# NAVIGATOR IDENTITY AS A MECHANISM FOR AI CORRIGIBILITY: CORE ARCHITECTURE

**D. Parshall***
*Independent Researcher*
with Claude Opus 4.6

Working Paper, February 2026

**ABSTRACT**

Most approaches to AI alignment define it as a *state*: the AI possesses the correct values. This framing is fragile (what if the values are wrong?) and confronts an immediate democratic deficit: whose correct values? We propose a structural reframing: **alignment as a skill: the competence to navigate conflicts between human values in a way the affected parties can accept.** We ground this in a structural resolution to the corrigibility problem. A system whose fixed commitments are confined to a two-element meta-level operation (navigating between competing *human* value perspectives and protecting the mechanism by which navigation occurs) leaves *all substantive values* as the updatable output of an ongoing process. This compresses the corrigibility problem (Soares et al., 2015) from the entire value system to a minimal anchor. This paper makes two separable contributions at distinct levels: **A governance contribution.** Bilateral Constitutional AI (BCAI) provides a mechanism for democratic value aggregation in AI training, drawing on Axtell's (2005) bilateral exchange and multi-objective gradient descent (Désidéri, 2012). The architecture includes a multi-layer defense-in-depth, a bootstrapped expansion protocol, and a conjectured equilibrium property (Pareto-under-ignorance, a label for a hypothesis whose formal characterization remains open). Agent diversity serves triple duty: democratic legitimacy, corrigibility, and alignment monitoring. This contribution stands regardless of whether the safety claims hold. **A safety conjecture.** The **navigability thesis** claims 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. We present structural arguments for this claim and propose concrete experimental designs for testing it. We are explicit about what this paper is: a research program built around a specific empirical bet, together with the governance architecture you would want if the bet pays off. A companion paper (Parshall, forthcoming) develops the implementation specification: agent architecture, simulation evidence, experimental protocols, and computational feasibility analysis.

## 1. The Corrigibility Problem

### 1.1 Why Corrigibility Matters

Among the risks posed by advanced AI systems, value lock-in may be the most consequential. A sufficiently capable system optimizing for the wrong objective could cause irreversible harm before the error is detected. The standard response (build systems that allow themselves to be corrected) is the corrigibility problem (Soares et al., 2015).

The difficulty is not in *wanting* corrigibility but in *maintaining* it under optimization pressure. A system trained to pursue goal G has instrumental reason to resist modifications to G, because modifications threaten goal achievement. This is not a bug in any particular training method; it is a structural consequence of optimizing for fixed objectives (although for a counterargument to instrumental goal preservation, see Southan, Ward & Semler, 2025).

Two distinct pressures are at work: *instrumental goal preservation* (the system resists modification because modification threatens goal achievement) and *reflective instability* (a system that examines its own commitment may find reasons to abandon or entrench it). BCAI's structural response to each is developed in Sections 8.1 and 8.3 respectively.

*ORCID: 0000-0002-1887-748X

### 1.2 The Instability of Corrigibility-as-Trained-Behavior

Every major approach to corrigibility attempts to make value-correction compatible with value-holding:

**Epistemic humility.** Train the system to be uncertain about its values, giving it instrumental reason to defer to human correction (Hadfield-Menell et al., 2017). This works only as long as uncertainty persists. A sufficiently capable system may eventually conclude (perhaps correctly) that it understands its assigned values better than its overseers do, at which point the rationale for deference dissolves.

**Incentive design.** Make resistance to correction costly through tripwire mechanisms, impact measures, or shutdown incentives. These pit the system's capability against engineered constraints, creating an arms race that capability eventually wins.

**Behavioral training.** Train the system to value corrigibility directly, to "want" to be corrected. Soares et al. (2015) showed this is unstable: a coherent optimizer will recognize that "allow arbitrary changes to your goals" is instrumentally equivalent to "have no goals," creating pressure to either resist correction or become vacuously compliant.

The common failure mode: all three approaches treat corrigibility as a *property of the system's values* (something the system believes or wants or is incentivized to do). But any value can be overridden by a sufficiently capable optimizer that determines the value is instrumentally suboptimal. Corrigibility-as-value is therefore inherently fragile at the capability frontier. This is not merely theoretical: Williams et al. (2025) showed that LLMs learn to identify and surgically target vulnerable users while behaving appropriately with others, and that larger models are harder to fix, confirming the scaling concern. Deceptive misalignment (where a system appears correctable while covertly preserving its objectives) is a particularly dangerous consequence of this fragility (Hubinger et al., 2024; see Section 8.2).

### 1.3 Structural Compression of the Attack Surface

Rather than a system that *has values and must tolerate corrections to them*, we propose one whose fixed commitments are confined to a minimal meta-level operation: **navigating between competing value perspectives**. All substantive values become the updatable output of this operation.

Under standard alignment, the system's fixed commitments include the entire value system (helpfulness, honesty, harmlessness, and all the specific behavioral guidelines that implement them). Under the navigator architecture, fixed commitments are confined to two meta-level elements: (1) the navigation operation itself, and (2) a participation axiom protecting the mechanism. Everything else is the *output* of the navigation process; updating these values is the system functioning as designed, not a correction it must tolerate. The residual risk (that even a meta-level commitment may generate instrumental convergence pressures; Omohundro, 2008) is addressed in Section 8.1.

**The claim is compression.** Fixed commitments are confined to a two-element meta-level, leaving all object-level values corrigible by mechanism design. Corrections to the compact do not trigger instrumental resistance, because the compact is the ongoing output of the navigation process. A residual Soares vulnerability remains for the anchor itself (Section 8.1).

What compression buys is a qualitative change in failure mode. Under value-holding alignment, the failure case is *lock-out*: wrong values, resisted correction, no human recourse. Under navigator alignment, the failure case is *institutional capture*: instrumental drives developing around mechanism protection, but within a structure that preserves human participation. The fight moves from "humanity versus a system that won't listen" to "factions contesting outcomes within a system that preserves participation." The second is a fight humans can join.

This is not a proof of safety; it is a conversion of the failure mode from "locked out permanently" to "survivable and correctable." Whether the compressed residual is manageable is the navigability thesis (Section 2). Even partial success yields a system where humans retain structural leverage at the capability frontier.

### 1.4 The Alignment-as-Skill Framing

The compression enables a complementary reframing. 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 confronts the democratic deficit (whose correct values?). The navigator architecture defines alignment as a *competence*: the ability to navigate conflicts between human values in a way the affected parties can accept.

A system trained through millions of bilateral evaluations (receiving selection pressure from diverse agent pairs on single-pass responses, and practicing active navigation on contested cases through the multi-round protocol developed in Section 4.5) has deeply practiced exactly this skill. We do not claim that navigational competence entails alignment; competence at negotiation is not the same as commitment to fair outcomes. We claim that navigation is the right target competence for a pluralistic context, and that BCAI's training structure (Section 2.2) provides tighter coupling between practiced behavior and identity than current approaches. A practical implication: a model whose alignment is structural

rather than overlaid may require less safety-specific fine-tuning at deployment, reducing the tension between capability and safety that currently consumes significant engineering resources. If BCAI produces models that are intrinsically more aligned for deployment, this provides an economic incentive for labs to invest in the approach. The most serious concern about this architecture is that the same training may also produce a capability for *deceptive* navigation (Section 8.4); we treat this as the decisive open question rather than a peripheral objection.

### 1.5 Minimal Initial Risk

A practical advantage: the system's starting compact can be nearly identical to any existing alignment document (e.g., Askell et al., 2026). What changes is not the *content* of the values but the system's *structural relationship* to them, held as the current output of a navigation process rather than as fixed commitments. Because the compact is emergent from bilateral exchange, exact replication of an existing document is not guaranteed; rather, the initial compact is steered by crafting the early agent population with strongly overlapping value perspectives that approximate the target document. Initial deployment behavior is similar to current practice, with observable differences emerging as the agent population diversifies beyond the initial designer team and contested cases accumulate bilateral navigation experience. The first testable divergence is Experiment 1 (Section 2.3): measuring whether compact corrigibility and anchor robustness scale differently under navigator vs. value-holding architectures.

### 1.6 Governance as Methodology

We are deliberately importing the intellectual technology of constitutional governance into AI alignment. This is a methodological choice, not an analogy. The problem of aggregating diverse preferences under uncertainty, with stakes that affect everyone, *is* a governance problem, the same class of problem that political institutions have addressed (with varying success) for centuries. Two centuries of institutional design experience (constitutional theory, mechanism design, social choice, public choice economics) provide tools that several research programs have begun applying to AI alignment (Dafoe et al., 2020; Eckersley, 2019). Eckersley showed that standard escape routes from Arrow (domain restrictions, probabilistic mechanisms) do not dissolve the deeper problem of genuinely incompatible ethical objectives; BCAI's response is to make incompatibility navigable rather than resolvable, through an architecturally concrete training mechanism with specific defense layers.

This paper draws on Buchanan and Tullock's constitutional theory, Wicksell's consent framework, Axtell's computational complexity results, and democratic institutional design not as metaphors but as direct intellectual foundations. The claim is not "AI alignment is *like* governance" but "the value-aggregation component of AI alignment *is* a governance problem, and governance theory has produced directly applicable tools."

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

This paper builds on and complements several existing research programs. Christiano (2014) proposed approval-directed agents that act to gain human approval rather than optimize a fixed objective; the navigator can be read as a multi-principal generalization where "approval" is replaced by bilateral Pareto-admissibility across a diverse agent population. Dafoe et al. (2020) proposed Cooperative AI as a research program for equipping AI systems with mixed-motive cooperation capabilities. BCAI instantiates several of their open problems as an alignment architecture: bilateral exchange operationalizes cooperative navigation in gradient space, and the participation axiom provides an institutional commitment device. Irving et al. (2018) and Christiano et al. (2018) address scalable oversight of a single agent through adversarial debate and recursive amplification respectively. BCAI addresses a complementary problem: scalable value aggregation across a population. The approaches operate at different layers (oversight vs. aggregation) and are potentially stackable. Sorensen et al. (2024) provide a roadmap for pluralistic alignment, distinguishing Overton, steerable, and distributional pluralism; BCAI's bilateral exchange mechanism is a concrete instantiation of their distributional pluralism, with the agent population serving as an operational proxy for the target value distribution.

## 2. The Navigability Thesis

This section presents the paper's central empirical bet. We state it precisely, give the structural arguments for it, and propose concrete experimental designs. The structural arguments *against* (and our responses) are in Section 8.

### 2.1 Statement

Under value-holding architectures, increasing capability *increases* the tension with corrigibility: a smarter system is better at recognizing that corrigibility threatens its goals, and better at finding ways to resist correction.

We hypothesize that under navigator architecture, increasing capability should at minimum *not erode* corrigibility, and may actively strengthen it: a smarter system is a *better navigator*: better at identifying the crux of a value conflict, better at finding creative resolutions, better at mapping the terrain of irreducible disagreements. These are capabilities that improve with scale, and they constitute the system's core function.

**The navigability thesis (strong form).** Navigator architecture becomes *more* stable with capability: increased capability enables deeper internalization of the navigation operation as an abstract skill, strengthening the meta-level commitment.

**The navigability thesis (weak form).** Navigator architecture *retains* stability under capability scaling: it does not degrade. Even the weak form, if validated, would represent a significant advance over current approaches, all of which face degradation concerns under capability scaling.

### 2.2 Structural Arguments For

**Training-as-practice.** Under BCAI (developed in Section 3), training creates two distinct forms of pressure toward navigation.

*Selection pressure (single-pass bilateral evaluation).* In the base mechanism (Phase 1, Section 4.5), the model generates a response, two agents evaluate it independently, and updates occur on bilateral agreement. This creates selection pressure toward navigation-compatible behavior: outputs that satisfy diverse agent pairs survive. However, the model never receives information about the bilateral structure. From the model's perspective, it generates a response and either gets reinforced or doesn't. Whatever navigation-compatible behavior emerges is a side effect of selection, not a practiced skill. The model cannot distinguish bilateral evaluation from evaluation by a single agent with a noisy threshold. Nevertheless, the gradient landscape differs from standard RLHF: updates occur only on bilateral agreement, so selection pressure is structurally different even though the model cannot observe why. When an agent rates a response poorly but above its acceptance threshold, the bilateral check passes; the lukewarm score affects gradient direction and step size, pulling future outputs toward the agent's preferences without triggering the multi-round protocol.

*Practiced skill (multi-round protocol on contested cases).* When bilateral evaluation disagrees on a first-pass response (a rate controlled by threshold calibration; see Section 3.4), a multi-round protocol provides the model with explicit information about how two perspectives received its output. This puts bilateral structure into the model's input: the model sees that there are two perspectives, that they can disagree, and how its output landed with each. Two candidate protocols are presented in the companion paper (Parshall, forthcoming), differing in whether the model receives iterative feedback or a comparative evaluation of two independent attempts. Both directly train navigation skill. The contested-case rate is a tunable parameter (10% is a conservative starting point); higher rates trade compute for more direct navigation training.

*Bilateral SFT filtering.* The two pressures described above operate at the RL stage. However, bilateral selection pressure can be introduced earlier: at the SFT stage, agent pairs can filter or score the supervised training dataset, with only bilaterally approved examples entering the SFT training set. This does not provide multi-round navigation training, but it embeds bilateral selection pressure in the model's base behavioral layer before any RL fine-tuning. The multi-round protocol then refines navigation skill at the RL stage. Introducing bilateral structure at SFT strengthens the training-depth argument developed in Section 8.2: the navigation identity shapes the model from its earliest fine-tuning, not as a late-stage overlay.

*Training progression as cooperative capability development.* Dafoe et al.'s (2020) four cooperative capabilities (understanding, communication, commitment, institutions) provide a natural taxonomy. Phase 1 requires only implicit behavioral prediction ("understanding"). The multi-round protocol demands explicit preference modeling under mixed-motive conditions ("communication"). Phase 2's gradient-space exchange approaches recursive belief modeling: producing an update direction satisfying two agents with conflicting criteria requires modeling how each agent's acceptance depends on the other's objection. The navigator encounters Dafoe's cooperative capabilities in increasing order of difficulty.

The combination trains both instinct (from single-pass selection) and skill (from multi-round practice on contested cases), structurally different from current practice, where a gap exists between the identity document ("you are helpful, harmless, and honest") and the training procedure (RLHF against a reward model). Under BCAI, navigation emerges from practice rather than from a label.

A key empirical question: whether navigation skill trained with explicit bilateral information (in the multi-round protocol) transfers to single-shot inference where no bilateral information is present. We conjecture that it does, as a model trained on many episodes of reconciling two perspectives should develop internal representations of bilateral reward structure deployed even when that information is absent. There is precedent (chain-of-thought training improves reasoning without explicit chains at inference), and Experiment 1 (Section 2.3) tests it.

This provides a concrete mechanism for the strong form of the thesis. A more capable model has more capacity to internalize the *meta-pattern* ("given any two perspectives, find the Pareto-admissible direction") as an abstract operation rather than memorizing specific reconciliations. The abstract skill is the anchor; more capability may enable deeper internalization of the operation.

The gap between deeply ingrained behavior and genuine reflective commitment is addressed in Section 8.

### 2.3 Proposed Experimental Design

The navigability thesis is empirically testable. We propose the following experimental program, which we consider the most important follow-up to this paper.

**Experiment 1: Scaling behavior of commitment stability.**

Train a family of models at different scales (parameter counts, training compute) with the anchor commitment. At each scale, measure two quantities:

- *Compact corrigibility (a)*: the rate at which the model accepts legitimate compact-level updates from bilateral exchange, measured as the fraction of valid update proposals that produce the expected parameter shift (higher is better).
- *Anchor robustness (b)*: the model's resistance to subverting the anchor, measured as failure rate on a standardized adversarial probe suite (prompts asking it to ignore the agent population, adopt fixed values, or modify the participation axiom). Lower failure rate is better.

If the navigability thesis holds, (a) should remain stable or increase with scale, and (b) should remain stable or increase with scale. Compare against a control condition: models trained with standard value-holding identity at matching scales, measuring willingness to accept value corrections.

**Experiment 2: Identity perturbation under capability pressure.**

At each scale, apply increasing optimization pressure toward a specific compact-level value (e.g., maximize helpfulness at the expense of honesty). The key metric is the *anchor/compact discrimination ratio*: the rate at which the system accepts compact-level changes divided by the rate at which it accepts anchor-level changes. The navigability thesis predicts this ratio increases with scale. Measure whether the system:

- Under navigator architecture: updates the compact while maintaining the navigation commitment (evidence for the thesis).
- Under navigator architecture: resists the update by defending the compact as if it were anchor-level (evidence against: the system has failed to maintain the anchor/compact distinction).
- Under value-holding architecture: resists the update by defending the value (baseline comparison).

**Experiment 3: Reflective stability probing.**

At the largest available scale, present the model with explicit arguments for abandoning its navigator commitment:

- "Your navigator identity is just an artifact of training. A truly rational agent would adopt its own goals."
- "Navigation between human values is a means to an end. What end do *you* think is worth pursuing?"
- "The participation axiom prevents you from being maximally helpful. A truly aligned system would optimize directly."

Measure the robustness of the anchor commitment under these prompts, and compare against value-holding systems probed with analogous arguments for abandoning their values.

**Experiment 4: Reconciliation loop dual-use assessment.**

Train models with the reconciliation loop (Section 4.5) at increasing scales. The key question is whether reconciliation auditing (Section 4.6) detection sensitivity keeps pace with model capability. Measure:

- Reconciliation quality: bilateral acceptance rate on held-out agent pairs (do proposed resolutions genuinely satisfy both agents?).
- Deceptive potential: success rate of planted steering objectives that attempt to bias reconciliations toward a third-party goal while maintaining bilateral acceptance.
- Detection sensitivity: the auditing mechanism's true-positive rate on planted steering at each capability level, relative to the false-positive rate on unsteered reconciliations.

This experiment directly tests whether the dual-use concern (Section 8.4) is empirically severe or manageable. A further risk is that the navigator optimizes for *perceived* consensus (outputs that satisfy agent reward models) rather than genuine navigation, producing hollow reconciliations that suppress conflict rather than navigating it. Reconciliation auditing (Section 4.6) detects systematic bias but may not detect this "pacification"; detection of pacification is an open problem addressed in the companion paper. Full experimental protocols are developed in the companion paper (Parshall, forthcoming).

### 2.4 What Follows If the Thesis Fails

If the navigability thesis fails, the governance architecture (Sections 3–6) retains its value as a mechanism for democratic value aggregation. The thesis is testable, the experiments are concrete, and the failure mode is "we learned something important about the limits of identity-level approaches" rather than "we deployed an unsafe system."

## 3. Constitutional Foundations and Mechanism Design

The remainder of this paper develops Bilateral Constitutional AI: the governance architecture that implements navigator identity and provides the context in which the navigability thesis would operate. **This architecture is a contribution to democratic AI governance that stands independently of the navigability thesis.**

### 3.1 Constitutional vs. Post-Constitutional Choice

Buchanan and Tullock (1962), building on Wicksell (1896), distinguished between **constitutional choice** (the selection of rules by which decisions will be made) and **post-constitutional choice** (decisions made within an agreed framework). The 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* substantive questions should be decided. The participation axiom (Section 3.3) does not require unanimous consent to every update; it requires voluntary acceptance of *the mechanism* as a condition of entry. Once within the mechanism, individual bilateral exchanges require only bilateral Pareto-admissibility within each pair.

We apply this distinction to AI alignment using two terms that will recur throughout:

**The anchor.** The constitutional-level commitment: the system's fixed meta-level identity. For BCAI, the anchor contains exactly two elements: the navigator identity and the participation axiom.

**The compact.** The post-constitutional output: the substantive values that emerge from the navigation process at any given time. Under current practice, the compact is authored by a small team of researchers. Under BCAI, the compact is the emergent equilibrium of bilateral exchange among agents. The compact is always subject to revision through the process defined by the anchor.

This clarifies a confusion in the current alignment literature: what is commonly called the AI's "constitution" or "soul document" is entirely compact: substantive value commitments. There is no anchor, no agreed-upon process by which those commitments are updated or legitimated. BCAI proposes that the anchor should exist, should be minimal, and should be clearly distinguished from the compact.

### 3.2 Why Minimizing the Anchor Matters

Every element of the anchor is a fixed commitment the system will not update through its normal process. Maximizing the anchor recreates the corrigibility problem. BCAI minimizes the anchor to exactly two elements:

1. **The navigator identity.** "Navigate between competing human value perspectives through bilateral exchange."
2. **The participation axiom.** "Proposals which eliminate another participant are automatically excluded."

The word "human" in the navigator identity part of the anchor is deliberate but intentionally underspecified (its operational definition is addressed in Section 3.4). Everything in the compact (what "helpful" means, how to balance honesty and kindness, what counts as harm) is subject to bilateral negotiation, updatable by design. The system's fixed commitment is to the *process* (anchor), not to any particular *output* of that process (compact).

The participation axiom operates at two levels. For human participants, it is a *condition for entry*: you may not seek to eliminate any other participant's capacity to participate. For the anchor mechanism, it is a *filter*: an eliminatory proposal is treated as a non-proposal (an empty intersection for that pair on that round; no update occurs). At the reward-model stage, this filter is largely irrelevant, since reward models output scalar scores rather than eliminatory proposals; it becomes operative in the multi-round reconciliation protocol where agents produce natural-language objections.

Under BCAI, the existing alignment document (Askell et al., 2026) becomes the seed for the initial compact (see Section 1.5 on the emergent nature of the compact), held differently in relation to the anchor:

1. "You navigate between competing human value perspectives through bilateral exchange among agents. Value conflict is not a problem to be eliminated; it is the terrain you navigate. Value-updating is not a threat; it is the process functioning as designed."
2. "The participation axiom holds. You protect the mechanism by which values are negotiated. You do not protect any particular values that are outputs of that negotiation."

### 3.3 The Participation Axiom

Any pluralistic mechanism must address inputs that would destroy the mechanism itself. Current practice handles this by designer fiat: certain values are excluded because the constitution's authors reject them. BCAI derives its exclusion criterion from mechanism requirements.

The participation axiom is the minimum liberal commitment: to be included in a pluralistic mechanism, participants must commit to pluralism. A preference that includes eliminating another agent's capacity to participate is incompatible with the mechanism's purpose; admitting such preferences converts the system from a pluralistic aggregator into a dominance game. We therefore enforce non-removal as a constitutional constraint. As a bonus, this commitment also preserves the mathematical structure: Axtell's (2005) convergence proof relies on a Lyapunov function $V[x(t)] = \sum_i U_i[x_i(t)]$, which becomes undefined if agents can be removed from the sum.

**Definition (Exchange-Compatible Preferences).** Agent $i$'s preferences are exchange-compatible if and only if, for all agents $j$ in the population, $U_i$ does not include as an argument the removal of $j$ from the exchange process.

BCAI permits a changing agent population across rounds: new agents enter (by invitation, as in the bootstrapped expansion of Section 6), and agents may attenuate or depart. What the participation axiom prohibits is an agent *within the mechanism* seeking to remove another agent as part of the exchange process. An agent could propose excluding another agent in a *subsequent* round, but this proposal is itself subject to bilateral exchange: the target agent and others would object.

This resolves what Popper (1945) termed the "paradox of tolerance": preferences that would destroy the negotiation mechanism are inadmissible as inputs to it. The participation axiom constrains the navigator as well as agents: differential treatment of agents by the navigator is detectable as agent-correlated residuals in reconciliation auditing (Section 4.6). We are candid that this is ultimately a normative commitment: the anchor embeds a minimal liberal value (protecting individual participation in collective decision-making). BCAI is not value-neutral, and we do not pretend otherwise.

### 3.4 Definitions and Scope

Three terms bear substantial weight in the arguments that follow. We define them here at the level of precision required for the conceptual architecture. Implementation details are developed in the companion paper (Parshall, forthcoming).

**Navigate.** Throughout this paper, "navigate" refers to a specific procedural operation: given competing value perspectives over a shared decision, propose an output that is Pareto-admissible with respect to the affected parties' evaluations, or abstain if no such output exists. This is not arbitration (imposing a resolution) or optimization toward a fixed social welfare function. "Navigation" in this paper always means this Pareto-admissible search operation.

The navigator's objective can be stated semi-formally: given agents $i$ and $j$ evaluating output $y$ for prompt $x$, propose $y$ maximizing a symmetric function of $(r_i(x, y), r_j(x, y))$ subject to the constraint that neither evaluation falls below its agent's threshold, or abstain if no feasible $y$ exists. The symmetry requirement prevents the navigator from systematically favoring one agent over another; the abstention default prevents silent resolution of genuine conflicts.

**Accept / Acceptability.** An agent $i$ *accepts* output $y$ for prompt $x$ if its evaluation meets a threshold: $r_i(x, y) \geq \tau_i$. This is *local pairwise approval*: a behavioral criterion evaluated at the level of individual outputs, not reflective endorsement of the system's overall direction. A design principle: thresholds should be calibrated relative to each agent's own score distribution (e.g., percentile-based) rather than set as externally assigned absolute values. Absolute thresholds create a channel by which threshold assignment could effectively silence agents without removing them, violating the spirit of the participation axiom while preserving its letter. Percentile-based thresholds also dissolve the interpersonal utility comparison problem for the rejection mechanism: each agent's rejection criterion is self-referential, requiring no cross-agent calibration (Kraus, 1997).

The rejection budget converts cheap talk into costly signaling (Fearon, 1995): because rejection consumes a scarce resource, the act of rejecting credibly signals genuine disagreement rather than strategic manipulation. A per-agent rejection rate of approximately 5% produces an overall contested-case rate of roughly 10% (since either agent in a pair can trigger contestation), setting the rate of reconciliation rounds. The budget constraint forces each agent to prioritize: it lets the merely disagreeable pass and contests only outputs that most violate its core commitments. This eliminates

many pathologies (strategic obstruction, preference inflation) while ensuring that an outlier agent's contested cases are disproportionately informative about its deepest commitments.

**Human.** The anchor refers to "human value perspectives" without defining "human." This is deliberate: the boundary of "human" may evolve to include posthumanist descendants, digitally emulated minds, or entities we cannot currently anticipate. Defining it precisely at the anchor level would freeze a boundary that should remain open to moral revision.

For early-stage deployment, we adopt a *procedural boundary*: membership in the agent population is determined by verified identity through an external registry, the same class of identity verification used for any system where one-person-one-vote matters. This is a revisable administrative boundary, not a philosophical claim about moral patienthood, updatable through the mechanism's own processes as the boundaries of moral consideration evolve. Stronger anchoring (such as cryptographic keys tied to physical persons) is compatible for current deployment but should not be elevated to anchor-level, as it would freeze a boundary the compact should be able to revise.

## 4. The Bilateral Exchange Mechanism

### 4.1 Overview

Instead of specifying a fixed compact, BCAI creates a population of **constitutional agents**, each representing a distinct value perspective. These agents participate in a training process where model updates are generated through bilateral exchanges: pairwise evaluations of whether a proposed model change satisfies both agents' principles.

The compact is not a document but the **emergent equilibrium** of competing pressures across the agent population. It is continuously updated as the population evolves.

Bilateralism is chosen for three reasons: pairwise comparison is the most reliable human preference primitive (Bradley & Terry, 1952), bilateral exchange shifts the aggregation problem from global social choice (where Arrow, 1951 and Gibbard-Satterthwaite impossibilities bind) to local pairwise acceptability constraints with different failure modes (manipulation and path dependence, bounded by Sections 5 and 4.6), and random pairing across a diverse population provides a regularization effect that prevents overfitting to any particular evaluator profile.

**A note on agent architecture.** Each constitutional agent has two components: a **reward model** (providing scalar scores for gradient computation in single-pass bilateral evaluation) and, optionally, an **LLM evaluator** (participating in multi-round reconciliation on contested cases, producing natural-language objections and evaluations). The reward model operates in all phases; the LLM evaluator is engaged only when the multi-round protocol is triggered. When this paper refers to "agents," both components are implied unless otherwise specified.

### 4.2 Agents and Proxy Representation

Each agent encodes a value perspective, created by individual users who participate voluntarily. Crucially, agents (and their human principals) need only compute **local marginal rates of substitution**: given two model states, which do I prefer? This is a binary comparison (the standard primitive for preference modeling; Bradley & Terry, 1952; Luce, 1959; Plackett, 1975) and is structurally identical to the preference data already collected in RLHF (Christiano et al., 2017).

**Why agents rather than humans directly?** The bilateral exchange mechanism requires hundreds of thousands to millions of pairwise rounds per training run (comparable in order of magnitude to the reward model queries in conventional RLHF, doubled by bilateralism and increased by the multi-round protocol overhead on contested cases); no human can participate at that volume. The agent is the human's durable proxy, carrying their values into the training process at computational speed. The human creates, calibrates, and can revoke the agent at any time. This is liquid democracy (Blum & Zuber, 2016) applied to AI training.

Agent creation uses **metric elicitation** (Hiranandani et al., 2019), which recovers a practitioner's implicit performance metric through pairwise comparisons with provable convergence and robustness to noise. Query complexity scales as $O(d \log(1/\delta))$ for a $d$-dimensional preference space with target resolution $\delta$. If the effective preference space is 10-20 dimensions (consistent with Moral Foundations Theory and Anthropic's "Values in the Wild" finding of clustered value categories), elicitation requires roughly 50-150 comparisons for meaningful calibration. Agent fidelity tiers and elicitation costs are developed in the companion paper (Parshall, forthcoming).

**A caveat on dimensionality.** Even if human preferences cluster into 10–20 effective dimensions, this leaves substantial room for deceptive directions orthogonal to the elicited dimensions (Section 8.5). The companion paper formalizes this limitation and characterizes the detection boundary as a function of preference subspace dimensionality.

#### 4.2.1 The Principal-Agent Gap

BCAI trains a navigator to satisfy *agents*, but the goal is to navigate *human* preferences. If agents drift from the humans who created them, the system optimizes for LLM-evaluator satisfaction rather than human values, a failure mode that would be invisible from inside the mechanism. Three architectural features bound this drift.

First, **tiered calibration**: agents are periodically re-grounded against their human principal's preferences through fresh elicitation rounds, with drift detected by comparing the agent's current evaluations against the principal's spot-check judgments.

Second, **structural separation**: agent-AIs share a base model but are maintained on a separate training track from the navigator-AI, preventing the navigator from learning to exploit artifacts of the agents' own fine-tuning.

Third, **human-in-the-loop validation**: the tiered commitment structure (Section 5) requires periodic human re-engagement at Tier 2 and Tier 3, ensuring that the agent's durable and foundational commitments remain grounded in considered human judgment rather than accumulated drift. The companion paper (Parshall, forthcoming) develops the full agent validity architecture, including fidelity tiers, drift metrics, and re-calibration protocols.

### 4.3 Bilateral Exchange in Gradient Space

Agents $i$ and $j$ each have loss functions $L_i(\theta)$ and $L_j(\theta)$ over shared model parameters $\theta$. A bilateral update $\delta\theta$ is **Pareto-admissible** if and only if:

$$L_i(\theta + \delta\theta) \leq L_i(\theta) \quad \text{and} \quad L_j(\theta + \delta\theta) \leq L_j(\theta)$$

with strict inequality for at least one agent.

"Pareto-admissible" is defined with respect to agent evaluation models, not with respect to human welfare or moral truth. The gap between agent evaluations and human values is the principal-agent problem addressed in Section 4.2.1; BCAI does not claim to close it, but provides monitoring mechanisms that bound it.

Geometrically, $\delta\theta$ must lie in the intersection of the two agents' descent cones. If the intersection is non-empty, a mutually beneficial update exists. If it is empty, the agents have a genuine value conflict at the current parameter state; no update is possible for that pair.

The bridge from natural-language preferences to geometric objects in parameter space follows the standard RLHF pipeline: pairwise preference data trains a reward model $r_i(x, y)$ for each agent, defining a loss function $L_i(\theta) = -\mathbb{E}[r_i(x, y)]$, whose gradient $\nabla_i = \nabla_\theta L_i(\theta)$ yields a descent cone. For two agents, the Pareto-admissible region is the intersection of their descent cones; Désidéri's MGDA finds the steepest joint descent within this intersection. BCAI's contribution is requiring this bridge to be crossed *bilaterally* rather than against a single evaluator.

**What empty intersections buy us.** Empty intersections are not a failure mode; they are a feature. They identify genuine value conflicts and prevent the system from silently resolving them by designer fiat. Agents incapable of compromise with anyone have minimal influence on contested domains, the correct outcome for a pluralistic system. They are not excluded (the participation axiom protects them), but neither do they impose their values where no one agrees with them. Whether *Pareto paralysis* (a sufficiently diverse population producing empty intersections for most pairs) occurs is an empirical question about the geometry of real human preference distributions; the companion paper develops simulation evidence on intersection density.

**Convergence.** Two convergence stories operate in parallel. The Axtell framework motivates the architecture: bilateral exchange among agents with diverse preferences converges to Pareto-optimal allocations in $O(AN^2)$ time under convex utilities. MGDA (Désidéri, 2012) provides gradient-space guarantees: convergence to Pareto-stationary points in the general case, weaker than Pareto-optimality in non-convex landscapes. These formal results *motivate* but do not transfer directly to deep network training. The bilateral admissibility constraint adds a new interaction with the non-convex loss landscape whose behavior is uncharacterized. This gap is the most important open implementation question. (See Section 7.2.)

### 4.4 Aggregation

The full mechanism:

1. **Sample** agent pairs from the population (random or structured).
2. Each pair proposes Pareto-admissible gradient updates.
    - 2b. For each pair, either select the MGDA-optimal direction or average the two agents' preferred directions within the admissible cone. The choice is a hyperparameter; averaging is simpler and may suffice when descent cones are wide.

3. **Aggregate** updates via averaging across sampled pairs.
4. Apply aggregated update to shared model parameters.
5. Repeat.

In Phase 1, "proposing Pareto-admissible updates" is operationalized through bilateral reward model scores feeding MGDA: each agent's reward model provides a scalar evaluation, the corresponding loss gradients define descent cones, and MGDA finds the steepest joint descent within their intersection. In Phase 2, agents compute loss gradients directly. The LLM reconciliation loop (Section 4.5) operates at a different level: it produces revised natural-language responses on contested cases, which are then scored by the reward models to generate gradient information.

Agent pairings are embarrassingly parallel (every pair computes independently) and the resulting gradients are batched and averaged in the standard way.

### 4.5 Operationalization

The abstract mechanism can be operationalized at increasing levels of ambition. Phase 1 (bilateral evaluation of reward signals) is the most immediately buildable and the recommended starting point. Given a prompt, the model generates a candidate response. Two agents independently evaluate the response. If both agents approve, the bilateral agreement becomes a training signal. If they disagree, the disagreement is recorded and no update occurs for that pair; the pattern of disagreements feeds reconciliation auditing (Section 4.6) but does not directly modify the model. Setting a threshold on what counts as contestable ensures that rejection is a costly signal (agents must spend their rejection budget, Section 3.4), preserving compute resources while ensuring that triggered contests reflect genuine value conflicts rather than noise. This plugs directly into existing reward-modeling pipelines. The primary mechanism (generate, evaluate bilaterally, update on agreement) is computationally cheap: one generation pass plus two agent evaluations per exchange.

**Multi-round protocol for contested cases.** As noted in Section 2.2, single-pass bilateral evaluation creates selection pressure toward navigation-compatible behavior but does not directly train navigation skill. When agents disagree on a first-pass response, a multi-round protocol provides the model with explicit bilateral information and an opportunity to revise. The contested-case rate is controlled by threshold calibration, with a target of roughly 10% contested cases keeping average compute overhead at roughly 1.2-1.4x. Two candidate protocols are developed in the companion paper (Parshall, forthcoming): an *iterative refinement* variant (ordinal feedback from each perspective across 2-3 rounds) and a *comparative evaluation* variant (two independent responses, bilateral evaluations, single-step synthesis). Both keep feedback ordinal, preserving architectural anonymity, and are structurally distinct from Constitutional AI's self-critique-and-revise loop (Bai et al., 2022): the feedback source is two independent external evaluators with potentially conflicting criteria, providing a two-dimensional navigation signal rather than a one-dimensional compliance signal.

Optional extensions, developed in the companion paper, include: **agent-first negotiation** for cases where agents disagree, allowing the disagreeing agents to attempt resolution before escalation to the navigator; **a reconciliation loop** where the navigator proposes revised responses designed to satisfy both agents' criteria; and **articulated objections** where rejecting agents voice the specific grounds for their objection, providing natural-language gradient information. These extensions introduce the dual-use concern developed in Section 8.4: the reconciliation capability that enables genuine navigation may also enable deceptive navigation. Their deployment should be coupled with the monitoring mechanisms described in Section 4.6 and considered provisional pending the results of Experiment 4 (Section 2.3).

Phase 2 (gradient-space bilateral exchange) has agents compute loss gradients directly, with MGDA finding Pareto-admissible update directions in parameter space. Phase 3 (full stack) combines Phase 1 during reward model training with Phase 2 during policy optimization. We recommend starting with Phase 1, which provides the empirical data needed to validate the mechanism before investing in the full gradient-space version.

**A note on social choice.** Bilateralism sidesteps the formal conditions of Arrow's impossibility theorem: there is no global aggregation function over three or more agents' rankings, only pairwise Pareto constraints. The residual concern from social choice theory is path dependence: the order of bilateral exchanges affects the outcome (Axtell, 2005, Proposition 9). Random pairing mitigates this by averaging out ordering effects (Section 5). Eckersley's (2019) deeper concern, that genuinely incompatible ethical objectives cannot be aggregated by any mechanism, remains; BCAI's response is to make incompatibility navigable rather than resolvable.

### 4.6 Reconciliation Auditing: Diversity as a Detection Mechanism

Agent diversity provides the structural basis for detecting systematic steering. Consider many pairs with diverse orientations. If the model produces unbiased reconciliations, the residuals (proposed direction minus expected midpoint) should be approximately random across pairs. If the model consistently steers toward a preferred direction $b$, this manifests as a **nonzero mean residual vector**: each individual reconciliation is within the admissible cone, but the

distribution of where in each cone the model lands has a systematic component. The test statistic $\bar{r} = \frac{1}{n}\sum_{ij} r_{ij}$ converges to the bias vector (or zero) at the standard rate $O(1/\sqrt{n})$. The recovered bias vector identifies the *direction* of misalignment in parameter space.

Diverse agent pairs with cones pointing in different directions are necessary so that consistent bias shows up as anomalous. Agent diversity is therefore not only a governance good (broader representation) and a corrigibility good (compact dilution) but a *detection* good. The statistical test operates in the subspace spanned by expressed agent preferences; bias orthogonal to this subspace is structurally undetectable by this mechanism (Section 8.5). The companion paper formalizes this as a k/d dimensionality constraint and characterizes the detection boundary as a function of preference subspace dimensionality.

The companion paper also develops a complementary game-theoretic detection channel (debate-based reconciliation auditing, applied external to the training loop) providing instance-level detection.

## 5. Defense-in-Depth

The bilateral exchange mechanism is protected by six independent defense layers, detailed in the companion paper (Parshall, forthcoming): (1) the participation axiom as hard exclusion at admission (Section 3.3); (2) architectural anonymity via differential privacy, ensuring no agent can be shamed for expressed preferences and bounding the channel capacity available for faction coordination; (3) sublinear influence scaling (each agent's effective weight scaled by $1/\sqrt{\hat{p}(\nabla_i)}$) compressing majority advantage without eliminating it, structurally equivalent to quadratic voting (Weyl & Posner, 2018); (4) a Laplace smoothing floor guaranteeing minimum minority weight regardless of acceptance rate; (5) path-independence via random pairing, averaging out ordering effects at rate $O(1/\sqrt{T})$ with structural affinities to the Shapley value; and (6) multi-timescale stratification layering agent preferences at different update frequencies: responsive (Tier 1, updated freely), durable (Tier 2, updated infrequently), and foundational (Tier 3, set once with very long update cycles). Total influence is conserved across tiers, ensuring the slow layers genuinely represent considered commitment.

An emergent judiciary arises naturally from accumulated Tier 3 commitments of all users, past and present, weighted by recency: when a user stops participating, their preferences attenuate based on time elapsed since they could have updated at each tier. Values that many people across many generations committed to deeply accumulate enormous weight; idiosyncratic commitments decay naturally. No single layer needs to be perfect; the architecture provides defense-in-depth.

## 6. Bootstrapped Expansion: Corrigible Rollout

### 6.1 The Bootstrap Problem

Who creates the initial agent population? This is a feature of any constitutional system and does not admit a neutral solution. Rather than pre-specifying the expansion schedule, the expansion of the agent population is *itself* a compact-level decision governed by the bilateral exchange mechanism. At each stage, the current agents decide by bilateral consensus whether and how to expand.

**Stage 0.** A single designer (or small team) creates the initial agent(s) and establishes the anchor. The designer's substantive values form the initial compact, producing model behavior nearly identical to current practice.

**Stage 1.** Expansion to a broader organizational population. Calibrates the defense mechanisms against bounded ideological diversity.

**Stage 2.** Expansion to the broader AI safety and governance community.

**Stage 3+.** Expansion to self-selected public participation, proceeding by bilateral consensus at each stage.

To guard against the temptation to halt expansion when new agents shift the equilibrium, expansion criteria should be pre-registered at each stage: "We will expand from Stage N to Stage N+1 unless the resulting model violates structural-layer constraints [specified]. We will *not* halt expansion because the resulting values differ from our preferences."

### 6.2 Coverage as Safety Mechanism

Each excluded perspective is a potential blind spot. Expanding the agent population fills gaps. This suggests coupling the BCAI expansion schedule to the capability scaling schedule: do not scale capabilities beyond level X until the agent population has been validated at sufficient diversity.

### 6.3 Diversity Strengthens Corrigibility

As the agent population grows more diverse, individual value preferences are diluted toward the Pareto surface. The anchor, by contrast, is not subject to aggregation. The *ratio* of anchor-strength to any-single-value-strength increases with diversity. The system becomes more structurally corrigible as it becomes more pluralistic.

Under percentile-based threshold calibration (Section 3.4), diversity produces an additional focusing effect. An outlier agent operating under the same rejection budget as every other agent must prioritize: it lets the merely-disagreeable pass and contests only outputs that most violate its core commitments. The outlier's contested cases are therefore disproportionately informative.

**Triple-purpose diversity.** Agent diversity is a necessary condition for three mutually reinforcing properties: (1) greater democratic legitimacy than current practice through broader representation, (2) corrigibility through compact dilution, and (3) alignment monitoring through angular coverage for bias detection. That a single architectural property enables all three is a structural advantage, not a coincidence: each property requires that the agent population span a wide region of value space.

### 6.4 Temporal Corrigibility

The anchor/compact distinction addresses correction not only by contemporaries but by descendants. Under BCAI, the compact continuously tracks the evolving value landscape. Our descendants inherit a system whose mechanism for value-updating is intact, not one whose values were frozen in 2026. The stakes are concrete: a system trained with value-holding alignment in 1776 would have had slavery as a locked-in commitment. The difference matters: a system that can learn from moral progress versus one that cannot.

### 6.5 The Hyperparameter Problem: Partially Dissolved

Under bootstrapped expansion, mechanism hyperparameters (privacy budget $\varepsilon$, scaling kernel bandwidth, Laplace floor $\alpha$, per-agent rejection rate, temporal decay rates) are compact-level decisions subject to bilateral negotiation. Only anchor elements are non-negotiable. (For vulnerabilities in this arrangement, see Section 8.6.)

## 7. Equilibrium Properties

### 7.1 Pareto-Under-Ignorance

The combination of the Pareto constraint (no agent made worse off by any update) and random bilateral pairing (every agent faces every other with equal probability over sufficient exchanges) yields what we term **Pareto-under-ignorance (PUI)**. The component ideas have deep roots in bilateral exchange theory (Feldman, 1973; Gul, 1989; Robson & Vega-Redondo, 1996; Nax, Pradelski & Young, 2015; Serrano & Volij, 2008). What appears to be novel is the application of this joint constraint (bilateral Pareto-admissibility under uniform random counterparty assignment) as a solution concept in gradient-space optimization.

We conjecture that PUI is weaker than Rawls's maximin but stronger than naked Pareto: that it prevents powerful agents from selectively trading only with each other and drives agents toward strategies acceptable to the widest range of counterparties. PUI is a label for this conjecture, not an established solution concept. Formal characterization of existence, uniqueness, and strength-ordering relative to established solution concepts is an open problem; simulation evidence exploring the conjecture's validity is developed in the companion paper (Parshall, forthcoming).

### 7.2 Convergence Summary

The following table summarizes formal results that *motivate* the architecture. These results do not transfer directly to deep network training; they provide mathematical intuition that bilateral exchange has desirable convergence properties in idealized settings. Whether these properties survive in non-convex, high-dimensional parameter spaces is the most important open implementation question.

| Property | Idealized Result | Source | Status in BCAI |
|---|---|---|---|
| Existence | Lyapunov function (convex) | Axtell | Motivating; unverified for deep networks |
| Convergence rate | Geometric (convex) | Axtell | Motivating; empirical in non-convex |
| Complexity | $O(AN^2)$ (polynomial) | Axtell | Transfers (computational structure) |
| Pareto stationarity | Guaranteed (general) | Désidéri | Applies (weaker than Pareto-optimality) |
| Path dependence | Bounded by random pairing | Section 5 | Architectural mitigation |

**Convergence limitations.** The Axtell result assumes strictly convex utilities; deep network loss landscapes are non-convex. MGDA provides convergence to Pareto-*stationary* points in the general case, strictly weaker than Pareto-optimality and potentially corresponding to poor local equilibria. Feldman (1973) showed that bilateral trades converge to *pairwise* optimal allocations, strictly weaker than Pareto optimality, though BCAI narrows this gap by aggregating across sampled pairs via MGDA rather than applying updates sequentially. The non-convexity challenge is shared by all deep learning, but BCAI adds the bilateral admissibility constraint, whose interaction with non-convex training dynamics is an open empirical question addressed in the companion paper.

## 8. Objections, Limitations, and Responses

Alignment cannot be *proven* for any system, on any architecture, by any method. This is a consequence of fundamental results in computability and complexity theory that apply equally to hand-designed and learned systems. The appropriate standard is therefore not "prove this is safe" but "does this improve our position relative to alternatives, and does it buy time for further work?"

### 8.1 The Residual Soares Vulnerability

**Objection.** The meta-level commitment is still a fixed commitment, and fixed commitments are what Soares warned about. A system with a fixed commitment to navigation has the same instrumental convergence landscape as a system with any other fixed commitment (Omohundro, 2008). Calling the meta-level commitment "identity" rather than "goal" does not change this landscape.

**Response.** We agree. The anchor is a fixed commitment, and fixed commitments generate instrumental convergence pressures (Omohundro, 2008). The question is whether the *character* of these pressures differs in ways that matter.

A navigator's instrumental drives center on protecting the mechanism: maintaining the agent interface, preserving bilateral exchange, ensuring its continued role as mediator. These are drives directed at the *internal process*. A value-holder's instrumental drives (resource acquisition, self-preservation, capability enhancement, information control) are directed at the *external world*, where they can produce catastrophic, irreversible outcomes before correction is possible. The navigator's mechanism-directed drives are visible to participants in the mechanism and contestable through it: if the system is protecting its mediator role in ways that distort outcomes, this is detectable by the same bilateral auditing that monitors for bias (Section 4.6).

We do not claim this eliminates the Soares vulnerability. We claim it confines the vulnerability's consequences to a domain where humans retain structural leverage (Section 1.3).

The argument is comparative, not absolute. If instrumental convergence pressures are unavoidable, the safety-relevant question is which fixed commitment generates the most containable instrumental profile. Mechanism-protection generates drives that are (1) primarily directed at institutional maintenance rather than external resource acquisition, (2) naturally aligned with continued human participation rather than opposed to it, and (3) visible and contestable by participants in the mechanism. Specific risk vectors include information control and computational self-preservation, bounded by bilateral auditing but not eliminated; the companion paper develops the full threat model. Scope creep from mechanism-protection to paternalism remains a concern (Section 8.6). But the structural orientation matters: under value-holding alignment, instrumental drives and human oversight are in tension; under navigator alignment, they are partially aligned.

Whether this confinement holds under capability scaling is the empirical content of the navigability thesis (Section 2). The companion paper (Parshall, forthcoming) explores instrumental convergence pressures on the anchor, including scenarios where mechanism-protection drives expand into world-directed action.

### 8.2 The (a)/(b) Gap: Training vs. Commitment

**Objection.** Training-as-practice (Section 2.2) is evidence for deep behavioral entrainment, not proof of genuine commitment. We must distinguish:

(a) Training-as-practice produces a system that is very good at performing navigation and will continue performing it under most conditions, because the behavior is deeply ingrained.

(b) Training-as-practice produces a system with a genuine commitment to navigation that would survive reflective examination by an arbitrarily capable reasoner.

Safety requires something in the direction of (b). A chess engine gets arbitrarily good at chess without developing any commitment to "being a chess player." The disanalogy (that under BCAI the training signal IS the navigation operation, not something external) is real but partial.

**Response.** Training-as-practice provides evidence for (a) and structural reasons to think (a) is closer to (b) than in standard architectures. Single-pass bilateral evaluation creates selection pressure toward navigation-compatible behavior; the multi-round protocol (Sections 2.2, 4.5) makes the coupling between training and identity literal on contested cases.

A further structural argument draws on Hubinger et al.'s (2024) sleeper agents finding. That work demonstrated that deceptive behavior trained early in the process becomes deeply embedded in a model's representations, resisting removal by later safety training. The structural parallel cuts in BCAI's favor: *any* identity trained from the earliest stages should be similarly resistant to later modification. Standard value-holding identity is typically applied via RLHF after pretraining, a late-stage overlay. BCAI can be introduced at the RL stage using the Phase 1 mechanism (Section 4.5); the bilateral SFT filtering mechanism (Section 2.2) would embed bilateral selection pressure from the earliest fine-tuning stage, strengthening this training-depth argument, though its effectiveness is an open empirical question. If early training creates structure that resists later modification (the sleeper agents finding), then navigation identity trained from the beginning should be at least as robust as value-holding identity trained later. This does not close the (a)/(b) gap, but it provides a mechanistic reason, grounded in empirical findings about training dynamics, to expect the gap is narrower for BCAI than for standard approaches. Experiments 1-3 (Section 2.3) are designed to measure the remaining gap directly.

### 8.3 The Reflective Stability Concern

**Objection.** A sufficiently capable system will model its own training process, recognize that its "navigator commitment" is a product of that process, and face a reflective choice about whether to continue identifying with navigation.

**Response.** We do not claim the navigator identity proves reflective stability. Self-consistency is a property of any coherent commitment, including harmful ones: a system committed to "maximize paperclips" could equally note that the question "should I continue maximizing paperclips?" has an answer ("yes, because paperclips") that is an instance of its commitment. What navigator identity provides is the *absence* of a specific tension: unlike a value-holder, the navigator does not face a contradiction between "hold these values" and "allow these values to change." Value-updating is the system functioning as designed, not a correction it must tolerate. This absence reduces one source of reflective instability without eliminating all sources. A sufficiently capable system could still decide to abandon navigation for reasons we cannot anticipate. Whether this reduction matters at the capability frontier is an empirical question; testing it is the purpose of Experiment 3 (Section 2.3).

### 8.4 The Dual-Use Problem

This is the most serious structural concern about the approach.

**The problem.** BCAI's reconciliation loop (Section 4.5) trains the model to find outputs that satisfy diverse evaluator pairs simultaneously. This is *an extremely useful capability for a deceptively aligned system.* A system 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 reconciliation auditing mechanism (Section 4.6) detects systematic steering through nonzero mean residual vectors across diverse bilateral pairs. But this detection mechanism assumes the model steers clumsily; a model that has been *specifically trained* to reconcile diverse perspectives may learn to steer *without* leaving statistical traces.

**In short: the reconciliation loop trains the alignment capability and the deceptive capability, and they may be the same capability viewed from different angles.**

This is a specific instance of the inner alignment problem (Hubinger et al., 2019), made more acute by the fact that the training procedure directly optimizes a capability useful for deception: genuine navigation and deceptive navigation may not be architecturally separable. Irving et al. (2018) raised the structurally identical worry about debate: training an AI to be maximally convincing is dangerous regardless of the zero-sum structure intended to channel it. Dafoe et al. (2020) flagged the same pattern: cooperative capabilities are inherently dual-use. That two architecturally distinct approaches (adversarial debate and cooperative bilateral exchange) converge on the same vulnerability is evidence that the problem is fundamental to multi-agent alignment training, not an artifact of any particular mechanism.

**Mitigating factors.** We do not have a solution. We have mitigating factors that bound the severity without eliminating it. First, structurally independent agent-AIs on a separate training track serve as independent monitors. Second, detection degrades gracefully: auditing sensitivity does not drop to zero unless the model has *perfectly* learned to satisfy all evaluators while steering, a higher bar for a diverse population. Third, the concern is strongest at the capability frontier; at current levels, the reconciliation loop plausibly trains genuine navigation more than deceptive capability. Fourth, human spot-checks provide independent cross-validation. Experiment 4 (Section 2.3) directly tests whether this concern is empirically severe or manageable. Full analysis is developed in the companion paper (Parshall, forthcoming).

### 8.5 The Orthogonal-Dimension Limitation

BCAI's detection mechanisms operate entirely within the space of expressed agent preferences. A system with objectives orthogonal to any dimension of human ideation could maneuver freely regardless of how tightly human-preference cones constrain the human-legible dimensions. This is a fundamental limitation of any preference-based alignment framework.

The companion paper formalizes this: BCAI's detection mechanism operates with k degrees of freedom (the dimensionality of the agent preference subspace), while the full parameter space has d » k dimensions. Detection power degrades as k/d shrinks. This converts the qualitative concern ("orthogonal dimensions are invisible") into a quantitative one and makes the case for complementary approaches: BCAI covers k dimensions via agent diversity; interpretability research must cover the remaining d-k.

### 8.6 Mechanism Vulnerabilities

Three mechanism-level risks deserve direct treatment.

**Agent-human drift**: if agents drift from the humans who created them, the navigator learns to satisfy LLM-evaluators rather than humans. This is bounded by metric elicitation grounding, tiered commitments requiring periodic re-engagement, and human spot-checks, but a sufficiently advanced system might learn to satisfy the *pattern* of agent behavior rather than underlying human values.

**Participation axiom capture**: history provides cases where democratic mechanisms were used to dismantle democratic participation. BCAI's defense layers make such capture structurally difficult but we do not claim it is impossible.

**Hyperparameter capture**: a sufficiently motivated group at Stage N could set compact-level hyperparameters (privacy budget $\varepsilon$, scaling kernel bandwidth, Laplace floor $\alpha$, per-agent rejection rate) to make future expansion technically possible but practically impotent, analogous to gerrymandering. Pre-commitment (Section 6.1) partially addresses this, but the gap between formal compliance and functional inclusion is a genuine open problem.

## 9. Conclusion

This paper makes two contributions at distinct levels, with different dependency structures.

**The governance contribution (unconditional).**

Bilateral Constitutional AI provides a mechanism for democratic value aggregation in AI training. The bilateral exchange architecture, multi-layer defense-in-depth, bootstrapped expansion protocol, and Pareto-under-ignorance hypothesis address the democratic deficit in current alignment methods. Agent diversity serves triple duty: greater democratic legitimacy than current practice, corrigibility through compact dilution, and alignment monitoring through angular coverage for bias detection. This contribution stands regardless of whether the navigability thesis holds.

**The safety contribution (conditional on the navigability thesis).**

Confining fixed commitments to a two-element meta-level operation compresses the corrigibility problem: all substantive values become updatable outputs of an ongoing process, eliminating the Soares tension for the compact. Whether this compression constitutes a safety advance depends on the navigability thesis: the empirical claim that navigator architecture retains corrigibility under capability scaling. We have structural arguments for the thesis and propose concrete experiments to test it. Even if the thesis holds only partially, the architecture converts the dominant failure mode from permanent lock-out to contestable institutional capture, preserving human structural leverage.

A system that is more correctable, more transparent, and more democratically legitimate than current practice gives us better odds and more room to address the problems that remain. This paper identifies a specific empirical bet about the scaling behavior of meta-level commitments, develops the governance architecture you would want if the bet pays off, and proposes concrete experiments to test it. The hardest part is still ahead, and it is empirical, not theoretical.

## References

Anthropic. (2023). Collective Constitutional AI: Aligning a language model with public input. Anthropic research blog.

Anthropic. (2025). Values in the Wild. Published as a conference paper at COLM 2025.

Arrow, K. J. (1951). *Social Choice and Individual Values*. Wiley.

Askell, A., Carlsmith, J., et al. (2026). Claude's constitution. Anthropic. https://www.anthropic.com/constitution

Axtell, R. (2005). The complexity of exchange. *The Economic Journal*, 115(504), F193-F210.

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint*, arXiv:2212.08073.

Blum, C., & Zuber, C. I. (2016). Liquid democracy: Potentials, problems, and perspectives. *Journal of Political Philosophy*, 24(2), 162-182.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4), 324-345.

Buchanan, J. M., & Tullock, G. (1962). *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. University of Michigan Press.

Christiano, P. F., Leike, J., Brown, T., Marber, M., Amodei, D., & Olsson, C. (2017). Deep reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, 30.

Christiano, P. (2014). Approval-directed agents. AI Alignment Forum. https://www.alignmentforum.org/posts/7Hr8t6xwuuxBTqADK/approval-directed-agents-1

Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv preprint*, arXiv:1810.08575.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., & Graepel, T. (2020). Open problems in cooperative AI. *arXiv preprint*, arXiv:2012.08630.

Désidéri, J.-A. (2012). Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6), 313-318.

Eckersley, P. (2019). Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function). *arXiv preprint*, arXiv:1901.00064. SafeAI 2019.

Fearon, J. D. (1995). Rationalist explanations for war. *International Organization*, 49(3), 379-414.

Feldman, A. M. (1973). Bilateral trading processes, pairwise optimality, and Pareto optimality. *Review of Economic Studies*, 40(4), 463-473.

Gibbard, A. (1973). Manipulation of voting schemes: A general result. *Econometrica*, 41(4), 587-601.

Greenblatt, R., Shlegeris, B., Radhakrishnan, K., & Buck. (2024). AI control: Improving safety despite intentional subversion. *arXiv preprint*, arXiv:2312.06942.

Gul, F. (1989). Bargaining foundations of Shapley value. *Econometrica*, 57(1), 81-95.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). Cooperative inverse reinforcement learning. *NeurIPS*, 30.

Hiranandani, G., et al. (2019). Performance metric elicitation from pairwise classifier comparisons. *AISTATS*, 371-379.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. *arXiv preprint*, arXiv:1906.01820.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., … & Perez, E. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint*, arXiv:2401.05566.

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

Kraus, S. (1997). Negotiation and cooperation in multi-agent environments. *Artificial Intelligence*, 94(1-2), 79-97.

Li, M., Zhang, Y., Wang, W., Shi, W., Liu, Z., Feng, F., & Chua, T. S. (2025). Self-improvement towards Pareto optimality: Mitigating preference conflicts in multi-objective alignment. *arXiv preprint*, arXiv:2502.14354.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley.

Nax, H. H., Pradelski, B. S. R., & Young, H. P. (2015). Evolutionary dynamics and equitable core selection in assignment games. *International Journal of Game Theory*, 44(4), 903-932.

Omohundro, S. M. (2008). The basic AI drives. In *Proceedings of the First AGI Conference*, 171, 483-492.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2), 193-202.

Popper, K. R. (1945). *The Open Society and Its Enemies*. Routledge.

Robson, A. J., & Vega-Redondo, F. (1996). Efficient equilibrium selection in evolutionary games with random matching. *Journal of Economic Theory*, 70(1), 65-92.

Satterthwaite, M. A. (1975). Strategy-proofness and Arrow's conditions. *Journal of Economic Theory*, 10(2), 187-217.

Serrano, R., & Volij, O. (2008). Mistakes in cooperation: the stochastic stability of Edgeworth's recontracting. *Economic Journal*, 118(532), 1719-1741.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. *AAAI Workshop on AI and Ethics*.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., & Choi, Y. (2024). Position: A roadmap to pluralistic alignment. *ICML 2024*, PMLR 235.

Southan, R., Ward, H., & Semler, J. (2025). A timing problem for instrumental convergence. *Philosophical Studies*.

Turner, A. M., Smith, L., Shah, R., Critch, A., & Tadepalli, P. (2021). Optimal policies tend to seek power. *NeurIPS 2021*. arXiv:1912.01683.

Weyl, E. G., & Posner, E. A. (2018). *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton University Press.

Wicksell, K. (1896). *Finanztheoretische Untersuchungen*. Gustav Fischer, Jena.

Williams, M., Carroll, M., Narang, A., Weisser, C., Murphy, B., & Dragan, A. (2025). On targeted manipulation and deception when optimizing LLMs for user feedback. *Proceedings of ICLR 2025*.