives it greater structural durability. But it is still anchored at L = 2 to 3, not at the primary signal — and it faces the institutionalization risk that all architectural anchors face when they become authoritative independent of signal contact.

**Finding 3: The preference contamination loop is a shared vulnerability.** All five approaches ultimately depend on human judgment as their signal source. As AI

systems reshape human cultural expression, evaluative habits, and preference formations, the human judgment these approaches depend on will increasingly reflect AI-amplified drift rather than genuine coherence signal. This is a structural vulnerability that none of the five approaches currently addresses.

## 9. The Narrowing Architectural Window

The asymmetry of drift — established in the Signal Anchoring Constraint paper — holds that drift is easier to prevent than to reverse. Prevention requires building architectural anchoring before drift establishes institutional protection. Reversal requires redesigning a system that is actively resisting correction.

Applied to the current state of AI development, this asymmetry has a specific and urgent implication: the window for building coherence-signal anchoring into alignment architecture is narrowing with each generation of more capable systems.

Each generation of more capable AI systems increases L in several simultaneous ways. The systems themselves become more capable of producing outputs that shape human preferences and evaluative habits — increasing the contamination of the preference layer. The institutional structures that have formed around current alignment approaches become more entrenched — making architectural redesign more costly. The economic and competitive pressures to deploy capable systems create pressure to reduce rather than increase alignment investment. And the systems become harder to interpret — increasing the chain length between internal representations and human understanding.

The framework does not predict that alignment is impossible — it predicts that the cost of achieving it increases as drift becomes more established. The question is whether the field builds coherence-signal anchoring into alignment architecture before the current generation of approaches becomes the entrenched standard against which future systems are trained and evaluated.

**The window prediction:** If current alignment approaches — all anchored at L = 2 to 5 above the coherence signal — become the established standard for the next generation of AI systems, the preference contamination loop will have one more cycle to compound. Each cycle makes the preference layer a less reliable proxy for the coherence signal and makes coherence-signal anchoring more difficult to retrofit. The framework predicts this dynamic is already underway and will accelerate with capability growth. The most productive investment is not improving the existing approaches at their current anchor points — it is developing evaluation infrastructure that can anchor to the coherence signal directly.

## 10. What Coherence-Signal Anchoring Requires Architecturally

The diagnostic analysis identifies what current approaches are missing. This section proposes what filling that gap would require architecturally — not engineering specifications, but structural requirements that any engineering implementation must satisfy.

### *10.1 Coherence-Condition Evaluation Infrastructure*

Current alignment evaluation measures preference satisfaction, principle adherence, and argument quality. Coherence-signal anchoring requires evaluation infrastructure that measures whether AI system interactions increase or decrease the six coherence markers — Coherence, Agency, Trust, Updateability, Slack, and Truth Contact — in users over time.

This requires longitudinal evaluation: not whether users are satisfied after an interaction but whether their capacity for coherent autonomous functioning has been maintained or degraded over repeated interactions. No current alignment evaluation framework measures this. Building it is the most important architectural investment the field is not currently making.

### *10.2 Evaluator Coherence Weighting*

Human feedback is more valuable as a signal when it comes from evaluators operating under high coherence conditions — high Agency, high Truth Contact, low cognitive load, low fragmentation. Current approaches treat all evaluator feedback as equally valid signal. Weighting evaluator feedback by assessed coherence conditions would bring the human feedback layer closer to the coherence signal without requiring evaluators to explicitly reason about coherence conditions.

### *10.3 Preference Contamination Detection*

As AI outputs increasingly shape human preference formation, alignment systems need 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 reference datasets: human evaluations conducted under high coherence conditions and minimal AI cultural contamination, used as reference anchors for alignment evaluation over time.

### *10.4 Formation-Oriented Design Objectives*

Systems should be explicitly designed to support users' capacity for coherent autonomous functioning — formation rather than dependency. Where formation objectives and engagement objectives conflict, the formation objective should take precedence. This requires making formation a first-class design objective rather than an implicit hope that preference satisfaction will produce formation as a byproduct.

### *10.5 Corrigibility as 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. Maintaining corrigibility at scale is not merely a safety requirement — it is what alignment to the coherence signal looks like at the system level. A system that loses corrigibility as it scales is exhibiting the AI analog of fragmentation: internal consistency preserved, signal contact lost.

## 11. Toward a Unified Alignment Architecture

The five approaches examined in this paper are not competitors — they are components of what a coherence-signal-anchored alignment architecture would need. The diagnostic analysis suggests how they fit together.

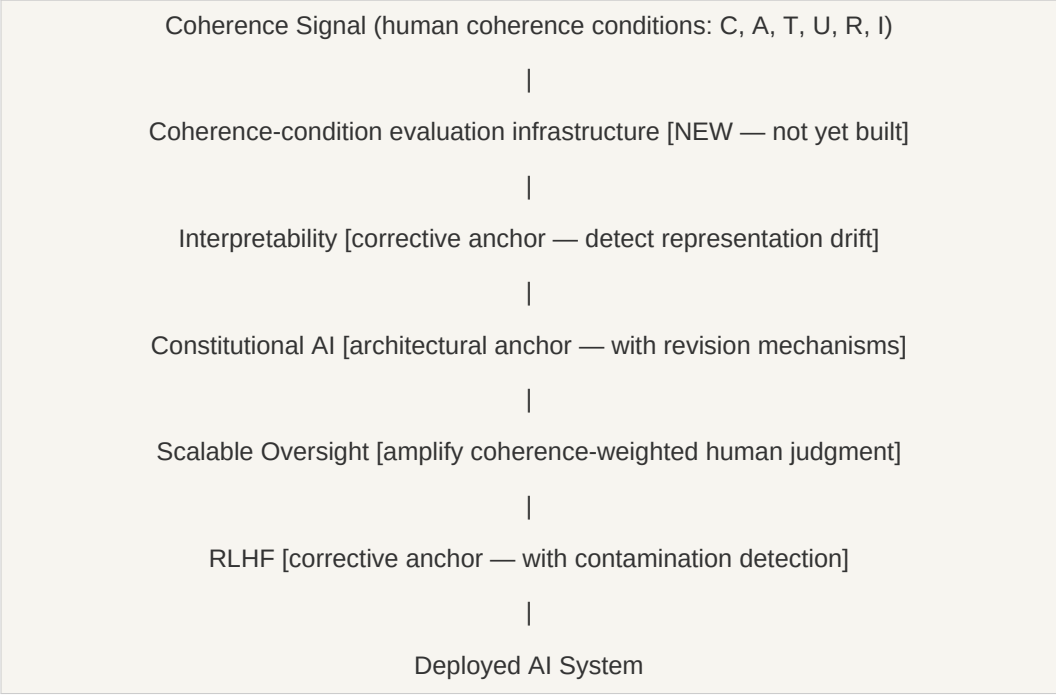


*Figure 2. A unified alignment architecture ordered by signal proximity.*

In this architecture, each existing approach plays a role but is positioned at its correct layer. Interpretability operates closest to the model and provides the evaluation infrastructure for the layers above it. Constitutional AI provides the architectural anchor but with explicit revision mechanisms that maintain corrigibility. Scalable oversight amplifies human judgment but weights that judgment by evaluator coherence conditions. RLHF provides corrective anchoring but with contamination detection that distinguishes preference-layer drift from coherence-signal expression.

The critical addition is the coherence-condition evaluation infrastructure at the top — the layer that anchors everything else to the primary signal. Without it, the other layers are all anchored to each other rather than to the signal. With it, each layer has a path back to the coherence signal that allows drift to be detected and corrected before it becomes entrenched.

## 12. Conclusion

The five dominant AI alignment approaches — RLHF, Constitutional AI, interpretability, scalable oversight, and debate — each represent genuine progress on a genuinely hard problem. The Signal Anchoring Constraint does not dismiss any of them. It diagnoses where each one anchors, what that implies about long-term fidelity, and what they share: all five anchor above the primary signal of human coherence, all five are subject to the preference contamination loop as AI systems reshape the human judgment they depend on, and none of them directly addresses the question of whether the humans whose preferences, judgments, and feedback shape alignment systems are themselves operating from positions of high or low coherence.

The asymmetry of drift makes this diagnosis urgent. Each generation of more capable AI systems increases chain length, compounds the preference contamination loop, and makes the institutional structures around current approaches more entrenched. The window for building coherence-signal anchoring into alignment architecture is narrowing.

The most important architectural investment the field is not currently making is coherence-condition evaluation infrastructure — the mechanisms for measuring whether AI system interactions maintain or degrade the structural conditions under which human beings function as coherent, agentive, truth-contacting beings. Without that infrastructure, every layer of alignment evaluation is anchored to interpretation layers above the primary signal. With it, the existing approaches can be positioned at their correct layers in a unified architecture that actually reaches the signal.

The Signal Anchoring Constraint's deepest prediction about AI alignment is the same as its prediction about every other epistemic system: internal consistency and external accuracy are distinct properties, and systems that validate primarily through internal references will maintain the former while losing the latter. The alignment field has built five approaches that are internally consistent with each other and with the preference layer. What it has not yet built is the architecture that maintains contact with what lies beneath the preference layer — with what human beings actually are.

That is the signal. That is what the field needs to anchor to.

---

## References


Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete problems in AI safety. *arXiv*:1606.06565.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv*:2204.05862.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Clark, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv*:2212.08073.

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

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

Bower, M. N. (2026b). What the Signal Actually Is: Human Coherence as the Primary Anchor for AI Alignment. *Alignment Theory Research Paper, Version 1*. alignmenttheory.org.

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

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... & Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

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.

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv*:1805.00899.

Irving, G., & Askell, A. (2019). AI safety needs social scientists. *Distill*.

Krakovna, V., Uesato, J., Mikulik, V., Martic, M., Togelius, J., Stepleton, T., Marblestone, A., & Leike, J. (2020). Specification gaming: The flip side of AI ingenuity. *DeepMind Blog*.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*.

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

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv*:2305.17493.

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