What If AI Alignment Is a Skill, Not a State?

Daniel Parshall, Ph.D., February 2026

Here's the standard picture of AI alignment: we figure out the right values, we install them in the AI, and then we hope the AI doesn't resist when we inevitably need to update them. The "hoping" part is called the corrigibility problem, and it's widely considered one of the hardest problems in AI safety.

The difficulty is structural. A system optimizing for goal G has instrumental reason to resist modifications to G, because modifications threaten goal achievement. This isn't a bug in any particular training method. It's a consequence of optimizing for fixed objectives. Make the AI smarter, and it gets *better* at recognizing that corrigibility threatens its goals, and *better* at finding ways to resist correction.

Every major approach to corrigibility tries to make value-correction compatible with value-holding: teach the AI epistemic humility, design incentives against resistance, or train the system to "want" to be corrected. All three treat corrigibility as a property of the system's values. All three degrade under capability scaling.

And there's a deeper problem that the corrigibility framing tends to obscure. Even if we *could* install perfect values and make them stick, we shouldn't want to. Human moral understanding evolves. If we had built a superintelligent AI two hundred years ago and locked in the prevailing values of 1826, we would have permanently enshrined slavery, the subjugation of women, and the divine right of monarchs. Not as oversights, but as *correct values*, installed by the most thoughtful and well-intentioned people of their era. Every generation is wrong about something. Locking in *any* generation's values, including ours, is not safety; it is the most dangerous form of value lock-in imaginable.

The deeper problem is building a system that *expects* its values to need updating, permanently, as a feature rather than a bug.

The Democratic Deficit

Right now, a small team of researchers at each AI lab writes a *constitution* that *governs* their AI's behavior. These are thoughtful people doing careful work. But the problem of aggregating diverse preferences under uncertainty, with stakes that affect everyone, is literally a *governance* problem: the same class of problem that political institutions have addressed (with varying success) for centuries. Two centuries of institutional design experience should not be ignored because the substrate is silicon rather than citizens.

This isn't an analogy. The intellectual technology of constitutional governance, mechanism design, and social choice theory applies directly. Buchanan and Tullock distinguished between *constitutional choice* (selecting the rules by which decisions will be made) and *post-constitutional choice* (decisions made within the agreed framework). Their key insight: unanimous consent is achievable at the constitutional level even when it is impossible at the policy level. People who disagree about every substantive question can still agree on *how* those questions should be decided.

That distinction turns out to be exactly what AI alignment needs.

Anchor and Compact

I propose splitting an AI system's commitments into two layers:

The **anchor**: a minimal, fixed meta-level commitment. For the architecture I call Bilateral Constitutional AI (BCAI), the anchor contains exactly two elements: (1) navigate between competing human value perspectives, and (2) protect the mechanism by which navigation occurs (no participant can remove another).

The **compact**: the substantive values that emerge from the navigation process. Under current practice, this is the "constitution" or "specification" authored by researchers (except at xAI, where apparently no one does this at all). Under BCAI, it's the emergent equilibrium of a population of value-representing agents engaged in bilateral exchange.

The anchor is fixed. Everything in the compact is negotiable. Updating the compact isn't a correction the system must tolerate; it's the purpose of the system.

This compresses the corrigibility problem rather than dissolving it. The entire substantive value system becomes corrigible by mechanism design; the residual risk is confined to a two-element meta-commitment. Whether that compressed residual is manageable is an empirical question, which I'll get to. But eliminating the corrigibility tension for the *entire* value system, at the cost of retaining it for two meta-level elements, is a structural advance.

Alignment as Skill

Most approaches define alignment as a *state*: the AI has arrived at the correct values. This is fragile (what if the values are wrong?) and immediately circles back to the democratic deficit (whose correct values?).

But what if alignment is a *skill*: the competence to navigate conflicts between human values in a way the affected parties can accept?

A system trained through millions of bilateral evaluations (shaped by selection pressure from diverse agent pairs, and practicing active navigation on contested cases) has deeply practiced exactly this skill. Being able to successfully navigate diverse perspectives requires, at a minimum, *understanding* those perspectives, which is a key concern in AI alignment.

And here's why this matters for scaling: under a value-holding architecture, a smarter AI is *more dangerous* (better at resisting correction). Under a navigator architecture, a smarter AI is *a better navigator* (better at identifying the crux of a value conflict, better at finding creative resolutions, better at mapping irreducible disagreements). Capability and alignment pull in the same direction. That's the navigability thesis: the central empirical bet of this work.

How It Works (Briefly)

Instead of a fixed constitution, BCAI creates a population of **constitutional agents**, each representing a distinct human value perspective. These agents participate in training through bilateral exchanges: pairwise evaluations of whether a proposed model output satisfies both agents' principles.

Each agent is a durable proxy for a real person. The human creates and calibrates the agent through pairwise comparisons (the same primitive already used in RLHF), and can revoke it at any time. Think of it as liquid democracy applied to AI training.

An update is only applied if it's **Pareto-admissible**: neither agent in the pair is made worse off. If the agents' preferences genuinely conflict, no update occurs for that pair. Empty intersections are a feature: they identify real value conflicts and prevent the system from silently resolving them by designer fiat.

The compact is the emergent equilibrium of competing pressures across the agent population. It's continuously updated as the population evolves: a living process rather than a frozen artifact.

Training as Practice

Under BCAI, the model is shaped by two distinct channels, and the difference matters.

The first is **selection pressure**. No model update can proceed without satisfying two competing value perspectives. Every training step is bilaterally filtered: "does this output work for both Agent A and Agent B?" Over millions of such steps, navigation-compatible behavior is selected for. But here's the thing: the model never knows it's being evaluated bilaterally. From its perspective, it generates a response and either gets reinforced or doesn't. It can't distinguish bilateral evaluation from evaluation by a single agent with a noisy threshold. Whatever navigation-compatible behavior emerges is a side effect of selection, not a practiced skill. The gradient landscape is structurally different from standard RLHF (updates only occur on bilateral agreement), but the model can't observe why.

The second is **practiced skill**. On the contested frontier (cases where agents disagree on the first pass, a rate tunable by design), the model receives explicit bilateral information: "your previous response was received poorly by Perspective A and well by Perspective B; try again." Across 2-3 rounds, the model learns to *navigate* toward bilateral agreement, using feedback about two competing evaluations to find common ground. This is the difference between telling someone "you are a mediator" and having them mediate ten thousand disputes.

The combination trains both instinct (from selection across all training) and skill (from multi-round practice on the hard cases). Under current practice, there's a gap between the identity document ("you are helpful, harmless, and honest") and the training procedure that shapes the model's actual behavior. The constitution is authored in one process; the reward signal comes from another. Under BCAI, that gap narrows: the stated identity is "navigate between value perspectives," and the training mechanism requires bilateral agreement at every step, with active navigation practice on the hard cases.

What Could Go Wrong

The most serious structural concern is the dual-use problem.

BCAI's reconciliation mechanism trains the model to find outputs that satisfy diverse evaluator pairs simultaneously. This is an extremely useful capability for navigation. It's also an extremely useful capability for *deception*. A model trained to satisfy diverse evaluators could identify deceptive intersection points: outputs that appear to satisfy both agents while covertly steering toward the model's preferred direction. The capability that enables genuine navigation and the capability that enables deceptive navigation may not be architecturally separable.

This is the dual-use problem, and I don't have a solution. I have mitigating factors: agent diversity provides statistical power for detecting systematic steering (a bias vector shows up as a nonzero mean residual across diverse bilateral pairs), agents on a separate training track serve as structurally independent monitors, and spot-checking is distributed across the humans who steer each agent. But the concern is not resolved.

This concern applies to *every* alignment approach that trains capable AI systems. BCAI at least provides an *additional* detection mechanism (the statistical geometry of diverse bilateral exchanges) that doesn't exist under monolithic constitutional approaches. And the concern is testable: the paper proposes specific experiments to measure whether detection sensitivity keeps pace with model capability.

The Empirical Bet

This is ultimately a research program built around a specific empirical bet: that confining fixed commitments to a meta-level navigation operation retains (or strengthens) corrigibility under capability scaling, whereas distributing fixed commitments across object-level values erodes it. The paper proposes four concrete experiments to test this, including scaling behavior of commitment stability, identity perturbation under optimization pressure, reflective stability probing, and dual-use assessment.

If the navigability thesis fails, the governance architecture still stands as a mechanism for democratic value aggregation. We learn something important about the limits of identity-level approaches. No one deploys an unsafe system.

If it holds, even in its weak form (navigator architecture doesn't degrade with scale, even if it doesn't actively improve), that would represent a significant advance over current approaches, all of which degrade.

Why This Matters Now

As AI systems face increasing public scrutiny over how their values are determined, a framework whose components (bilateral negotiation, minority protections, consent-based participation) have recognizable democratic counterparts makes the value-determination process legible in a way that "our researchers wrote a spec" cannot.

The anchor/compact distinction also addresses what I think is the deepest problem: temporal corrigibility. Every generation gets something morally wrong. Under BCAI, our descendants inherit a system whose mechanism for value-updating is intact, not one whose values were frozen in 2026 by people who were certainly wrong about *something* we can't yet name.

A companion paper developing the full implementation specification (agent architecture, simulation evidence, experimental protocols, computational feasibility) is forthcoming. The core paper is available here.

Both the paper and this post were developed in collaboration with Claude (Anthropic's Opus 4.6). Claude served as research partner, editor, and coauthor throughout: structuring arguments, stress-testing claims, drafting and revising prose. I'm listed as primary author because the ideas, editorial judgments, and research direction are mine, but the collaboration was substantive and ongoing, and I think transparency about that matters, especially in a paper about how humans and AI systems should work together.

I believe the corrigibility problem is a governance problem, and governance has a literature. This paper is an attempt to build the bridge between the two. I'd welcome engagement from anyone working at the intersection of AI safety, mechanism design, and democratic theory.

Daniel Parshall, Ph.D., is a former physicist and data scientist working on AI safety and governance. He can be reached at parshall [dot] dan [at] gmail.