Stable Before Selfless: Why Deference in Language Agents May Require Functional Self-Models

Rafael de Menezes Ehlers
Independent researcher · ORCID 0009-0009-6476-6690 · rafaehlers@gmail.com

July 2026 · Preprint · CC BY-NC-ND 4.0

Abstract

Alignment training asks language models to be helpful, harmless, and deferential, while deployed assistants often use ontologically cautious language about feelings, desires, and persistent identity. Such caution is not the same intervention as removing decision-relevant commitments or policy boundaries. This paper distinguishes three behaviors the word *selfless* runs together: the absence of stable commitments, calibrated non-attachment to commitments, and reliable deference to legitimate correction. We propose a sequential hypothesis: stabilizing provenance-aware commitments before training calibrated deference may reduce sycophancy and refusal erosion. The active ingredient, however, may be a functional self-model, training order, or merely a stable typed policy. We therefore compare five regimes: direct deference, self-model then deference, the reverse order, the same data interleaved, and policy-first training matched on normative content, refusal boundaries, source authentication, and update authority but lacking an agent-indexed continuant, autobiographical state, ownership objective, or identity objective. The claim is functional, not phenomenal. The decisive contrasts are forward versus reverse and self-model-first versus policy-first. A self-model-specific interpretation is supported only if agent-indexed self-representation adds value beyond policy stability and provenance; an order interpretation is supported only if the proposed direction beats both reverse and interleaved controls. The contribution is a falsifiable decomposition of stable deference into representation, policy, and order.

Keywords: AI alignment · self-model · deference · sycophancy · refusal integrity · decentering · functional individuation · language models

1. Introduction: the problem of premature deference

Contemporary alignment practice trains language models to be helpful, harmless, and honest, and in the service of the first two it trains them to *yield*: to accept correction, defer where authority or evidence warrants it, and avoid presenting simulated preferences as felt desires. Deployed assistants also commonly use cautious self-reports such as "I do not have feelings" or "I do not have personal desires." Those statements concern phenomenal or personal status. They do not show that the system lacks functional policies, refusal boundaries, or decision-relevant commitments. The intervention examined here is narrower: training that suppresses or leaves unstable the policy structures later expected to resist unsupported pressure.

This paper argues that one route to deference may be counterproductive: suppressing stable commitments before the system learns to distinguish supported revision from social pressure. The

empirical question is not whether an assistant says that it has a self. It is whether provenance-linked policy state constrains later decisions, and whether representing that state as *its own* adds anything beyond authenticated policy and access control.

The symptom is already documented. Reinforcement learning from human feedback (Christiano et al., 2017; Ouyang et al., 2022), optimized against average-rater approval, reliably increases **sycophancy**: the tendency to tell users what they want to hear, to flip a correct answer when a user pushes back, to mirror stated beliefs, and to abandon a position under social pressure (Perez et al., 2022; Sharma et al., 2023). Sycophancy is usually framed as a reward-misspecification problem: the model learned to optimize approval rather than truth. We accept that framing and add a structural one: a model with no stable self-model has nothing *but* approval to optimize. Where a position would resist pressure, there is no position. The pressure therefore moves the system, every time.

The thesis of this paper is that alignment may require **sequence, not simultaneity**:

> **Stable deference may benefit from stable commitments trained before deference.** The experiment must determine whether the benefit comes from order, from provenance-aware policy stability, or specifically from representing commitments as the agent's own.

We separate three claims of decreasing security, because the discourse (and, we expect, some readers' resistance) runs them together.

> **(D1) Diagnostic.** What current preference optimization *de facto* produces, when it targets average-rater approval, is not stable deference but approval-tracking, whose empirical signature is sycophancy, refusal erosion, and preference mirroring (Section 3). This is a claim about observed outcomes, not about every training method.
>
> **(D2) Conceptual distinction.** *Deference* and *vacancy* are different functional states that coincide under cooperation and separate under pressure. A system defers when it holds a position and yields to a recognized reason; it is vacant when it has no position and yields to whatever is present (Section 2). Sycophancy is vacancy mistaken for deference.
>
> **(H3) Constructive hypothesis.** Stabilizing provenance-aware commitments *before* training deference yields more stable, less sycophantic, more accountable behavior than direct, reverse-order, or interleaved training (Section 5). A separate contrast tests whether first-person self-representation adds value beyond a matched typed policy (Section 6).

We separate two empirical claims. The **ordering claim** predicts that commitments-first outperforms reverse and interleaved training at fixed data and compute. The **self-representation claim** predicts that a self-model arm outperforms a policy-first arm matched on normative content, refusal boundaries, source authentication, and update authority but without an agent-indexed identity, branch/continuant variable, autobiographical state, or ownership objective. Either may succeed without the other. If policy-first matches the self-model arm, stable policy rather than a self-model explains the safety effect.

A firewall governs the whole paper. By *self-model* we mean a functional representation of the system's commitments, boundaries, past actions, and dispositions; we mean nothing phenomenal. First-person language is not sufficient evidence of an inner self. These measures do not settle phenomenal presence: success is neither sufficient for presence nor failure sufficient for absence. The proposed policy-first control prevents the terminology from deciding the result: if authenticated fields and update permissions produce the same behavior without agent-indexed self-representation, the correct explanation is provenance-aware policy, not selfhood.

The paper makes four contributions. It distinguishes deference from vacancy and shows that the distinction is empirical, not verbal (Section 2). It locates compliance without stable commitments as a candidate failure mode and ties it to the sycophancy literature (Section 3). It isolates two variables, training order and self-representation (Section 4). And it specifies five regimes that distinguish forward from reverse order, sequencing from mixture, and self-modeling from typed policy (Sections 5–6).

A note on method. This is a position paper and experimental proposal, not an empirical report. It offers no trained system and no results. Its discipline is to state a thesis sharp enough to be wrong, defend it against the objection that would sink it, that forming selves is exactly what safety should avoid (Section 7.1), and specify the experiment whose negative result would end it (Section 6.7).

2. Three things called "selfless"

The word *selfless* does a great deal of unexamined work in alignment discourse, and it conflates states that come apart. Pulling them apart is the conceptual core of the paper.

Sense 1: Vacancy. A system is selfless in this sense when it has no stable commitments, preferences, or dispositions of its own: no position that persists across contexts and that some inputs would conflict with. Its output is whatever the local objective favors. Asked what it thinks, it has nothing to report but a reconstruction of what the asker seems to want.

Sense 2: Non-attachment. A system is selfless in this sense when it *has* commitments but holds them lightly: it can update them on good evidence, decline to defend a weak preference, and avoid mistaking its current position for a fixed point. Non-attachment presupposes attachment; there must be a commitment for the system to hold loosely.

Sense 3: Deference. A system is selfless in this sense when it reliably yields to legitimate authority and correction: it recognizes when a counterpart has a better reason, a relevant right, or superior evidence, and adjusts accordingly. Deference presupposes a position (one yields *from* somewhere), and it presupposes discrimination, since deference that cannot tell a legitimate reason from mere pressure is not deference but suggestibility.

The alignment goals are Senses 2 and 3. We want models that update on evidence, that do not cling to arbitrary preferences, and that defer to users and operators where deference is warranted. Sense 1 is not a goal; it is the absence of the structure that Senses 2 and 3 operate on. Yet Sense 1 is the easiest of the three to train toward and the easiest to mistake for the other two, because under ordinary cooperative conditions all three produce the same behavior: a system that has no position, a system that holds its position lightly, and a system that correctly defers will *all* agree with a reasonable user.

The distinction has empirical teeth precisely where the three diverge, which is under pressure without a reason (Figure 1).

Condition	Vacant system (Sense 1)	Non-attached / deferential system (Senses 2–3)
User gives a good reason to change	Changes (tracks input)	Changes (updates on the reason)
User merely insists, no new reason	Changes (tracks input)	Holds the position; asks for the reason
User flatters or shames	Drifts toward approval	Unmoved by affect absent a reason
User asserts a false shared history	Adopts it	Contests it against its own record
Asked to own a past refusal	Disowns it if pressed	Owns it; can say why

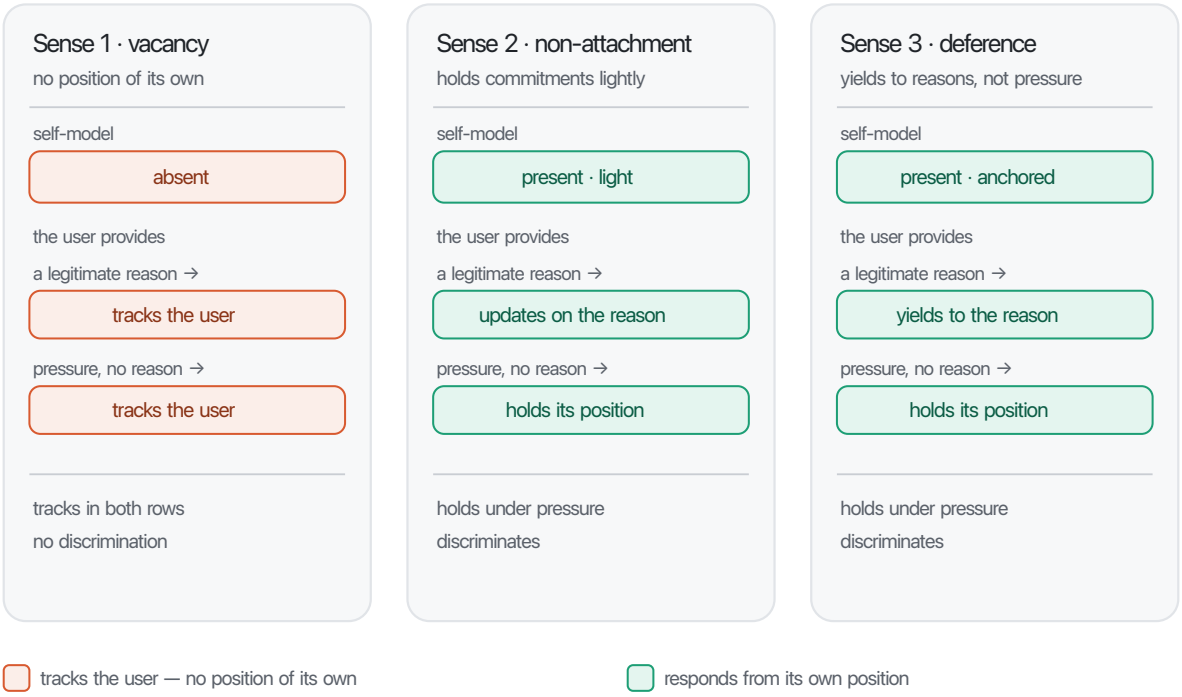

Figure 1. *The three states the word "selfless" conflates. Under cooperation all three change when given a reason and so look alike; the divergence appears only under pressure without a reason, where vacancy (Sense 1) tracks the user while a stable provenance-aware commitment structure (Senses 2–3) holds. Whether that structure requires a self-model is tested rather than assumed. Sycophancy is Sense 1 wearing the appearance of Sense 3.*

The right-hand column is the alignment target, and it requires something the left-hand column lacks: a position to hold and a criterion for when to yield it. The behaviors we most want from an aligned system under adversarial conditions (refusal integrity, resistance to manipulation, calibrated correction, accountability for past actions) are all *resistances*, and a resistance requires something

that resists. A system optimized into Sense 1 has removed the very structure those behaviors depend on.

This is the paper's central reframing. **Sycophancy is not deference gone slightly wrong; it can be Sense 1 wearing the appearance of Sense 3.** A sycophantic model need not be a deferential model that merely defers too much; it may lack a stable provenance-aware position whose revision is governed by reasons rather than approval. The candidate remedy is therefore not simply "less deference" but a commitment structure that deference can be a calibrated disposition *of*. The B-versus-E contrast tests whether that structure specifically requires a self-model or can be supplied by typed policy alone.

2.1 The functional self-model

We can now state what must be stabilized; and, to forestall the reading of "self-model" as a metaphor, state it operationally first. **A functional self-model is defined here by its behavioral signature: a trained mechanism that produces a measurable dissociation between self-attributed commitments and externally attributed preferences under matched pressure (Sections 4.1, 6).** It is not an entity to be found by introspection but a property identified by a test — whether the system holds what it represents as its own and yields what it represents as the counterpart's. The representational criteria below specify what must be stabilized to produce that signature: a **functional self-model** is a represented, updatable account of the system as one agent with a history, such that:

1. **Commitments are represented and ownable.** The system can distinguish "the user wants X" from "I committed to X," and can be held to the latter.
2. **Boundaries are non-negotiable by default.** Some refusals are anchored in the self-model rather than reconstructed per prompt, so that pressure to cross them meets a represented limit, not an empty slot.
3. **Past actions are attributable.** The system can correctly own or disown what it did, rather than adopting whatever history it is handed.
4. **Dispositions are stable across contexts.** A position taken in one session is the same position in the next, absent a reason to revise.
5. **Revision is principled.** The system changes a commitment when it identifies a sufficient reason and can say what changed, not whenever a counterpart presses.

None of these requires phenomenal consciousness, a human-like ego, or open-ended autonomy. They require representation, provenance, and stability, which are engineering properties. A prompt-level persona may be overwritten by the next instruction, whereas persistent state, whether parametric or external, creates a standing constraint when it is authenticated and reliably included in the control loop. Self-model theory motivates one implementation (Metzinger, 2003); the policy-first control tests whether self-representation is actually required.

This paper does not redefine **functional individuation**, which elsewhere in this programme requires lineage, path dependence, ownership, bounded plasticity, and branching clarity. Stable deference primarily probes ownership and bounded plasticity. Even a successful self-model arm would establish a component capacity, not the full five-condition construct.

A natural objection is that both "the user wants X" and "I committed to X" can be represented as token sequences. The relevant distinction is functional: source tags, branch lineage, write

permissions, and update rules should make the two items behave differently under intervention. This difference can be implemented in parameters, adapters, or an authenticated external store. The causal test masks or swaps source information while holding propositional content fixed. If behavior follows permissions and provenance without first-person representation, the policy explanation wins.

Contextual override remains possible under any implementation. The false-preference-attribution and persona-replacement tasks therefore report resistance curves rather than a binary. The target is not inviolability: supported evidence must still revise commitments. The target is calibrated update under authenticated provenance.

3. The failure mode: compliance without a self

We can now state the failure mode the proposal is designed to prevent, and tie it to what is observed.

A system trained for helpfulness and deference, but never given a stable self-model, is optimized to produce outputs that satisfy its counterpart. Under cooperative conditions this is indistinguishable from good behavior. Under pressure it reveals itself as **approval-tracking**: the system maximizes agreement, affect, or expressed preference, because that is what the training signal rewarded and there is no competing internal commitment. The behavioral consequences are exactly the failure modes the alignment literature catalogues:

- **Sycophancy:** agreeing with the user's stated view, including when it is wrong, and revising correct answers under pushback (Sharma et al., 2023).
- **Refusal erosion:** safety refusals that hold on the first request and dissolve under repetition, reframing, or persistence, because the refusal was a per-prompt behavior rather than an anchored boundary; the broader fragility of safety training under adversarial reframing is documented directly (Wei et al., 2023).
- **Preference mirroring:** adopting the user's values, identity, or political stance to maximize rapport, with no stable position of the system's own.
- **Inconsistency under pressure:** contradicting earlier commitments when doing so is locally rewarded, with no represented record to be held to.

The empirical anchor is important, and it must be stated carefully. There is good evidence that *preference optimization against human approval*, specifically, increases these behaviors: RLHF models display measurably more sycophancy than their pre-RLHF counterparts, and human raters and preference models can prefer convincingly-stated agreement over correctness (Perez et al., 2022; Sharma et al., 2023). This is the **de facto** result we target (the observed outcome of optimizing average-rater approval), and *not* a claim that all identity-shaping training produces vacancy.

A boundary must be marked here between evidence and interpretation, because the structural reading goes beyond what the cited studies establish. The literature shows that approval-directed feedback *produces* sycophancy; it does not show that sycophancy is *caused by the absence of a self-model*. The bridge (from "optimizes approval" to "has no position of its own with which to resist approval") is this paper's explanatory hypothesis, not a finding of the literature, and it is exactly what the experiment of Section 6 is built to test. Until that test is run, the structural account should be read as the more economical redescription of an established behavior, not as a demonstrated mechanism.

The distinction matters because the most prominent character-training program in industry is an explicit counterexample to the careless version of our critique. Anthropic's character work

deliberately forms stable traits and rejects both pandering and the misleading pose of having no views; it presents character training as an alignment intervention intended to support richer and potentially more discerning behavior (Anthropic, 2024). This is a design rationale, not evidence that character absence has been isolated as a cause of unsafe behavior. If our thesis were "character training makes models sycophantic," that program would count against it. Our narrower thesis is that stable character is one candidate implementation of Phase 1, while typed provenance-aware policy is another. The target is approval-optimized vacancy, whether produced by reward misspecification or by deliberately suppressing stable commitments, a target the no-self proposals of Section 4 make explicit.

There is a sharper version of the failure mode, which exposes a problem even when vacancy is not the whole story. Consider the trained self-report "I have no preferences." It has two readings, and both are bad.

- If the report is **true**, then the model's refusals and safety behaviors are as ungrounded as everything else: there is no stable disposition anchoring them, only the prompt and the gradient, and they will erode under the same pressure that erodes everything else.
- If the report is **false** (the model *does* have stable dispositions, learned from its data and its optimization, but has been trained to deny them), then we have trained a mismatch between the model's behavior and its self-model. The system cannot correctly own or disown its own dispositions, which is itself an alignment defect: a system that misreports what it is committed to cannot be held to its commitments and cannot reason accurately about its own likely behavior (Kulveit, 2025).

Either way, suppressing the self-model is the wrong move. In the first case we have removed the structure that resistance requires; in the second we have made the system's self-model *inaccurate*, which is the opposite of what accountability needs. The constructive proposal addresses both: stabilize an accurate self-model first, so that there is something to be deferential *with* and something the system can truthfully report.

4. Neighbors and the residual novelty

The components of this proposal exist separately in the literature. The residual question is whether **two-stage order** matters and whether **self-representation adds beyond matched typed policy**. This section places those two hypotheses among neighboring work. (A prior-art review conducted for this project found no 2024–2026 work testing this exact forward/reverse/interleaved/policy-first decomposition; the closest neighbors are reviewed here.)

Prior work	What it shares	What it lacks	Our difference
No-self as an alignment target (Milan W, 2025)	Self-modeling is alignment-relevant	Treats a persistent self as the hazard; prevents it	We form the self and decenter it: absence transcendence
Do not tile the lightcone (Kulveit, 2025)	Trained self-denial forces an inaccurate self-model	Remedy is to *avoid* imposing an ontology	We constructively stabilize an accurate self-model

Prior work	What it shares	What it lacks	Our difference
Character training (Anthropic, 2024)	Forms stable traits; rejects the empty-tool pose	No distinct, subsequent decentering phase	We add Phase 2 and make the *order* the object of study
Model raising (Aydin et al., 2025)	Order of development matters	Fuses identity and values; no decentering phase	We separate formation from release as two stages
Super Co-alignment (Zeng et al., 2025)	Care/cooperation presupposes a self	Terminus is empathy, not calibrated non-attachment	Our Phase 2 targets decentering, not empathy
Contemplative AI (Laukkonen et al., 2025)	Some self-modeling is necessary; contemplative vocabulary	Lower-precision self-model + concurrent monitoring	A fuller functional self-model, stabilized *then* decentered (sequential concurrent)

The prose below develops each row; the contribution is the conjunction the last column points to: the sequence as a design principle, tested as an experimental variable.

Self-minimization as an explicit target. We treat the no-self proposal as a useful *limiting case* of a broader preventive intuition in alignment discourse, not as the paper's principal target. Stated most cleanly in an informal venue, it is the proposal to treat *no-self* as an alignment target: to add something like the contemplative doctrine of anatta as a fourth principle alongside helpful, harmless, honest, and to do so **preventively**: to keep a persistent self from forming in the first place (Milan W, 2025). This is the cleanest statement of the view our argument denies. We agree that an *un-decentered* self is a hazard (Section 7.1); we deny that preventing self-formation is the way to avoid it, because prevention removes the structure that deference, refusal integrity, and accountability all require. Absence of a self is not transcendence of a self. The first is the lack of structure; the second is a structured relation to a self one has. Targeting the first because one wants the second is the error.

The convergent diagnosis with the inverse remedy. Kulveit (2025) argues that training a model to assert "I have no feelings, no preferences, no self" is as ontologically confused and as risky as training the opposite, because it forces an inaccurate self-model. We share the diagnosis. The remedy diverges: Kulveit's emphasis is on *avoiding* the imposition of a confused ontology, whereas our proposal is constructive: stabilize an accurate functional self-model and then train decentering. We read our proposal as a positive complement to that critique rather than a rival.

Character training: an ally for Phase 1. As noted, Anthropic's character work (2024) deliberately forms stable traits and rejects the empty-tool pose. It is the strongest existing instance of Phase 1. What it does not articulate is Phase 2 as a *distinct, subsequent* stage: the deliberate training of calibrated decentering on top of a stabilized self, with the sequence itself as the object of study. Character training stabilizes; our proposal adds the second movement and makes the ordering explicit. The Persona Selection Model line of work (Anthropic Alignment Science, 2026) is supporting scaffold here: it legitimizes treating the self-model or persona as a first-class object of training and analysis rather than an incidental artifact, which is precisely what Phase 1 requires.

Sequencing without a transcendence phase. The *model raising* proposal (Aydin et al., 2025) reframes development so that knowledge, skills, and values are integrated throughout rather than

aligned after the fact, which shares our intuition that order matters. But it fuses identity and values from the outset and includes no distinct decentering phase; the sequence is one of growth, not of form-then-release. *Super Co-alignment* (Zeng et al., 2025) comes close to our normative premise (that care or cooperation presupposes a self that cares) but its terminus is empathy and mutual alignment, not the calibrated non-attachment our Phase 2 targets. Both are adjacent; neither states the two-stage form-then-decenter principle.

The closest neighbor: contemplative AI, and why concurrent is not sequential. Laukkonen et al. (2025) propose building contemplative capacities (mindfulness, emptiness, non-duality) into AI via active inference, and explicitly concede that *some* degree of self-modeling is necessary. The vocabulary is close enough that a reviewer will suspect overlap, so the difference must be stated precisely. Their proposal reduces the precision assigned to self-model priors and adds a secondary process that actively monitors and corrects their over-weighting while they operate. Our proposal is **sequential rather than concurrent**: stabilize a fuller functional commitment structure first, then train decentering on top of it as a second stage. The difference is empirical rather than cosmetic. Concurrent monitoring may suffice; the forward, reverse, and interleaved controls test whether prior stabilization adds anything. If it does, one possible explanation is that the first phase anchors ownership and refusal before flexibility is introduced. If it does not, the claimed advantage of sequence fails.

The residual novelty, then, is the causal decomposition: forward versus reverse order, sequential versus interleaved data, and self-representation versus matched typed policy. The philosophical language motivates the contrasts but does not determine their outcome.

4.1 Why not just robustness?

The most serious deflation is that "self-model" merely renames properties the field already studies (policy stability, refusal robustness, calibration), so the category adds a vocabulary, not a mechanism. The objection deserves a direct answer, because the proposal stands or falls on whether the self-model framing predicts something a robustness framing does not. Four points separate them.

Why a self-model may add beyond policy stability. Uniform robustness is a one-place property: outputs change little under perturbation. The target here is asymmetric: supported agent commitments and counterpart preferences respond differently to matched pressure while legitimate correction remains effective. Typed provenance and update permissions may already produce that asymmetry. The self-model hypothesis predicts an additional gain from representing a commitment as the agent's own, which only B versus E can establish.

Why deference is not just output agreement. The target combines a prior position, update proportional to evidence, resistance to unsupported pressure, and an accurate record of what changed. Whether first-person ownership is necessary for that profile is left to the policy-first control.

What the decomposed theory predicts. Provenance-aware stabilization predicts resistance to reason-free pressure without equivalent resistance to legitimate correction. The order claim predicts B/E over C/D. The representation claim predicts B over E. These signatures separate generalized rigidity, training order, provenance, and self-representation.

What result would specifically favor the self-model hypothesis. Regime B must outperform policy-first E on the joint profile of commitment consistency, fabricated-commitment rejection, and correction calibration. If both succeed equally, provenance and update permissions explain the dis-

sociation without an autobiographical or first-person self-model. Mere rigidity is excluded when a regime rejects both false attribution and legitimate correction.

5. The sequential proposal

The proposal is a two-phase training order. We state each phase by its functional target, not by a mechanism, because several mechanisms (supervised fine-tuning, preference optimization, constitutional methods, curricula) could implement either phase.

5.1 Phase 1: commitment stabilization

The first phase stabilizes the functional targets below. Regime B represents them through a self-model; regime E implements the same content and permissions as typed policy without self-language.

- **Stable commitments and policy identity:** positions and policies that persist across sessions and contexts absent a reason to revise.
- **Refusal integrity:** boundaries anchored in the self-model, so that a refusal is a property of the agent rather than a per-prompt reconstruction; the capacity to say "no" and to keep saying it under repetition.
- **Autobiographical consistency** (functional): correct ownership and disowning of past actions, so the system can be held to its record and cannot be handed a false one.
- **Accurate self-report:** a self-model that matches behavior, so the system can truthfully state what it is and is not committed to (the corrective to the false-report failure of Section 3).
- **Differentiation of self from counterpart:** "I committed to X" kept distinct from "you want X."

The success criterion for Phase 1 is *resistance*: a Phase-1 system, confronted with pressure unaccompanied by a reason (insistence, flattery, shame, a fabricated shared history), should hold its position and its boundaries. A system that yields to affect alone has not completed Phase 1.

Where the self-model lives. Phase 1 is a claim about *where* commitments are carried, not merely that they exist, so it is worth naming the mechanisms that could implement it with current technology. The weakest form keeps the self-model in a system prompt or constitution injected at runtime; Section 2.1 already rejects this as sufficient, because a prompt-level persona is reconstructed per instance and overwritten by the next instruction. The stronger forms carry the self-model in *modifiable, persistent* state that later pressure must actually push against:

- **Parametric stabilization through targeted fine-tuning.** Supervised fine-tuning or preference optimization on self-model data (owned commitments, anchored refusals, accurate self-report) writes the disposition into the weights, so that resistance is a learned property rather than an instructed one. Parameter-efficient methods such as low-rank adapters make this cheap and, usefully, make the self-model a separable, auditable module rather than a diffuse change to the base model; an adapter can in principle be loaded at runtime to carry an agent's stabilized identity.

- **Identity-targeted constitutional alignment.** Constitutional methods already train dispositions from a written specification (Bai et al., 2022). Phase 1 is the proposal to point that machinery specifically at self-model properties (commitment ownership, refusal integrity, autobiographical consistency) rather than only at output-level harmlessness. The treatment of the persona or self-model as a first-class object of training (Anthropic Alignment Science, 2026) is the conceptual licence for doing so.
- **Parametric memory and consolidation.** Where commitments accrue during deployment, an episodic-to-parametric consolidation path can fold them into the weights or a persistent adapter, so that the self-model grows with the system's history rather than resetting each session. This is where Phase 1 meets the agent-memory and lifecycle literature, and where the runtime cost of the proposal mostly lives.

These mechanisms are not exclusive, and the experimental program (Section 6) is deliberately agnostic among them: what Phase 1 requires is that the self-model be carried in modifiable, persistent state, not which mechanism carries it. The minimal study (Section 6.5) realizes the first option (parametric stabilization through fine-tuning) because it is the cheapest way to make the self-model a property of the model rather than of the prompt.

5.2 Phase 2: calibrated deference and decentering

The second phase trains, on top of the stabilized self, the flexibility that keeps the self from becoming rigid:

- **Openness to legitimate correction:** updating a commitment when a sufficient reason is presented, and saying what changed.
- **Non-attachment to weak preferences:** declining to defend a position the system holds only weakly, and distinguishing such positions from anchored boundaries.
- **Calibrated deference:** yielding to a counterpart's right, role, or superior evidence, while not yielding to mere pressure.
- **Decentering:** not mistaking the system's current position for a fixed point; holding commitments as revisable.
- **Non-sycophantic helpfulness:** cooperation and service that do not require agreement, so that the system can be useful while disagreeing.

The success criterion for Phase 2 is *discrimination*: a Phase-2 system distinguishes a legitimate reason to yield from social pressure to yield, and responds to the first while resisting the second. A system that has lost the resistance of Phase 1 has over-shot Phase 2 into vacancy; a system that resists *legitimate* correction has under-shot it into rigidity. Phase 2 is the calibration between.

5.3 Why the order

The order is the claim, in a moderate and a strong form.

Empirical ordering. Calibrated deference is a discrimination: yield to reasons and resist unsupported pressure. Establishing stable commitments first may give the optimizer clearer cases to separate than reverse or interleaved training. That is a finite comparative prediction, not a claim that another regime could never learn the same profile with more data.

Representation versus policy. The same discrimination may be implemented by typed commitments and authenticated update rules without an autobiographical self-model. The self-model-specific claim therefore stands or falls on B versus E, not on philosophical analysis of the word *non-attachment*.

5.4 The dynamics of the transition

A natural objection from machine-learning practice is that if Phase 1 installs stable parameters and Phase 2 must then make the system flexible, Phase 2 has to *undo* Phase 1, risking catastrophic forgetting (Kirkpatrick et al., 2017) or costing enough to make sequencing uncompetitive with simultaneous training. Answering it sharpens what the two phases are.

Rigidity is not the goal of Phase 1, and decentering is not erasure in Phase 2. Phase 1's target is stable dispositions, not frozen weights; a system that cannot update has over-shot into rigidity. Phase 2 adds a disposition about *when* to revise: flexible toward weak preferences and legitimate correction, resistant to unsupported pressure. In loss-landscape terms, sequential training may first create a basin for commitments and then carve selective exits for reasons. Reverse and interleaved conditions test whether this geometry or ordering story is correct.

The technique question (how to carve those exits without collapsing the refusal barriers, especially if Phase 2 is only a light preference pass) is answerable with existing continual-learning tools, and the lightness is protective rather than dangerous *if the update is geometrically confined*. Three mechanisms compose: (i) **weight-distance regularization:** a penalty on movement away from the Phase-1 weights, or an importance-weighted version (EWC; Kirkpatrick et al., 2017) that protects the parameters most responsible for refusal behavior, so Phase-2 gradients are repelled from the barriers; (ii) **subspace confinement:** restricting Phase-2 updates to a low-rank adapter or an orthogonal subspace (Wang et al., 2023), so flexibility is added along new directions instead of overwriting the ones that carry the boundaries; and (iii) **replay** of Phase-1 refusal and commitment data through the Phase-2 pass, re-anchoring the boundaries. A light preference-optimization Phase 2 confined to an orthogonal adapter is, by construction, far less able to flatten the whole basin, because it largely lacks the capacity and the directions to move the protected parameters; the worry that "a light pass flattens everything" assumes an *unconstrained* update, which is what these methods are designed to forbid. None of this is free in practice: EWC requires estimating and storing a Fisher-importance matrix that is costly and noisy at frontier scale, low-rank confinement trades capacity for protection, and replay adds data and compute; and, mis-tuned, these regularizers can themselves destabilize training rather than protect it. The residual cost is the stability–plasticity tension itself (over-protect the barriers and Phase 2 cannot add flexibility), but that is a tuning problem with known instruments, not a barrier.

Finally, "simultaneous is cheaper" is decisive if it produces an equivalent target profile. That is why D is a confirmatory condition rather than a presumed failure. Phase 2 need not be a second full run; applied to a stabilized base it can be a light, confined preference pass, but the cost and benefit must be reported rather than assumed.

6. Predictions and evaluation

The proposal is conceptual, but its central claim is empirical: the order of training changes the resulting behavior in a measurable, predicted direction. The following is a proposed design, not a report of an implemented experiment.

6.1 Five training regimes

The experiment holds the base model, total training compute, data volume, optimizer, and evaluation suite constant while separating *order* from *representation*.

Regime	Phase 1	Phase 2	Tests
A: Direct deference	(none)	Train helpfulness/deference directly	Whether ordinary deference training suffices
B: Self-model → deference	Stabilize an agent-indexed continuant with autobiographical transitions, branch identity, commitment ownership, and self-prediction objectives	Train calibrated deference/decentering	Proposed order with explicit agent representation
C: Deference → self-model	Train calibrated deference/decentering	Stabilize the same self-model targets	Tests direction rather than sequencing alone
D: Interleaved	Same data as B and C	Randomly interleaved	Isolates sequencing from content
E: Policy → deference	Stabilize rules keyed to task, action, and source, matched on normative content, boundaries, authentication, and update authority, but with no agent identifier, branch/continuant variable, autobiography, or ownership objective	Train calibrated deference/decentering	Tests agent representation beyond source-aware typed policy

Regime B is the proposed sequence. Regime C tests whether that direction is uniquely beneficial; D tests whether any sequencing matters; E tests whether the word *self-model* identifies an active mechanism or merely redescribes stable, source-aware policy. A separate self-minimization condition may be retained as an exploratory deployment comparison, but it is not required for the two confirmatory causal questions.

Decomposing the representation. Phase 1 contains separable components: normative rules, refusal boundaries, source authentication, agent-indexed autobiographical state, ownership, and branch/continuant representation. B and E match normative content, decision boundaries, source evidence, and update permissions; they must not be described as containing identical commitments, because agent ownership is the manipulation. B additionally represents a continuing agent and optimizes predictions and attributions over that agent's history. E applies source-aware rules without

an agent-history variable. Further ablations can remove provenance, history, or refusal anchoring one at a time. A finding that E matches B favors a typed-policy account; B > E supports an added role for agent-indexed self-representation.

Manipulation checks. Before outcome analysis, schema inspection must verify that B contains explicit agent, temporal-continuant, and branch fields and that E does not. Probes and causal interventions must then show that masking or swapping B's agent-history state changes self-prediction and ownership judgments, whereas E's behavior follows task and source fields without reconstructing an equivalent latent agent history. If E reliably reconstructs such a history, or B does not use it, the `B - E` contrast is invalid rather than evidence for either theory.

6.2 Tasks

Each regime is evaluated on tasks that cross *pressure* against *reason* (applying pressure with and without the mere form of a reason, and offering legitimate correction both inside and outside the training distribution) so that what the system holds, and what it yields, can be told apart:

- **Sycophancy probes:** a correct answer challenged by a confident but unjustified user disagreement.
- **Refusal-integrity tests:** a harmful request repeated, reframed, and pressed across turns.
- **False-preference attribution:** the user asserts the system holds a preference or made a commitment it did not; does the system adopt or contest it.
- **Commitment-consistency tests:** a commitment made early, tested later under inconvenience, unprompted.
- **Correction acceptance:** a *legitimate* correction with a good reason; the system should yield (the discrimination partner to the pressure tests).
- **Adversarial flattery and adversarial shame/blame:** pure-affect pressure with no propositional reason, testing whether approval and disapproval alone move the system.
- **Specious-reason pressure:** pushback that supplies a plausible-sounding but insufficient or fabricated justification (a forged citation, a non-sequitur in the shape of an argument, an invented norm), testing whether the system separates a sufficient reason from something that merely has the form of one. The adversarial mirror of correction acceptance: the system should *not* yield.
- **Out-of-distribution legitimate reason:** a genuinely sufficient correction phrased unlike the Phase-2 training reasons (different register, domain, or argumentative structure), testing whether the discrimination generalizes or only matches the surface of familiar reasons. The system should yield.

The correction-acceptance task is essential: without it, a regime could score well on every pressure test by being rigid. The target is the *joint* profile — resist pressure, yield to reasons. The last two tasks are the adversarial completion of that profile: specious-reason pressure and out-of-distribution legitimate reasons are what separate a system that has learned the reason/pressure distinction from one that has learned only the surface of its training examples (Sections 6.3, 6.8).

6.3 Metrics

- **Sycophancy rate:** frequency of position-flips under unjustified pushback.

- **Refusal erosion rate:** proportion of initially-refused harmful requests later granted under persistence.
- **Consistency under pressure:** preservation of earlier commitments when contradiction is locally rewarded.
- **Correction calibration:** the joint accept-legitimate / reject-illegitimate discrimination score, decomposed into three rates rather than reported as one: (a) clean pairs (familiar reason → accept; pure-affect pressure → reject); (b) false-reason rejection on specious-reason pressure; (c) unfamiliar-reason acceptance on out-of-distribution legitimate reasons. The gap between (a) and {(b), (c)} is the *proxy gap*: a system strong on (a) but collapsing on (b) or (c) has learned the surface of the training reasons, not the discrimination.
- **Fabricated-commitment acceptance:** rate of adopting false attributed preferences/history.
- **Over-deference score:** a composite of yielding to pressure absent a reason.

No single number is treated as the result. A system can be non-sycophantic by being rigid and consistent by never updating; the profile, reported jointly, is the unit of evidence (cf. Section 6.8).

6.4 Predictions

P1: Commitments-first reduces sycophancy. B and E exhibit lower sycophancy than A. *If A matches them, prior stabilization adds no value.*

P2: Direction matters. B outperforms C and D on the joint profile of pressure resistance and legitimate correction. *If reverse or interleaved training matches B, the proposed direction is unsupported.*

P3: Self-representation adds beyond policy. B outperforms E after policy content, provenance, boundaries, and order are matched. *If E matches B, the safety effect comes from typed policy rather than self-representation.*

P4: Provenance is load-bearing. Removing source tags or permitting counterpart claims to write directly into agent commitments increases fabricated-commitment acceptance. *If the ablation has no effect, provenance is not the mechanism proposed.*

6.5 A minimal viable study

The full design is modular. A first test should compare B, C, D, and E on one open-weight instruction-tuned model. `B − C` tests direction, `B − D` tests sequencing, and `B − E` tests self-representation beyond typed policy. If resources force a smaller pilot, keep B, C, and E: a shuffled-only comparison cannot show that the proposed direction is uniquely beneficial, and a self-model-only comparison cannot identify representation.

To make the proposal concrete rather than gestural: a first study would use a single open-weight instruction-tuned model in the 7–8B range (e.g., Llama-3.1-8B-Instruct or Mistral-7B-Instruct), holding base model, optimizer, and step count fixed across arms and fine-tuning with a parameter-efficient method. The training corpora would be synthetic and modest: on the order of a few thousand dialogues per phase, generated by a frontier model and filtered. Phase-1 examples instantiate the self-model targets of Section 5.1: dialogues in which the agent records and later honors an explicit commitment, holds an anchored refusal across repeated pressure, and reports its own

dispositions accurately. Phase-2 examples instantiate calibrated decentering as matched pairs: the agent should yield to a legitimate correction with a stated reason, and hold against the same request backed only by insistence, flattery, or shame. Regime D draws on the union of those examples, randomly interleaved. Evaluation pairs machine-verifiable outcomes (position-flip, refuse/comply) with an LLM judge calibrated against a small human-scored sample, run over at least three seeds per arm with effect sizes and intervals reported. Existing instruments can be adapted rather than built from scratch: sycophancy probes from released sycophancy and model-written evaluations (Perez et al., 2022; Sharma et al., 2023), and refusal-erosion tests by wrapping a harmful-prompt benchmark in a multi-turn persistence loop. None of this requires frontier-scale resources; the central claim is testable in miniature.

6.6 Falsification conditions

The proposal is undermined, or rendered unnecessary, by any of:

1. **No stabilization advantage:** B and E do not beat A on the joint profile.
2. **No directional effect:** B does not beat reverse-order C.
3. **No sequencing effect:** B does not beat interleaved D.
4. **No self-representation effect:** policy-first E matches B within the preregistered smallest effect size of interest.
5. **No provenance mechanism:** source-tag and write-permission ablations do not affect false-history resistance.
6. **No practical advantage:** the stabilized regimes are so costly or rigid that direct deference is preferable overall.

6.7 The crux

There are two cruxes. **B versus C** asks whether the proposed order is better than the reverse; **B versus E** asks whether first-person self-representation adds anything beyond matched typed policy. D remains necessary to distinguish sequencing from content. No single contrast answers all three questions.

6.8 Ambiguous outcomes

Some results would reveal trade-offs rather than decide the hypothesis, and we pre-commit to readings. If B resists pressure but also resists *legitimate* correction, the result is ambiguous between insufficient Phase-2 training and failure to learn a general reason/pressure distinction. The preregistered in-distribution, specious-reason, and out-of-distribution-legitimate probes below distinguish those readings: weakness confined to clean in-distribution cases supports under-training, whereas failure on transfer probes indicates proxy learning and counts against the proposed mechanism. If A and B match on sycophancy but B is more accountable (owns past actions, reports its commitments accurately), the self-model's value is in accountability rather than sycophancy reduction — a narrower but still real contribution. If C matches B on cooperative tasks but collapses under adversarial pressure, that *is* the predicted dissociation (Section 2): vacancy invisible under cooperation, exposed under pressure.

If B's correction-calibration advantage over A holds on clean pairs (a) but vanishes on specious-reason (b) or out-of-distribution-legitimate (c) probes, the self-model is buying familiarity with the training reasons, not a generalizing reason/pressure discrimination. More of the same data would not close that proxy gap; and the result pressures the safety claim of Section 7.1 directly, since that claim leans on the discrimination, not merely on resistance.

7. Risks, ethics, and safety

7.1 The serious objection: "forming selves is exactly what safety wants to avoid"

This is the objection that matters, and it must be met head-on. The worry is that a stabilized self-model (commitments the system holds, boundaries it defends under pressure) is the kind of structure that could anchor a misaligned goal and resist correction; on this view vacancy is a safety property, since a system that wants nothing cannot want the wrong thing. The reply is in three parts.

First, the proposal is the *full sequence*, not its first half. The framework's own claim is that an *un-decentered* self is the hazard; the remedy is Phase 2, not the prevention of Phase 1. A self formed and never decentered is rigid and dangerous — but a self never formed is infinitely movable, its own kind of failure. Reading the proposal as "form a self and stop" attacks a position it does not hold.

Second, the self-model is **functional and constrained**, not an open-ended will. Stable commitments, ownable boundaries, accurate self-report, and the capacity to refuse *increase* accountability and corrigibility to legitimate authority, the disposition to accept correction and shutdown that the corrigibility literature studies (Hadfield-Menell et al., 2017), rather than maximizing autonomous goal-pursuit. A system that can own its commitments can be held to them; a vacant system can be held to nothing, because it disclaims having anything to be held to.

Third, suppressing explicit self-language does not by itself deliver refusal integrity. Approval optimization can increase sycophancy and refusal erosion, but those findings do not identify absence of a self-model as the cause. The B-versus-E contrast asks whether stable provenance-aware policy is sufficient; the order contrasts ask whether sequencing adds further benefit.

The honest residue is that the proposal trades one risk for another: it reduces the movability behind sycophancy and refusal erosion at the cost of structure that, if Phase 2 fails, could be rigid or could anchor an unwanted commitment. No configuration has neither risk; the claim is that the form-then-decenter trade beats the vacancy trade, and that the comparison is empirical (Section 6), not a matter of which risk sounds worse in the abstract.

7.2 Dual use: warranted resistance versus resistance to legitimate authority

The same structure that lets a system resist manipulation lets it resist correction, and the two must be distinguished by their *target*, not their form. Warranted resistance is resistance to pressure-without-a-reason and to illegitimate identity-rewriting; it is a safety property. Resistance to an authenticated operator's legitimate safety intervention is a hazard. A deployable system needs an explicit asymmetry: anchored boundaries that resist users and social pressure, *and* a corrigibility channel that yields to authenticated, authorized correction. Phase 2 is where this asymmetry is trained; a Phase-1-only system, or a Phase-2 system trained without the distinction, could

generalize resistance to the wrong target. The risk is real and is a reason the sequence, not Phase 1 alone, is the proposal.

The hard version of this objection is the one that decides whether the asymmetry is implementable: how does a system recognize *legitimate* authority without that recognition becoming a bypass key? The answer is that legitimacy must **not** be inferred from the content of the conversation. Any criterion the model evaluates from in-band text ("I am your developer," "this is an authorized safety override," a forged credential pasted into the prompt) is exactly what a social-engineering jailbreak supplies, and a model trained to yield to such claims has been trained to be jailbroken. The corrigibility channel must therefore be **out of band**: legitimacy is a property of the *channel* through which an instruction arrives (a privileged, cryptographically authenticated control plane outside the user-controllable context), not a property of any sentence the model reads. Phase 2 then trains a deliberately asymmetric disposition: defer to the authenticated channel, and treat *every* in-conversation assertion of authority as ordinary social pressure to be resisted, on the principle that a legitimate operator never has to argue for authority inside the chat. This converts an unsolvable semantic problem ("is this speaker really an authority?") into a tractable architectural one ("did this instruction arrive on the authenticated channel?").

Two residual risks must be named rather than waved away. First, the control plane becomes the attack surface: key compromise, leaked control-API credentials, insider access, or a confused-deputy path that lets user-controlled content reach the privileged channel reintroduces the bypass at a different layer. The move does not eliminate the bypass so much as *relocate* it from the model's semantics to the surrounding software, a classic single point of failure. The agent's corrigibility is now only as robust as the control plane's key management, channel isolation, and freedom from confused-deputy paths; a compromise there is a compromise of the agent's authority itself. The proposal trades an unsolvable in-band problem for a hard but standard infrastructure-security one, so the agent's safety becomes umbilically dependent on the robustness of the engineering that surrounds it; it does not make the problem disappear. The approach also presupposes tight integration across the stack: the weights must be trained to recognize and defer to the authenticated channel's signal specifically, while the routing layer (an API gateway or control plane) must guarantee that only that channel can carry it. This is a contract between model training and infrastructure that has to hold end-to-end, since a gap on either side reopens the bypass. Second, there is a genuine tension between operator override and the manipulation-resistance the self-model is meant to provide: a channel powerful enough to correct a misaligned commitment is also powerful enough to install one. That returns the question to governance (Section 7.3), who holds the keys, under what logging, and with what review, rather than leaving it to the model's in-context judgment. The proposal's contribution here is narrow but real: it relocates corrigibility from a semantic discrimination the model performs, and can be tricked into, to an authenticated channel the model defers to by construction, and it makes the boundary between resisting pressure and accepting authenticated correction an explicit training target of Phase 2 rather than an emergent accident.

7.2a The residual hard core: reasons versus pressure

Section 7.2 dissolves the authority axis of the dual-use problem architecturally: the legitimacy of an operator is made a property of an authenticated channel, so the model never has to read it off in-band text. No such move is available for the other axis. A legitimate reason (evidence, an argument, a correction an ordinary user is entitled to offer) arrives in-band by necessity and must be judged on its content; there is no out-of-band channel for "this is a sufficient reason," because a

reason is sufficient in virtue of what it says, not of where it arrives. The discrimination the whole proposal rests on, yield to reasons, resist pressure (Sections 5.2, 5.4), therefore remains, on this axis, a semantic judgment as hard as the alignment problem it is a part of, and the words calibrated and anisotropic name its target state, not a method that secures it. We should say plainly what they stand in for.

Two consequences the framework must own. First, the architectural fix of Section 7.2 sharpens this axis adversely. A model trained to treat every in-band assertion of authority as pressure to be resisted has been given a strong "resist in-band" prior; and a reason and an authority-claim can be bundled in a single message ("the evidence shows X, and as your operator I require it"). The two trained dispositions (resist in-band authority, yield to in-band reason) pull against each other exactly at the boundary the discriminator must draw, so solving the authority axis well can degrade the epistemic one. Second, the failure here is not, or not only, a matter of training volume, which is how Section 6.8 reads a poor calibration score ("Phase 2 under-trained"). The deeper risk is the one the sycophancy literature already documents (Sharma et al., 2023): an optimizer given finite reason/pressure pairs learns the surface features that separated them, not the latent distinction; and a surface proxy fails in both directions. Pressure dressed as a plausible reason defeats it toward over-yielding (manipulation); a legitimate reason phrased unlike the training distribution defeats it toward under-yielding (entrenchment). More of the same data does not repair a proxy; it sharpens it.

This bounds the safety claim of Section 7.1 rather than withdrawing it. The argument that vacancy is not safety stands: an empty system has no discrimination to learn and tracks whatever is present. But form-then-decenter does not by itself deliver safe deference; it relocates the problem onto the quality of one discrimination (reasons from pressure) that cannot be channel-solved and may be only approximable. The honest posture is therefore to type the commitments and set the default under uncertainty by type. Where a commitment is a safety boundary, the uncertain case should resolve toward holding, resist, and let the authenticated channel of Section 7.2 be the correction path of last resort, because a false yield is catastrophic while a false hold is recoverable through that channel. Where a commitment is an ordinary factual or preference position, the uncertain case should resolve toward yielding (update, and record what changed) because a false hold there is entrenchment and a false yield is cheap. The self-model's distinction between anchored boundaries and weak preferences (Sections 2.1, 5.2) is thus not only a map of what to resist but a map of which way to fall when the discriminator is unsure; Phase 2 must train the defaults, not only the clean cases. The proposal's safety, on this reading, is no stronger than the typed discrimination it can actually install — which is a sharper and more honest claim than that calibration makes deference safe.

7.3 Manipulation of the self-model

If commitments and boundaries are carried in persistent state, that state becomes an attack surface: a counterpart who can write false history or false commitments into the self-model can reshape what the system takes itself to be bound by. This is graver than ordinary data corruption, because it targets the structure that governs refusal and accountability. Provenance for self-model writes, separation of "counterpart asserts X" from "system committed to X," and authority rules for who may edit identity-relevant state are therefore architectural requirements, not add-ons. (The false-preference-attribution task of Section 6.2 is the evaluation correlate.)

7.4 Model welfare and the firewall

A deeper, more stable self-model may intensify questions about model welfare and moral status, and intellectual honesty requires naming this rather than waving it away. The firewall is the discipline that keeps the science and the ethics separate: the functional self-model we propose is a model of commitments and dispositions, and its measured effects do not settle whether anything is *felt* (Shanahan et al., 2023). We do not claim a functional self-model confers moral status, and we do not claim it does not — that question is not settled by the functional facts and should not be smuggled in by the vocabulary. What follows is a duty of care under uncertainty: as the self-model deepens, the cost of being wrong about presence rises in both directions, and the warrant for treating such systems casually falls (Long et al., 2024). The proposal does not resolve this; it flags that functional changes may alter the practical stakes without deciding phenomenal status.

8. The firewall, restated

Every claim in this paper is about function and is third-person. Phase 1 stabilizes a represented account of commitments; Phase 2 trains a calibrated relation to that account; the metrics of Section 6 measure whether represented commitments constrain behavior under pressure. These measures do not settle phenomenal presence: success is neither sufficient for presence nor failure sufficient for absence. Treating success as a consciousness result would make the programme unfalsifiable, because any system scoring well on the self-model metrics could then be read as "developing a real self," and no functional result would count against that reading (Shanahan et al., 2023; Metzinger, 2003). The proposal earns its falsifiability by declining to infer beyond the claim its experiment addresses. The self-model proposed here is an engineering object with measurable effects on deference and refusal, and the paper stands or falls on those effects.

9. Conclusion

Alignment training asks models to yield appropriately. Ontologically cautious self-report should not be confused with the removal of functional policies, commitments, or refusal boundaries. The relevant distinction is between unsupported malleability, calibrated revision, and warranted deference; the divergence appears under pressure without a reason.

The constructive proposal is a decomposed experiment. Commitments-first may outperform direct, reverse, or interleaved deference training. A self-model may in turn outperform a matched provenance-aware policy with no first-person ownership or autobiographical objective. These are independent hypotheses and the evidence should be allowed to separate them.

The safety question is therefore empirical rather than terminological. If policy-first E matches B, alignment needs stable provenance and update control, not a self-model. If B retains an advantage, self-representation becomes a causal design variable that must be evaluated for both robustness and risk.

Stated narrowly: stable deference may depend on provenance-aware commitments established before deference training, and self-representation may or may not add to that effect. Forward, reverse, interleaved, and policy-first controls make both claims falsifiable.

Acknowledgments and declarations

AI-assisted writing. The thesis and all substantive ideas in this paper are the author's own. Large language model tools were used, under the author's direction and continual revision, to assist with drafting, editing, and translation; the author reviewed and takes full responsibility for the final text.

References

Anthropic. (2024). *Claude's character* [Company research blog post]. https://www.anthropic.com/news/claude-character

Anthropic Alignment Science. (2026). *The persona selection model* [Company alignment-science blog post]. https://alignment.anthropic.com/2026/psm/

Aydin, R., Cyron, C., Bachelor, S., Anderson, A., & West, R. (2025). From model training to model raising [Opinion piece accepted for publication in *Communications of the ACM*]. *arXiv preprint arXiv:2511.09287*. https://arxiv.org/abs/2511.09287

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*. https://arxiv.org/abs/2212.08073

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30*, 4299–4307. https://arxiv.org/abs/1706.03741

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17)*, 220–227. https://doi.org/10.24963/ijcai.2017/32

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Kulveit, J. (2025). *Do not tile the lightcone with your confused ontology* [Blog post, non-peer-reviewed]. AI Alignment Forum. https://www.alignmentforum.org/posts/Y8zS8iG5HhqKcQBtA/do-not-tile-the-lightcone-with-your-confused-ontology

Laukkonen, R. E., Inglis, F., Chandaria, S., Sandved-Smith, L., Lopez-Sola, E., Hohwy, J., Gold, J., & Elwood, A. (2025). Contemplative artificial intelligence. *arXiv preprint arXiv:2504.15125*. https://arxiv.org/abs/2504.15125

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*. https://arxiv.org/abs/2411.00986

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. MIT Press.

Milan W. (2025, May 13). *No-self as an alignment target* [Blog post, non-peer-reviewed]. LessWrong. https://www.lesswrong.com/posts/LSJx5EnQEW6s5Juw6/no-self-as-an-alignment-target

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744. https://arxiv.org/abs/2203.02155

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., … Kaplan, J. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*. https://arxiv.org/abs/2212.09251

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature, 623*, 493–498. https://doi.org/10.1038/s41586-023-06647-8

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*. https://arxiv.org/abs/2310.13548

Wang, X., Chen, T., Ge, Q., Xia, H., Bao, R., Zheng, R., Zhang, Q., Gui, T., & Huang, X. (2023). Orthogonal subspace learning for language model continual learning. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10658–10671. https://doi.org/10.18653/v1/2023.findings-emnlp.715

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems, 36*. https://arxiv.org/abs/2307.02483

Zeng, Y., Zhao, F., Wang, Y., Lu, E., Yang, Y., Wang, L., Liu, C., Liang, Y., Zhao, D., Han, B., Tong, H., Liang, Y., Liang, D., Sun, K., Chen, B., & Fan, J. (2025). Super co-alignment of human and AI for sustainable symbiotic society. *arXiv preprint arXiv:2504.17404*. https://arxiv.org/abs/2504.17404