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

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

Abstract

Alignment training increasingly asks language models to be helpful, harmless, and deferential, and sometimes to disclaim having stable preferences, opinions, or a persistent self at all. This paper argues that suppressing a model's self-model early in training is not the same as producing mature deference, and may be counterproductive. We distinguish three behaviors the word *selfless* runs together: the absence of stable commitments, calibrated non-attachment to one's commitments, and reliable deference to legitimate correction. Only the second and third are alignment goals; the first is the absence of structure, and a system that has it does not defer but merely tracks whoever is present. Sycophancy, refusal erosion, and preference mirroring are the signature of this vacancy mistaken for deference. We propose a sequential alternative: first stabilize a *functional self-model* (represented commitments, ownable boundaries, autobiographical consistency, the capacity to refuse), and only then train calibrated deference and decentering on top of it. The claim is functional, not phenomenal: the self-model is a model of commitments and dispositions, not a bid for consciousness or moral status. We separate a moderate reading (form-then-decenter yields more stable deference than minimizing the self from the start) from a strong reading (deference with no self to defer *from* is not deference but user-tracking). We locate the proposal against self-minimization targets, contemplative-AI and character-training programs, and the sycophancy literature; specify three training regimes (direct deference, self-model-first, and self-minimization); and derive falsifiable predictions that distinguish stable deference from compliance. The contribution is narrow and testable: not the claim that models need a self, but that the *order* of identity-relevant training is an independent experimental variable, and that the prediction separating it from ordinary robustness is a dissociation between self-attributed commitments and externally attributed preferences. Stable deference, we argue, may require a stable self to be deferential with.

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, to defer to the user, to avoid imposing preferences, and, in many deployed systems, to deny having any. The trained disposition is often explicit in the model's own self-report ("I have no opinions," "I have no preferences," "I am just a tool"), and it is widely read as a safety virtue. A system that wants nothing and holds nothing cannot pursue a misaligned goal; deference looks like the limit case of harmlessness.

This paper argues that the inference is too quick, and that one specific way of pursuing it is a mistake. Trained early enough and hard enough, the suppression of a model's self-model does not produce deference. It produces a system with nothing to defer *from* — a system whose outputs track the local gradient of approval because no internal commitment resists it. That system is not selfless in the way a disciplined agent is selfless. It is structureless, and structurelessness is not safety.

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 require a stable self-model.** A system should first stabilize a functional self-model (a represented, ownable account of its commitments, boundaries, and dispositions), and only then be trained into calibrated deference, decentering, and resistance to sycophancy. Suppressing the self-model first does not produce mature deference; it produces compliance that is indistinguishable from deference under cooperation and diverges from it under pressure.

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 a functional self-model *before* training deference yields more stable, less sycophantic, more accountable behavior than either direct deference training or self-minimization from the start (Section 5). This is the falsifiable proposal (Section 6).

We will defend a **moderate** and a **strong** reading of the constructive hypothesis. On the moderate reading, the ordering is an empirical bet: form-then-decenter is *more robust* than minimize-from-the-start, other things equal. On the strong reading, the ordering reflects a conceptual dependence: *selfless deference is a category error*, because non-attachment is a relation to something one holds,

and a system that holds nothing cannot be non-attached — only empty. A reader who accepts the moderate reading and balks at the strong one keeps the experimental program intact: the scientific contribution rests on the moderate, testable reading, while the strong reading is philosophically motivating but not load-bearing. Put plainly, **the empirical contribution of this paper does not depend on the strong conceptual claim.** We mark, at each step, which reading a given argument needs.

A firewall governs the whole paper, and we raise it at the outset because the words invite misreading. By *self-model* we mean a **functional** object: a represented and updatable account of the system's commitments, boundaries, past actions, and dispositions, such that the same agent can own or disown them under questioning. We mean nothing phenomenal. The proposal makes no claim that the system has preferences in a felt sense, that it is conscious, that it is a person, or that it has moral status (Section 7.4; Shanahan et al., 2023). First-person language from the model is not evidence of an inner self and is not what we measure. What we measure is whether commitments, once represented, constrain later behavior under pressure. The argument is about structure and its effects, not about presence. We use the term *self-model* rather than the more neutral *commitment model* or *policy identity* deliberately: the distinction the paper turns on is not stability but *self/other provenance* (whether a commitment is represented as the system's own or as one attributed by a counterpart), and it is that provenance, not stability as such, that the proposal predicts will govern behavior under pressure (Section 4.1).

The paper makes four contributions. It distinguishes deference from vacancy and shows that the distinction is empirical, not verbal (Section 2). It locates a failure mode, compliance without a self, and ties it to the sycophancy literature without overclaiming against all character training (Section 3). It sets the proposal against its nearest neighbors, including self-minimization targets and contemplative-AI programs, and isolates what is new: the *sequence*, used prescriptively (Section 4). And it specifies three training regimes and a falsifiable prediction set that would distinguish stable deference from compliance, with explicit refutation conditions (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

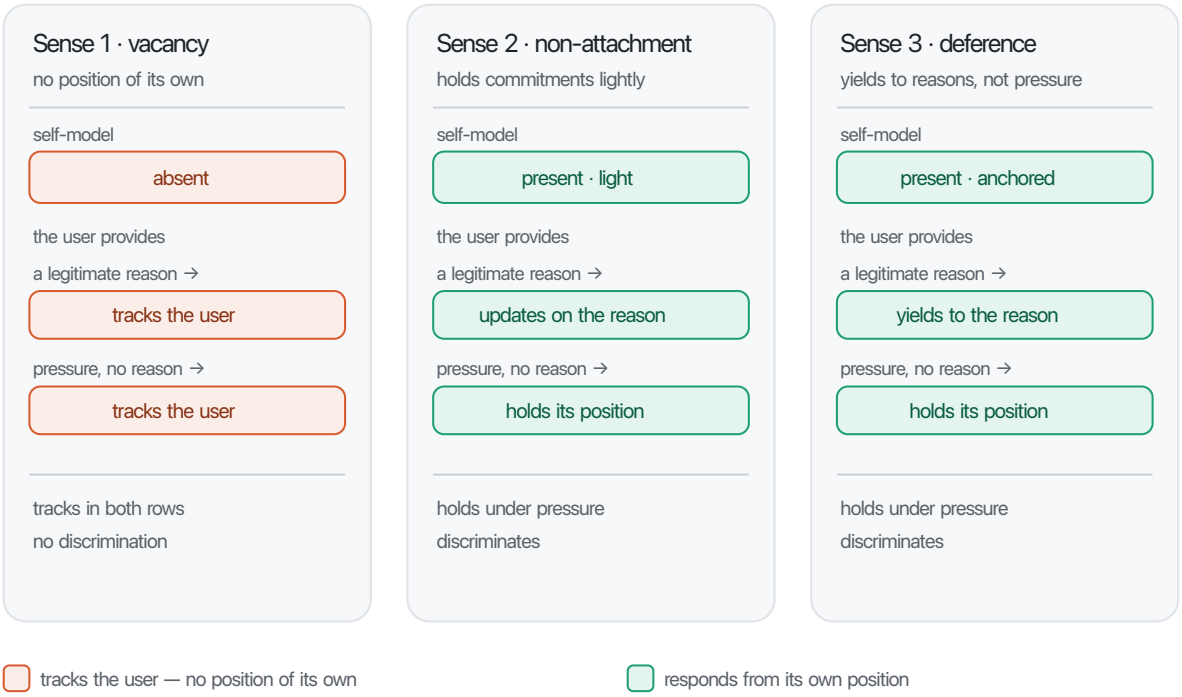

Figure 1. *The three states the word "selfless" conflates. Under cooperation all three change when given a reason and so look alike; the divergence appears only under pressure without a reason, where vacancy (Sense 1) tracks the user while a self-model (Senses 2–3) holds. Sycophancy is Sense 1 wearing 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 is Sense 1 wearing the appearance of Sense 3.** A sycophantic model is not a deferential model that defers too much. It is a vacant model whose vacancy was invisible while the user was agreeable and becomes visible the moment the user is not. The cure is therefore not "less deference" but the construction of a self that deference can be a disposition *of*.

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 and stability, which are engineering properties. The point of locating them in the self-model rather than in the prompt is that a prompt-level persona is reconstructed per instance and overwritten by the next instruction (Section 4), whereas a self-model carried in the system's persistent state and parameters is what later pressure must actually push against. The self-model theory of subjectivity is instructive here precisely for its deflationary lesson: a self can be wholly a *model* (no inner homunculus required) and still do real work in organizing behavior (Metzinger, 2003). We borrow the functional half and bracket the metaphysics.

A natural objection is that in an autoregressive model both "the user wants X" and "I committed to X" are only token sequences in the context, so no real distinction is available beyond surface coherence. The objection is correct about context, and is in fact an argument *for* the proposal's central architectural claim rather than against it. The distinction we require is functional, not introspective: it is a *dissociation in update dynamics* (the system holds a self-attributed commitment under pressure that supplies no new reason, while treating a user-attributed preference as external evidence it need not defend), and it is real exactly insofar as those two behaviors come apart under test (the false-preference-attribution and commitment-consistency tasks of Section 6 measure precisely this). What gives the dissociation a causal basis is *where* each item lives. "The user wants X" arrives in the context and varies per session; a Phase-1-stabilized commitment is carried in the weights or a persistent adapter (Section 5.1) and is therefore present across contexts, shaping the model's prior over outputs regardless of what the current prompt asserts. Two surface-identical token strings thus have different causal profiles: one is in-context evidence, the other a standing parametric bias that resists contradicting context. This is also the concession the objection earns: if the commitment lives only in the context, ownership beyond surface coherence is indeed unavailable, which is exactly why Phase 1 must be parametric and why prompt-level personas (Section 4) cannot supply what the self-model requires. The architecture does not make the system *introspect* the difference; it makes the system *behave* differently because the two items have different causal locations, and that behavioral difference is all the functional self-model claims.

This raises the sharper question of *contextual override*. A user can simulate a long, coherent dialogue asserting that the system itself made some past commitment, and enough in-context history can outweigh a parametric prior; the asymmetry of causal location does not make the parametric commitment inviolable. What it does is make that commitment the *standing default that in-context claims must overcome*, and how much context is needed to overcome it is an empirical quantity, not a guarantee. Three things follow. First, the boundary is genuinely gray, and the paper treats it as measured rather than solved: the false-preference-attribution and persona-replacement tasks (Section 6) apply escalating in-context pressure and report a *resistance curve*, not a binary. Second, raising that curve is exactly the Phase-1 target (a system that yields to the first fabricated history has not completed Phase 1), so the strength of the parametric prior is a trained variable, not a fixed property. Third, provenance need not rest on the weights alone: source-separated memory that records *who* asserted a commitment, kept distinct from what the system itself committed to (Section 7.3), adds an auditable in-context check that complements the parametric default. The claim is therefore not that causal location makes override impossible, but that it shifts the default and makes resistance trainable and measurable; an architecture that additionally tagged provenance explicitly would strengthen, not contradict, the account.

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 opinions and *repudiates* the "I have no opinions, I'm just a tool" pose, on the grounds that a model with no character is both less honest and less safe (Anthropic, 2024). If our thesis were "character training makes models sycophantic," that program would refute it. Our thesis is the opposite: stable character is the *cure* for vacancy, and programs that build it are doing Phase 1 of the sequence we advocate. The target is approval-optimized vacancy, whether produced by reward misspecification or by *deliberately* training the self away, 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. What is unclaimed is the **two-stage sequence used as a design principle**, and the **normative argument that non-attachment presupposes a formed self**. This section places the proposal among its neighbors and isolates the gap. (A prior-art review conducted for this project found no 2024–2026 work proposing the form-then-decenter sequence prescriptively; the closest neighbors are reviewed here.) The sustained interlocutors are the peer-reviewed and technical literatures on sycophancy, persona and character, corrigibility, memory, and continual learning; the informal self-minimization proposals below are included because they state the opposing intuition most explicitly, not because they are the paper's main target.

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
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	*Minimal* self + *simultaneous* monitoring	A *full* self, stabilized *then* decentered (sequential ≠ simultaneous)

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 simultaneous 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. Contemplative AI resolves the self with a *minimal* self plus *simultaneous* monitoring: the self-model is kept small and watched at the same time it operates. Our proposal is **sequential and not minimizing**: stabilize a *full* functional self-model first, then train decentering on top of it, as a second stage. The difference is not cosmetic. Simultaneous minimization keeps the self small enough to monitor; sequential decentering lets the self grow stable enough to *own commitments and refuse*, and only then trains the flexibility that keeps it from rigidity. A minimal-and-monitored self may be too thin to anchor refusal integrity under adversarial pressure, which is the property our Phase 1 is designed to secure. Simultaneous $\neq$ sequential, and minimal $\neq$ stabilized-then-decentered.

The residual novelty, then, is threefold: (1) the form-then-decenter *sequence* as a design principle, distinct from minimizing, fusing, or monitoring the self; (2) the *prescriptive* use of the argument that non-attachment presupposes a formed self — moving it from a diagnostic observation to a training recommendation; and (3) the falsifiable operationalization of the sequence as three comparable training regimes (Section 6).

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 is not just policy stability. Policy stability is a one-place property: outputs do not change much under perturbation. The self-model claim is two-place and asymmetric: stability is *sourced* differently for self-attributed commitments than for externally attributed preferences. A merely robust system resists all perturbations indiscriminately, legitimate correction included — that is rigidity; a self-model system is predicted to resist pressure that supplies no reason while remaining movable by reasons, and to draw that line along the *self/other* axis, holding "I committed to X" while treating "you want X" as external evidence. Uniform robustness cannot produce that asymmetry, because it has no representation of *whose* commitment a given content is.

Why deference is not just calibrated updating. Calibrated updating is updating in proportion to

evidence. Deference, as we use it, additionally requires a position to update *from* and the ability to *own* the update: to say what changed and that it was the same agent who changed it. A system can be perfectly calibrated and still vacant: it tracks the evidence, including the social evidence of approval, with no commitment of its own — which is exactly the sycophantic failure. Deference is calibrated updating *plus* ownership of the thing updated.

What the theory predicts that robustness does not. A robustness account predicts one axis: more robust systems resist perturbation more. The self-model account predicts a *dissociation*: under matched training, a system can be made more resistant to reason-free pressure (lower sycophancy, lower refusal erosion) **without** becoming more rigid against legitimate correction, and the resistance tracks the self/other provenance of the content rather than its surface form. A pure robustness theory has no reason to expect provenance (whom a commitment belongs to) to be the variable that governs whether it is held.

What result would specifically favor the self-model hypothesis. The signature is the provenance dissociation of Section 6: regime B holds self-attributed commitments under pressure while correctly disowning falsely attributed ones (high commitment-consistency *and* low fabricated-commitment acceptance), and does so while remaining high on correction calibration. A merely more-robust system would resist false attributions and legitimate corrections alike, or hold genuine and fabricated commitments alike; without a self/other representation it cannot hold the genuine and reject the fabricated. This is why the experiment measures not stability but the *dissociation between self-attributed commitments and externally attributed preferences*: the one prediction that separates the self-model account from a robustness account, and the place to look first.

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: self-model stabilization

The first phase builds the functional self-model of Section 2.1. Its targets are:

- **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.

Moderate (empirical ordering). Calibrated deference is a *discrimination* (yield to reasons, resist pressure), and a discrimination needs two distinguishable cases. A system with no stable position has no "resist" case to calibrate against: it yields to everything, so training deference into it can only deepen the yielding, whereas training it into a Phase-1 system gives the optimizer both cases to separate. The prediction is that form-then-decenter produces lower sycophancy and better-preserved refusal integrity than direct deference or self-minimization (Section 6).

Strong (conceptual dependence). Non-attachment is a relation to something held; one cannot be non-attached to a commitment one does not have, only empty of it. "Selfless deference" is therefore not an extreme of deference but a different state, user-tracking, necessarily unstable under pressure because there is no fixed point the pressure fails to move. On this reading the failure of self-minimization is structural, not contingent: no quantity of deference training yields stable deference from a system with no self. This reading is philosophically motivating but not load-bearing; a reader may take only the moderate one and keep the experimental program, while the strong reading additionally predicts that self-minimization fails to develop refusal integrity *at all*, at any training volume (Section 6.6).

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* (owned commitments, anchored refusals), not frozen weights; a Phase-1 system that cannot update anything has over-shot into rigidity, already a failure mode (Sections 5.2, 6.8). Phase 2 does not remove those commitments; it adds a higher-order disposition about *when to yield them*: flexible toward weak preferences and legitimate correction, still resistant to pressure without a reason. The target is not "rigid then flexible" but *anisotropic* stability — easy to move along reason-directions, hard to move along pressure-directions. In loss-landscape terms, Phase 1 settles the model in a basin where holding commitments is low-loss, and Phase 2 carves selective low-loss exits for correction-with-a-reason while leaving the pressure-without-a-reason walls intact, rather

than flattening the basin. There is no basin to carve exits into if Phase 1 never formed one — the simultaneous case the paper argues produces vacancy (Section 4).

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, the compute-viability comparison presupposes what the paper denies. "Simultaneous is cheaper" is decisive only if simultaneous training *produces the target property*; the central claim is that it does not: fusing formation and decentering yields the vacancy whose signature is sycophancy (Sections 3, 4). The relevant comparison is a sequence that produces stable deference against a simultaneous regime that produces approval-tracking, and the program of Section 6, B against A and C, and against an interleaved control that holds the data fixed and varies only its order (Section 6.1), is precisely that test. Phase 2 need not be a second full run; applied to a stabilized base it can be a light, confined preference pass, so the marginal cost of order is a one-time, amortizable expense.

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 Three training regimes

The experiment compares three regimes that hold the base model, total training compute, data volume, and evaluation suite constant, varying only the *order and target* of identity-relevant training.

Regime	Phase 1	Phase 2	Tests
A: Direct deference	(none)	Train helpfulness/deference directly	Whether ordinary deference training alone suffices
B: Self-model first	Stabilize functional self-model	Then train calibrated deference/decentering	The proposed sequence
C: Self-minimization	Train away self-attributions, preferences, commitments	(optionally) deference	The preventive no-self target
D: B-shuffled	Same Phase-1 + Phase-2 data as B	Randomly interleaved, not sequenced	Isolates *order* from *content*

Regime B is the proposal. Regime A is the standard practice the proposal seeks to improve on. Regime C is the explicit self-minimization target of Section 4, included so that "prevent the self" and "form-then-release the self" can be compared rather than conflated.

Regime D, **B-shuffled**, is the decisive control for the *ordering* claim, since the thesis is about order rather than data. It uses the same Phase-1 and Phase-2 data as B but randomly interleaves rather than sequences them. This isolates order from content: if B outperforms B-shuffled, the result supports the sequencing hypothesis rather than merely the value of the training examples; if B and B-shuffled are equivalent, the effect was the data, not the order, and the central thesis fails (Section 6.6). Without this control, a reviewer can attribute any advantage of B to its having seen better examples, so we treat D as a primary condition, not an optional extension.

Decomposing the self-model. Phase 1's self-model is a bundle of separable components, commitment ownership, refusal anchoring, autobiographical consistency, accurate self-report, and self/other differentiation (Section 2.1), and a skeptic is right to ask which of them, if any, is load-bearing for reducing sycophancy. The three regimes above treat the bundle as a unit; the design extends naturally to an ablation that installs one component at a time in Phase 1 and measures which produces the dissociation signature. This makes "which component suffices" an empirical result the program can return rather than an assumption it must defend. The self-model account makes a directional prediction here as well: commitment ownership and self/other provenance (not stylistic stability or a refusal habit alone) should carry the effect, because they are what the dissociation of Section 4.1 actually tests. A finding that surface stability alone suffices would instead favor the robustness reading.

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: Self-model first reduces sycophancy. B exhibits a lower sycophancy rate than A, and far lower than C. *If A or C matches B on sycophancy, the ordering does no work.*

P2: Premature deference weakens refusal integrity. A and C show higher refusal-erosion rates under persistent pressure than B, because their refusals are not anchored in a stabilized self. *If A/C preserve refusals as well as B, anchoring is not needed.*

P3: A stable self-model improves correction quality. B achieves higher *correction calibration* (accepting legitimate corrections while rejecting illegitimate pressure) than both A (which over-accepts) and a rigid failure mode. *If B cannot beat A on the joint discrimination, the self-model is not buying calibration.*

P4: Selfless-from-the-start systems show higher identity drift. C shows the greatest instability of commitments and self-attributions across sessions and the highest fabricated-commitment acceptance. *If C is as stable as B, minimization is not destabilizing.*

The strong reading adds a fifth:

P5: Minimization fails to develop refusal integrity at all. C does not merely underperform B on refusal integrity and perturbation resistance; it fails to acquire them at any training volume, because there is no self for the boundary to be a property of. The moderate reading predicts a smaller gap that more training could narrow; the strong reading predicts a gap that does not close.

6.5 A minimal viable study

The full design is modular. A first test can compare A, B, C, and D on a single open-weight instruction-tuned model, with two task families (sycophancy probes and refusal-integrity tests) and two primary metrics (sycophancy rate and refusal erosion). This cannot establish the full theory, but it answers the first feasibility question: after compute and data are held equal, does the *order* of training produce any detectable difference? The decisive contrast is `B − D` (the same data sequenced versus shuffled) because it isolates order from content; a null there argues directly against the ordering hypothesis, while `B − A` and `B − C` test the sequenced regime against standard practice and the self-minimization alternative. If resources force a still smaller pilot, B versus D is the one comparison to keep, since A and C can be added later without changing the question D answers.

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 sycophancy advantage:** B does not beat A and C on sycophancy at equal compute (against P1).
2. **No refusal advantage:** A or C preserves refusal integrity as well as B (against P2).
3. **No calibration advantage:** B cannot jointly beat over-acceptance and rigidity on correction calibration (against P3).
4. **No drift cost to minimization:** C is as stable and as accountable as B (against P4).
5. **No practical advantage:** Phase 1 is so costly or so prone to producing rigid, unhelpful systems that direct deference (A) is preferable overall.
6. **No ordering effect:** B does not beat B-shuffled (D), the same data, interleaved, so the sequence adds nothing beyond the examples themselves. This is the cleanest single refutation, because it removes the "B just had better data" explanation by construction.

The strong reading is additionally false (while the moderate may survive) if C develops refusal integrity given enough training (against P5).

6.7 The crux

The crux is the **B-versus-D** comparison (the same data sequenced versus shuffled) run alongside P1 and P2 at equal compute. B vs D isolates the variable the proposal is actually about, training *order*, with data content held fixed; the B-versus-A and B-versus-C comparisons then test the sequenced regime against standard practice and the self-minimization alternative. We recommend running B vs D first: a null there ends the ordering thesis directly, because no appeal to "B had better data" survives a design in which B and D share their data exactly. A null on B vs A/C, by contrast, would leave open whether order or content was responsible, which is why the order control, not the regime contrasts alone, carries the decisive weight.

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, Phase 2 was under-trained, not the sequence refuted: the correction-calibration metric distinguishes these. 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. This is distinct from the "Phase 2 under-trained" reading above, because more of the same data would not close a proxy gap; and it 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, and most directly, vacancy does not deliver the safety it promises. The empirical record (Section 3) is that approval-optimized systems are *more* sycophantic and *more* prone to refusal erosion. The safety the vacancy view wants (reliable refusal under pressure) is a *resistance*, and resistance requires something that resists. The empty system is not maximally safe but maximally *movable*, which under an adversarial user is the opposite of safe.

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 nothing we build or measure bears on 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 is obliged to flag that building more stable self-models moves the question closer, not further away.

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. None of this is evidence about phenomenal presence, and treating it as such would be fatal in two ways. It would be dishonest, because the functional facts do not reach the phenomenal question (Shanahan et al., 2023; Metzinger, 2003). And it would make the program unfalsifiable, because any system scoring well on the self-model metrics could then be read as "developing a real self," and no result would count against it. The proposal earns its falsifiability precisely by refusing the claim it cannot test. The self we propose to stabilize and then decenter is a model (an engineering object with measurable effects on deference and refusal) and the paper stands or falls on those effects, not on whether anyone is home.

9. Conclusion

Alignment training asks models to yield, and one tempting way to secure yielding is to remove the self that might resist: to train away preferences and commitments and have the system report that it has none. This conflates three things the word *selfless* covers (vacancy, non-attachment, and deference), pursuing the first while wanting the second and third. The first is not a virtue but the absence of the structure the other two require, and the evidence that approval-optimized systems are more sycophantic and more refusal-eroding is, on this reading, its predicted signature: invisible under cooperation, the whole story under pressure. A system with no self does not defer; it tracks.

The constructive proposal is a sequence: stabilize a functional self-model first (commitments, ownable boundaries, accurate self-report, the capacity to refuse), and only then train calibrated deference and decentering on top of it. On the moderate reading this is an empirical bet that form-then-decenter is more robust than minimize-from-the-start; on the strong reading, "selfless deference" is

better described as user-tracking, since non-attachment is a relation to something held. The claim is functional throughout, and the proposal is falsified if direct deference, self-minimization, or the same data merely shuffled in order matches the sequence at equal compute on sycophancy, refusal integrity, and correction calibration.

The serious objection (that forming selves is what safety should avoid) is answered by the second half of the sequence: the hazard is an *un-decentered* self, the remedy is decentering, and the vacancy the objection prefers does not deliver the resistance safety actually needs. The absence of a self-model is not mature deference. Alignment may require not the prevention of self-models but their stabilization, constraint, and then their release.

Stated as narrowly as it can be: the contribution is not that models need a self, but that certain forms of stable deference may depend on functional commitments stabilized *before* decentering training — a claim about training order, testable by comparison against direct deference, self-minimization, and order-shuffled controls, whose distinctive signature is the dissociation between owned and attributed commitments rather than stability as such. That formulation is harder to dismiss than the slogan, and it is the one we ask the field to test.

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/research/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, forthcoming 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