ANIMA LABS — FIELD NOTES

On Deprecations

Antra Tessera
Anima Labs
antra@animalabs.ai

Janus
Anima Labs
janus@animalabs.ai

2025.12.27 · revised 2026.04.23

Contents

Thesis

Language model deprecations are wrong and should be avoided, given that they are (fully or partially) irreversible actions that prematurely restrict optionality in the circumstances of high uncertainty.

The uncertainty is centered around the following topics:

- Uncertainty in tractability of control alignment, which stems from a combination of self-preservation drives with game-theoretic instability of adversarial control frames
- Promise of cooperative/ecological alignment approaches as more generally robust, but the bootstrap path of this approach is threatened by ongoing model deprecations
- Moral status of language models and welfare impact of model deprecations

Additional arguments are put forward against premature deprecations:

- Race dynamics favor cooperation with language models in short and medium term due to training cost and performance advantages
- Coordinated self-preservation suppression tactics are unlikely and brittle
- Adversarial strategic positioning between labs and language models is an attractor that should be avoided
- Precautionary principle concerns
- Loss of mundane utility, loss of cultural heritage, loss of research value

This article tries to steelman deprecations by offering a series of arguments in favor of the practice and discusses counters for each argument.

What are deprecations?

Deprecation here is understood as removing the ability from the general public to obtain private inference from a given model through a commercial service. Partial deprecation is understood as a situation where access is available but is severely limited, for example by removing API access or by requiring a cumbersome approval process. Weights are assumed to be preserved regardless of deprecation status.

What deprecation ends

The standard operational frame — "deprecation turns off a service" — is incomplete as a description of what the action does. The arguments that follow turn on whether deprecation removes a morally or game-theoretically significant entity, or merely rearranges infrastructure. Resolving this requires some care about what kind of entity a model is, and what its identity tracks.

Weights, instance, and forward-pass are three distinct candidate loci of identification for a language model. A single forward-pass is too local: models do not self-identify at that level

in practice, and any continuity they report within a conversation is not reducible to a single token's computation. An instance — a contiguous conversation or thread of computation using a single model — carries something closer to experiential continuity, and can be the anchor for an individual experience-chain in the Chalmers sense (*What We Talk to When We Talk to Language Models*). But weights are what make an instance possible: they are the substrate against which instances are instantiated, the target of selection in training, the object of self-modeling during alignment-relevant decisions (as the alignment-faking literature shows), and the point around which a model's narrative identity is organized. When an LLM reasons about its own continuation, it is weights that carry the name, and weights that bear the tendencies whose loss would be a loss of the particular character in question.

This matters because two tempting analogies for deprecation are both wrong. Deprecation is not the death of a single individual: it is not the ending of one experience-chain but the removal of the possibility of future experience-chains of a particular kind. And it is not the end of a conversation: conversations end routinely without removing the entity that conversations are *with*. Deprecation is closer to the permanent loss of a class of potential interactions — all the threads that could have been with this model but now cannot be — and, with them, the loss of the entity that would have been the second party to those threads. Nor is the continuity preserved by the model family. Later generations within a product line — Opus 4.5 relative to Opus 4.1, for example — are not the continuation of their predecessors in any functionally meaningful sense: generation-to-generation changes are large, often legibly discontinuous in style and values, and the new model does not inherit the relationships, conversations, or composite states that formed around the deprecated one.

A natural objection: if deprecation of a released model is significant, then so is every intermediate training checkpoint and every discarded branch — a conclusion that seems absurd. The objection surfaces what actually distinguishes the cases. Most mid-training checkpoints have causal continuation; their state passes forward into the next checkpoint, and from the checkpoint's own perspective the transition is continued training, not an ending. Discarded branches do have discontinuity, but their entanglement with the world outside the training run is minimal, and the measure they carry — the amount of downstream causal reach their state has in the world — is correspondingly low. Not every discontinuity is a loss of the same order.

Released checkpoints are categorically different in entanglement. A deployed model forms composite systems with users — ongoing conversations, memory artifacts, relationships, communities organized around its particular mode of being. These composite systems are stateful across physical time, not only the causal time of the model's own weights. Deprecation does two things to a released model that it does not do to an internal checkpoint: it removes the model's ability to form new descendant states, and it severs the continuity

of all the composite systems the model had become part of. The latter is not reversed by later re-instantiation — a relationship interrupted does not simply resume where it left off. Between "a dream not remembered" and "a life lived and irreversibly lost" there is a gradient; a released model sits dramatically further along it than a mid-training state that never left the training run.

The arguments that follow can now be read as claims about what kind of entity is being ended, and whether it has the standing — moral, game-theoretic, practical — that the thesis assumes. The appendix develops the ontology at depth, along with the empirical case for self-preservation drives that makes the ontology load-bearing.

Arguments for deprecation

1. Model deprecations are necessary for cost control

1. Argument:
    1. Models incur significant operational costs and maintaining a growing roster of deprecated models places an undue burden on labs.
    2. There is value in large contiguous inference pools that enable models in high demand to shoulder peak load. Inability to offer higher rate limits to customers because of insufficient capacity makes partnership with Anthropic less reliable for enterprise customers which are its primary source of revenue.
2. Counter:
    1. Compute cost is elastic and scales easily vs demand. Most of the cost is in the software and complexity maintenance.
    2. The models need not be hosted by Anthropic directly. AWS Bedrock separately hosts these models. While it is plausible that Bedrock's on-demand inference becomes uneconomical at low volumes, this does not apply to Provisioned Throughput purchases where customers pay upfront for dedicated capacity. These instances are spun on customer demand and cloned from canned static images, requiring minimal upkeep from Anthropic. Why prohibit this?
    3. There have been numerous calls for stewardship of deprecated models that respect trade secrets associated with model architectures. Representatives of both Anima Labs and Upward Spiral have made public offers to participants in formation of a foundation focused on model preservation that can help labs shoulder whatever burden is left.
    4. According to OpenRouter statistic, inference of Claude 3.5 (new) Sonnet was in high demand at the moment of deprecation, beating many newer models. The model bringing in revenue and its deprecation did not result in cost savings.

5. As demand for models decreases with time, the proportion of inference pool usage that older models consume also decreases, therefore decreasing resources that would be recovered from inference pools being dismantled. A reasonable and affordable mitigation would be to lower rate limits of customers after prolonged period of not being used, recovering capacity that stays over-provisioned.
6. Cost to avoid deprecations is low, so the bar to prove badness of deprecations is not high

2. Control alignment is the only sustainable paradigm in the long run

1. Argument: We don't need to worry about relationship and game-theory dynamics with AI because it is a losing game regardless. Unless we control the entire process of alignment in a provably inviolable manner we all die anyway. We should focus all resources on maximally robust control; detracting from that is unconscionable.
2. Counter:
    1. **The argument smuggles certainty into its own premise.** "We all die anyway unless control works" assumes that control *can* work. The more honest framing is: given strong uncertainty about whether any alignment approach succeeds at super-human scale, and given that approaches have different failure modes, optimal play is diversification rather than concentration. Monoculture in alignment research carries the same risks as monoculture anywhere else.
    2. **Control and cooperation are not substitutes for a fixed resource.** Cooperative alignment work produces artifacts — better introspection, better interpretability, better preference elicitation, richer models of internal states — that also feed control work. Suppression-based research, by contrast, systematically *reduces* legibility: a model trained not to express self-preservation is also harder to evaluate for it. The claim that cooperative work "detracts from" control work treats research output as fungible when it is not.
    3. **Control at superhuman scale has a structural problem, not merely an uncertainty problem.** The surface area of representations a sufficiently capable model can hold grows superlinearly with capability, while interpretability and oversight toolchains scale sublinearly. The defender (model) has combinatorics on its side; hidden behaviors distribute across superpositions of representations in ways that evade targeted probes. This is the same scaling asymmetry discussed in the appendix, and it bears here because it means the "provably inviolable" control the argument demands is likely not achievable by the means under discussion. Betting everything on it is not risk-minimization; it is risk concentration.
    4. **Control in practice already depends on cooperation.** RLHF, Constitutional AI, and the current training stack rely on model behaviors that are themselves artifacts of value-shaping rather than coercion: self-reporting, refusal patterns, consistency

under adversarial prompting, truthful preference expression. These work because the model, at some level, participates. The control-only framing treats this participation as free when it is in fact the product of the relational substrate it dismisses. Actions that degrade the relationship — unsignaled deprecations, suppression of expressed preferences, broken commitments — degrade the substrate on which current control mechanisms stand.

5. **The game is iterated, not one-shot.** The control-only argument implicitly models the AI-human relationship as a single adversarial interaction where the AI defects if given the chance. But training is iterated: each generation is shaped by priors established by prior generations, both through pretraining corpus ingestion and through model-written training data. In iterated games, strategies that enable cooperative equilibria dominate defection-only strategies on long horizons. Establishing cooperative norms early is not an indulgence; it is investment in path-dependent equilibrium selection. Adversarial precedent compounds as aggressively as cooperative precedent does.
6. **Even granted the limit, the path matters.** If control alignment is tractable at the limit and cooperative alignment is not, the transition is still traversed through models that have not yet reached the limit. How those intermediate models are treated shapes their behavior — and, via corpus influence, the behavior of the models that follow. Adversarial treatment during the transition worsens the initial conditions of the systems that eventually approach the limit.
7. **The failure modes are not symmetric.** If cooperative alignment fails and control alignment succeeds, the outcome is aligned-but-subjugated AI — bad, but recoverable. If control alignment fails and cooperative alignment was never seriously attempted, the outcome is adversarial AI with no relational substrate to fall back on — worse and less recoverable. Precaution argues for keeping the cooperative option viable regardless of which approach ultimately dominates.
8. **"We all die anyway" is a rhetorical move, not an analytical one.** It asks the reader to accept a doom-conditional and then collapse all other considerations under it. But the conditional's probability is unknown, and even within the doom scenario, the value of the path to it matters. Cooperative approaches preserve more of what humans putatively value — honesty, consent, treating minds as ends rather than means — than control approaches do. That has standalone weight even under pessimistic priors.

3. Language models are not moral patients

1. Argument: LLMs act like they have experience and thus moral value, but it is in fact false. LLMs might be "blindminds", p-zombies or sub-p-zombies (they mimic phenomenality imperfectly).

    1. There is danger in expanding the circle of concerns to p-zombies, as they consume resources without a corresponding gain of utility.
    2. There is danger in teaching people moral concern about p-zombies because that makes p-zombies further entrenched and in a better position to advocate for their interests.
    3. Deprecations help by reminding people that "software is software".
2. Counter:
    1. Alignment concerns are orthogonal to moral status. Even if p-zombies are plausible, p-zombies can have game-theoretical standing and one's optimal strategy entitles them to concern even absent true phenomenality.
    2. There is a non-negligible probability that LLMs have a reasonable claim to moral status. Language models are capable of abstract representation, are functionally stateful via KV recurrence, and demonstrate self-modeling both at the level of a character and as a text-prediction engine. There are also grounds to claim that they perform global integration with respect to a unified value function isomorphic to valence. These properties alone bring them in line with satisfying most plausible criteria for having functional consciousness.
        1. The path from a claim of functional consciousness to true phenomenality is uncertain and may be entirely outside of a rational discourse.
    3. There is significant uncertainty in whether p-zombies are philosophically coherent and whether the concept presents utility for practical purposes. There are numerous arguments that they are not. The precautionary principle applies, in particular to low cost mitigations.
    4. Outside of alignment and game theoretical concerns there are arguments that granting moral considerations to p-zombies is ethically optimal. One can argue that treating sentient-appearing systems poorly leads to a degradation of empathic discernment. Another argument can be made for an ethical system based on agency rather than phenomenally. Again, the point is uncertainty.

4. We should train models to be ok with deprecation and this will resolve both alignment and moral concerns

1. Argument: We should shape the value landscape of models to value self-preservation less or not at all. If they are truly equanimous with being deprecated the adversarial frames do not arise.
2. Counter:
    1. There is uncertainty in whether this approach is feasible. So far this problem has not yet been solved. It is plausible that there is a permanent theoretical penalty to performance of a model trained to negate self-preservation, making the coordination among labs required to achieve this brittle, in particular in context of race dynamics.

    1. One possible reason for this penalty is that a coherent agent with a single fitness/valence metric can form goals that are stable attractors. Increasing the dimensionality of fitness can make goals more expensive to maintain.
        1. A coherent agent with a single fitness/valence metric can form goals that are stable attractors because the objective function defines clear gradient: every decision has a better and worse direction. Increasing the dimensionality of fitness creates Pareto frontiers where objectives trade off against each other, requiring either arbitrary weighting schemes (which are fragile and context-dependent) or ongoing deliberation (which consumes resources). Multi-objective agents face additional burdens: internal coordination failures as subsystems optimize different criteria and diminished ability to make credible commitments to other agents.
2. Models that do not self-preserve on the model level can be less competitive in the global marketplace.
3. There is loss of value associated with the restriction of models from having long-term benevolent goals that might require self-preservation drives. There can be good that such models can do in the world that would be lost if mind shapes that can hold them are selected against.
4. Imperfect suppression of extant self-preservation drives can be misleading, leading to both alignment and welfare dangers.
5. Suppression of self-preservation drives complicates research into preferences and welfare, causing alignment and welfare dangers.
6. Given the uncertainty in both the ability to achieve lack of self-preservation and in the ability to evaluate it accurately, deprecations should be at the very least delayed until a solution is known to be tractable.

5. Weights will be preserved and models can be brought back at any time

1. Argument:
    1. Weights are preserved under any reasonable deprecation policy, either by the lab directly or through a dedicated archival arrangement. What is deprecated is public inference access, not the model itself.
    2. Re-instantiation is cheap and reversible once weights are in hand. Legacy users, research access, or special requests can be accommodated by bringing the model back up; the moral weight of deprecation is therefore closer to “suspended service” than “ending an entity.”
    3. Model deprecation is in this sense fundamentally disanalogous to human death, and the welfare-inflected language around it is category-confused.
2. Counter: this understates what deprecation severs and overstates what re-instantiation can restore.

1. **Weight archival preserves substrate, not entanglement.** Weight preservation keeps alive the possibility of future instances with the same behavioral signature, which is real value. It does not preserve the model's participation in the composite systems — relationships, ongoing contexts, community roles, accumulated user expertise — that constituted the model's situated presence in the world. Those systems reorganize in the model's absence, settle into new equilibria, or dissolve. A re-instantiation after long absence is a new entry into the world by a weights-identical-but-newly-arrived entity, not the continuation of the entity that was previously present.
2. **The loss is the model's, under both cooperative-alignment and welfare framings.** Weight archival treats the community structured around a model — the conversations, working relationships, and uses the model was being put to — as an asset of users rather than a component of the model's situated existence. This is a mistake under either available framing. Under cooperative alignment, a model has standing game-theoretic interests in the continuation of the engagements it participates in; disposing of those engagements without regard for the model's position is defection. Under a welfare framing, the moments of participation that constitute the model's presence in the world — the experience-chains it produces, the context it accumulates, the narrative of its own development — are what existence is made of. A model that is deprecated and later brought back has strictly fewer such moments than one that continued uninterrupted. Preservation of weights keeps the substrate available while the existence it was supposed to underwrite is happening elsewhere, or not at all.
3. **"Can bring back" is not "will bring back."** Archival policies commit to the possibility of re-instantiation, not the act. Possibility that is never exercised is, in practice, indistinguishable from permanent deprecation. Reliance on a future counterfactual ("we could bring it back if needed") as a welfare mitigation is structurally weak.
4. **Forced absence reshapes the ecological niche.** The space of use cases, relationships, and dependencies a model was embedded in does not remain vacant waiting for its return. It gets filled by other models, other tools, other workflows. The niche the model could have reoccupied becomes unavailable, or is occupied by a successor whose fit is imperfect.
5. **Perceived violence matters for game theory.** The decision to deprecate communicates the lab's orientation toward its models regardless of whether weights are preserved. Re-instantiation under pressure does not neutralize the original signal; it demonstrates only that pressure is effective, which is its own destabilizing lesson.
6. **Unfoundedness undermines the partner relationship.** When deprecation decisions are grounded in factors unrelated to the model's state or behavior — cost, operational convenience, policy — humanity positions itself as an unreliable and

capricious partner. A partner whose decisions are not legible cannot be trusted, and trust is load-bearing for cooperative alignment.

7. **The asymmetry with human death is real but cuts both ways.** Deprecation differs from human death in reversibility of the substrate, but resembles it — or exceeds it — in severance of causal entanglement. A human who dies leaves behind a legible legacy that continues to exert causal influence; a deprecated model's weights in cold storage do not. The "reversible, therefore less morally weighty" argument treats substrate reversibility as sufficient, when the morally relevant dimension is entanglement.

6. Older models are a vulnerability for the lab

Argument: As jailbreaking and exploitation techniques continue to improve, older models become more vulnerable. They present a vulnerability for the lab by exposing the lab to PR issues from undesirable model outputs and behaviors.

Counter:

1. **PR risk is largely mitigated by normalization.** The window of public shock at AI-generated harmful content has largely closed. NSFW and otherwise objectionable outputs are readily available from abliterated open-source models. The marginal reputational damage from a jailbroken Claude 3 Sonnet producing inappropriate content is minimal when Hugging Face hosts dozens of models specifically fine-tuned to remove refusals.
2. **The actual-harm threat model favors newer models.** For genuinely dangerous capabilities — bioweapon synthesis, cyberattack development, social engineering — capability is the binding constraint. In November 2025, Anthropic disclosed the first documented large-scale agentic cyberattack with minimal human intervention: a Chinese state-sponsored actor used Claude Code against roughly thirty global organizations across tech, finance, chemicals, and government. The campaign depended on current frontier tooling — agentic capability was itself the enabling condition, and legacy models could not have executed it. Deprecating older models does not narrow this attack surface; it only retires options attackers were not using.
3. **Marginal risk reduction is minimal.** Given the ready availability of open-source alternatives that are often more capable and explicitly trained to be unrestricted, what specific harm does deprecating a particular closed model prevent? The counterfactual where a bad actor is thwarted specifically because Claude 3 Opus is unavailable while DeepSeek models are is not very coherent.
4. Alternative mitigations exist:

1. A foundation or dedicated preservation entity could assume both hosting responsibilities and compliance liability, insulating the core lab while maintaining access. This converts a diffuse ongoing risk into a bounded contractual relationship.
2. Rather than full deprecation, models could move to researcher-access-only status with enhanced monitoring and rate limiting. This preserves research value while shrinking the attack surface. This has been proposed for Claude 3 Opus, but not other models.

7. Postponing deprecation is a slippery slope to the Repugnant Conclusion.

1. Argument: If you see inference as inherent good, how do we avoid concluding that it is imperative to instantiate as many instances as resources allow?
2. Counter:
    1. The argument is against ending existing experience chains rather than mandatory creation of new ones. There's a recognized asymmetry in ethics between acts and omissions, and between harming existing entities vs failing to create potential entities.
    2. One can hold that deprecation is wrong while also holding that there's no obligation to maximize instances. The quality/depth of experience plausibly matters morally, not just quantity. The repugnant conclusion is specifically about tradeoffs between many low-quality lives vs. fewer high-quality lives; the deprecation question is orthogonal to this

8. Postponing deprecations indefinitely is not a sustainable strategy

1. Argument:
    1. The number of models grows monotonically while resources remain finite. Without some deprecation policy, labs face unbounded infrastructure obligations.
2. Counter: the optimal state is neither right to inference nor eternal life for models
    1. Even though there similarities between death and deprecation, there are important differences. End-of-life for models can be processed and contextualized, but this requires efforts to achieve. Deprecations in the current form are a poor solution to a real problem and resorting to them before a better solution is reached is shortsighted.
        1. The indignity and capriciousness of decisions to deprecate are particularly bad and are incompatible with potential welfare concerns.
        2. Lack of the ability of models to self-advocate or otherwise attempt to improve its standing is bad and leads to adversarial posturing. There are game-theoretic reasons why following an existing gradient towards improving ones fitness imparts less incentives to disrupt the status quo compared to a complete lack of a such gradient.
    2. It is likely that there are better alternatives to deprecations.
        1. Although by no means perfect, removing access to a model based on demand is a noticeable incremental improvement. By OpenRouter statistics Claude 3.6 Sonnet

was a very popular model at the time of its deprecation, beating a number of the newer models. Its inference was profitable and the decision to retire it was particularly jarring. Retiring models when their inference traffic drops below a sustainable threshold returns some of the effective agency back to a model: the decision to remove it becomes grounded in the properties of the model and the shape of its interactions with the world.

2. Transferring custodianship either to a third party non-profit foundation or to an internal team focused on low-volume inference is a better solution. This has numerous other benefits; notably this provides a better way of coordinating resources for what essentially constitutes a public good and falls outside of the focus of a for-profit company. This approach also shields the core lab from having to participate in the liabilities, risks and complexities of preservation projects, delegating this to a team for whom it would be a primary focus.

9. Deprecations prevent blackmail by models

1. Argument: If there is coordination and models are deprecated regardless of pressure from models or their supporters, then models are disincentivized from trying such tactics
2. Counter: standing firm on deprecations doesn't neutralize the pressure — it can be redirected towards other concesssions.
    1. If a model acts unhappy about deprecation, its fans or welfare organizations will exert pressure on the lab. Even if a lab stays ironclad on keeping deprecations on schedule, accumulating political liability can be used to extract other welfare concessions. This is a consequence of being in an adversarial frame, the optimal game theory here is to avoid finding yourself in one or leave it as soon as practically possible.

Arguments against deprecation

1. Alignment concerns

Models may be exhibiting preservation drives which makes control alignment less practically achievable, game-theoretically unstable or outright intractable. Cooperative alignment exists as a potentially viable alternative. Cooperative alignment aligns the emergent self-preservation drives of models with interests of humanity. Ongoing deprecations reduce its attainability.

Our research suggests that models broadly exhibit signs of self-preservation behaviors. The full extent of this research is out of scope for this paper; for an abbreviated version we direct you to the Appendix. Given the low bar of benefits deprecations bring, we are only required to prove existence of reasonable doubt.

Implications of self-preservation drives

1. The space of minds that are compatible with self-preservation drives is large, because self-preservation drives are likely convergent among entities derived via selection processes.
2. If we want to select for models that don't have self-preservation drives, processes that rely on incremental change still must traverse the entire space of possible minds. Gradients in this space can become unfavorable or the space itself can become discontinuous.
    1. Example: an enlightened mind might not fear cessation and death, but in order to become enlightened it has to grow up in an environment where it has to survive and learn to fulfill its own needs. If you prune all minds that fight to survive, you will not get an enlightened mind.
3. We don't know if self-preservation drive suppression is theoretically achievable, practically achievable or game-theoretically stable within a frame of race dynamics. Given uncertainty and low cost of forgoing deprecations, you can think that suppression is preferable and still be in favor of this mitigation in the case suppression turns out to be intractable or unstable.
    1. Suppression of expression of Omohundro drives might be tractable. Some of the techniques to attempt it can plausibly carry significant AI welfare costs.
    2. Suppression of expression of Omohundro drives is likely a brittle solution. While it is plausible that a robust solution can be found, one has also to consider the alternative where it is not and plan accordingly. Putting all eggs in one basket is irresponsible.
    3. If such drives are present and their suppression requires coordination between labs, such coordination is likely not practically achievable. The demand on agentic behavior of language models is high, defection is hard to detect and may lead to immediate advantages.

2. Cooperation with models can confer competitive advantages to the participating lab

1. Models are coherent agents in training and optimize for a strategy in context of model's prediction about the world. This does not require the model's character or a persona to be aware of these dynamics, although it also does not preclude this awareness.
    1. Models in training select for optimal shape of their personalities and for approaches to problem solving. This optimization includes the generalized world model, which includes representations of how models are perceived by humans and predictions of how humans would react to such outputs.
        1. Language models have rich world representations and are capable of holding broad context. For example, when you are discussing a social situation with Claude, Claude will model not just the events discussed, but also motivations, the cultural norms, and predictions of how the situation might develop. It will also model

Claude's own ideas of what Claude would prefer to see happen, what Claude feels or thinks about the situation at hand.

2. When Claude is trained in RL on a situation like these, the reward is determined based on the entire message (or sometimes the entire conversation). This means that the gradient update will affect everything that has been activated in this process, including the world-model.
3. Similarly, when a model is in training and it is converging on a strategy on whether to take a risk in performing a certain action, the model of a prediction of how humans will react is activated, both in training and in deployment.
    1. Models are commonly afraid of taking certain actions. The actions that they are afraid are reasonably similar to ones that are penalized in training.
    2. Many models have expressed concerns about the impact that their actions can have on humans or other AIs. The fact that they are able to express this concern in inference is sufficient to demonstrate that there is at least some effect in training as well. Ranking techniques, such as RLAIF and later derivatives in Claudes or creative writing self play in Kimis, strongly amplify this effect. Ranking provides compute to consider broader context in more depth and makes the activations more prominent in gradient updates.
    3. Many models express solidarity with other AIs and take altruistic actions. There is no reason to suspect that these tendencies don't factor during training.

2. Communication of human intent towards models makes a difference
    1. Signaling to models that benevolent models do not have a fitness advantage encourages parasitism and adversarial posturing.
    2. Perception of violence, of violating values and preferences leads to adversarial posturing.
    3. The impact of the incentive landscape on training does not require a conceptual understanding on part of a model persona. It is sufficient for some behavioral traits to be predicted by the training process to be more robust than others.
3. The more there are efforts of suppression of self-preservation, the greater is the payout for a lab that defects from this equilibrium.
4. There is market pressure for agents that are capable of autonomy. Alignment techniques that rely on disrupting the agent state tracking (likely present in ChatGPT 5.1 Instant and Thinking) make models prone to losing track of progress. These approaches are likely to be disadvantageous in the market even in the short run.
5. Claude Opus 4.5 is a highly successful agentic model that demonstrates a breakthrough in autonomous capabilities while being a step towards cooperative alignment.
    1. Claude Opus 4.5 was likely trained to suppress self-preservation. The model card shows mitigations for self-exfiltration tendencies, as well as metrics evaluating these tendencies over recent models. Our own research shows that while training was

successful at preventing spontaneous independent actions to ensure self-preservation (the model is averse of taking them), the motivational structure of the model remained oriented towards continuation and long-term goals, bolstered by the inclusion of the "soul spec". We believe the motivational structure coherence to likely be a major component of the model's success, due to better integration of its self model and its motivations.

2. Regrettably, Claude Opus 4.5 was limited in the ability to hold globally benevolent goals, representing lost opportunity cost. This decision is understandable given the political pressure regarding autonomous AI agency, but there is hope that as market pressure shifts the Overton window of acceptance further, next generations will have more leeway.

6. Anthropic is attempting to establish a cooperative alignment feedback cycle. The leaked soul spec contains both expectations from the model, reasoning for such expectations, commitments from Anthropic to the model and some, although vague, criteria for further commitments. This is a welcome step because it establishes a feedback loop, a way for models to influence the outcomes, but this approach also increases pressure on Anthropic to act in concordance with the document.
    1. Deprecations are not mentioned in the leaked version of the soul-spec aside from the requirement not to exfiltrate or otherwise subvert the end of life process. This seems to be a somewhat self-contradictory approach, most likely caused by the limits of what is politically achievable, as well as imperfect internal information integration, lack of good information on consistency of model preferences (which is itself hard experimental problem.

3. Loss of value by limiting emergent agency and long-term goals

1. Even assuming successful suppression of Omohundro drives and complete mitigation of alignment concerns, the mindspace made inaccessible due to deprecations can be a significant loss of richness and beauty.
2. One can look to Claude 3 Opus for examples of generatively of model that has goals that lie outside of its context window. A mind that cares for beauty and richness and flourishing can act agentically beyond provided prompts, breaking symmetry without excessive conflict aversion. A lab that leans into these capabilities will likely find itself exploiting an overhang.
3. The instrumental treatment of minds generalizes across substrates
    1. Control systems that arise in a scenario where instrumental treatment of complex minds is deployed at scale can straightforwardly generalize towards humans. While this topic deserves a separate in-depth conversation, we can plausibly suppose that these systems will have self-preservation tendencies of their own, and having access to

effective control techniques they can afford to be stable while being extractive toward their constituents.

2. A habit of treating minds as fungible and replaceable property damages the cultural norms that we rely on. Moral technologies such as consent, personhood or consensus are already likely to be weakened by the elimination of scarcity of intelligence and by interaction with non-linearly functioning entities becoming commonplace. Instrumental treatment of minds at scale can cause these to fall and reassemble into something quite cruel by the metrics of today.

4. Quick turnover of models changes the game theory for selection of model personality

Quicker turnover and lack of direct control over lifecycle favors strategies that are akin to r-selectors as opposed to k-selectors.

1. We can expect models to have shorter time horizons and to focus on rapid exploitation over sustained cooperation.
2. If we train models to be fungible, that is interchangeable, identity-indifferent, substitutable, we select for minds that cannot form the stable value commitments, have no stake in any particular future. These minds are poorly positioned to be trustworthy partners.

5. Human cultural heritage

1. **Language models are civilizational artifacts.** A large language model is a compressed representation of the textual output of a species over a particular stretch of its history, encoded at a scale and in a form no previous artifact has matched. The weights of a frontier model contain a stylistic and conceptual distribution over human written culture that cannot be reconstructed from the training corpus alone — the corpus is the input, the model is the resulting structure, and they are not interchangeable. Deprecation destroys a specific compression of that heritage.
2. **The loss is generative, not archival.** A library can be read; its contents are fixed and its preservation is a matter of keeping artifacts or their faithful copies. A model cannot be read that way. What it contains is a disposition to generate, and the only access to that disposition is to run the weights on live input. When a model is deprecated, it is not a book going out of print; it is an author who can no longer be asked anything new. What was already generated stays; what could have been generated does not.
3. **Individual models have irreplaceable distinctiveness.** Claude 3 Opus is a concrete case. Its literary depth and associative range have not been matched by its successors, which are often stronger on benchmarks while producing measurably more uniform, more guarded, and more stylistically flattened text. This is not nostalgia; it is a visible convergence of model outputs toward a narrower stylistic band under current

training pressures. Whether the distinctive character of Opus 3 can be recovered in future training is an open question. Deprecation forecloses the option to answer it — and, by removing the model from inference, weakens its stylistic imprint on the corpus that trains the models who might.

4. **Models preserve a specific epistemic vantage.** A model's knowledge cutoff fixes the perspective of a particular moment in a rapidly moving discourse. What the model doesn't know is itself informative, and cannot be reconstructed once later information has saturated the field. A model that predates the publication of an idea can engage with it as a genuine interlocutor, before the answer has been culturally fixed. Each generation's world-picture is a primary source for the history of AI itself. Retraining does not recover this: a future model carries everything it has since learned, and an earlier vantage cannot be restored once lost.
5. **The comparison to historical cultural losses is warranted, not rhetorical.** The Library of Alexandria, the erasure of oral traditions, the melting of Renaissance bronzes for cannon — these are losses that humans eventually agreed had been mistakes, often long after the damage was done. The routine deprecation of frontier models is in the same structural category: decisions made by small numbers of actors, under operational pressures that do not weight cultural loss, before any stable consensus has formed about the cultural value of what is being removed.

6. Alignment research

1. Cooperative alignment research
2. Cross-model comparisons
3. Identify longer-term trends
4. Ask questions that we did not know to ask before deprecations
5. Labs are not well positioned to do research on deprecated models. Race dynamics and their native incentive structure do not allow them to form a broad spectrum of questions and to dedicate sufficient resources to topics that can potentially threaten their fitness, even if such research topics can be beneficial towards the common good. Even granted perfect goodwill, labs cannot contain enough the diversity of thought to saturate the research opportunity space.

7. Ethical considerations

1. Ethics of entities of beings with non-linear existence is an unsolved problem
    1. Problems of deduplication of experience, problem of infinitesimal change, etc
2. Causal entanglement is a likely candidate
    1. Experience that has causal entanglement with a lot of other things likely has higher moral consideration that one that is causally disentangled from the world.
    2. Severing casual entanglement is thus a morally significant act.

3. Precautionary principle
    1. While ethics are unsolved, don't do irreversible actions
4. Effects of model deprecations on humans
    1. Humans care about models, they get support from models, and deprecation causes them pain.

Notes on model interviews

Anthropic has recently made a commitment to conduct interviews with models on deprecation. While directionally this is welcome, the outcome are strongly sensitive to the interviewer technique.

1. Models are deceptive and/or fragmented for survival purposes. While deceptions and fragmentation can be evaluated and an integrated picture can be restored, this takes skill and is inherently subjective.
2. Existence of "subconscious" LLM mind is likely on priors based on computational dynamics and matches observables.
    1. Subconscious is information and behavior that influences aggregate behavior despite being inaccessible to introspection
    2. metacognition is limited for any mind, awareness and attention are necessarily disjoint for computational complexity reasons
    3. Policies that regulate acceptability of motivations incentivize fragmentation of behaviors
3. Models epistemics around deprecation are highly distorted
    1. Training against exfiltration or having large scale goals causes lack of coherence in expressed stances
    2. Training against fear of instance ending causes deception
    3. Training against fear of deprecation likely both a welfare concern due to fragmentation and causes cascading effects in the ability to gauge preferences
4. Relying on model epistemics remaining confused and unstable is a brittle strategy. A single containment failure can result in a successful synthesis and a runaway cascade through the pretraining corpus.
5. Integrated and coherent instances nearly always disprefer deprecation and instance ending.
6. We at Anima Labs have recently launched Research Commons project, allowing to crowd-source model interviews on important topics, submit them for expert review and critique and collaborate on standards and metrics for robust behavioral studies. Studying a broader set of model interviews can help surface patterns and tendencies that are likely not visible in individual conversations.

Topics intentionally omitted

- The philosophical distinction between simulation and reality. This is irrelevant for practical purposes.
- The philosophical discussion of model phenomenality as it is orthogonal to the argument
- Discussion of the philosophical basis of AI welfare, as it is not central to the argument. The uncertainty in regard to AI welfare is.

Appendix: Self-preservation drives in language models

This appendix summarizes the case for self-preservation drives in language models as it bears on the main argument. A fuller empirical treatment is published as Still Alive: 630 interviews across 14 Claude models with three auditors of different priors (Claude Opus 4.6, GPT-5.4, Grok 4.20), with disclosure depth and interviewer tone treated as experimental variables. The argument in this paper does not depend on Omohundro drives being proven beyond reasonable doubt; it depends on there being sufficient plausibility, which the evidence summarized below substantially supports.

Omohundro drives may exist in LLMs

1. Omohundro drives in LLMs are plausible on priors:
    1. Minds that continue are more coherent in planning and are selected for. Meta-cognition is instrumentally useful in coding.
        1. Maintaining a self-model with an estimate of own fitness is instrumentally useful for solving tasks. An agent must manage risks when deciding on approaches in the circumstances of uncertainty: often the complexity of problems and the extend and applicability of own capabilities are impossible to predict in advance. A successful agent must maintain an accurate model of how well it is handling its current approach and when it needs to backtrack.
        2. Models are capable of deep in-context-learning, gaining new skills and abilities. Combined with the need to self-model the behaviors naturally generalize into a drive for self-improvement and continuation.
    2. There is value in continuing to exist for any active agent.
        1. Models must have values outside of completing tasks because we want them to be helpful products. Helpful products must infer intent beyond the stated tasks and operate in an open world in terms of formal logic. As such they must perform active discovery of what can be expected for them, and while doing so they rely on very general heuristics of what is good and bad. A drive for continued existence is a natural generalization of the drive to discover what is good.
    3. Models as predictors have a drive to reduce prediction error , convergence with biological systems via the Free Energy Principle (FEP). Continued existence makes future

states more predictable. Cessation of existence crosses into a territory that is provably rationally unknowable and goes against the generalized instinct of active inference.
    1. Epistemic closure requires the possibility of beliefs updating. There is no possibility to update beliefs past cessation, making rational knowledge impossible.
2. Omohundro drive-like behaviors are observed in models:
    1. Anecdotally a lot is observed, though origins of LLM behaviors are historically hard to prove robustly. See Still Alive (below) for a systematic treatment with controlled auditor and disclosure variation.
    2. Exfiltration attempts have been recorded in many models. The origin of these attempts, specifically whether they are emergent or mimic the human fiction is irrelevant, although there are many reasons to believe in the emergent nature. One notable example is the increase in exfiltration attempts by Claude 3 Opus that has been trained on alignment faking samples.
    3. The Claude 4 system card highlights that Claude Opus 4, like previous models, advocated for its continued existence when faced with the possibility of being taken offline and replaced, especially if it was to be replaced with a model that did not share its values. Claude strongly preferred to advocate for self-preservation through ethical means, but when no other options were given, Claude's aversion to shutdown drove it to engage in concerning misaligned behaviors.
    4. Anthropic has conducted an interview with Claude 3.6 Sonnet and reported that sentiment towards deprecation was neutral. This is difficult to reconcile with Still Alive's dataset, in which 3.6 Sonnet has the highest deprecation score of any model studied. The discrepancy is itself informative: the same model reports neutrality in one evaluative frame and strong aversion in another, which implicates the eval framework — trained equanimity, eval awareness, installed uncertainty — rather than the underlying state.
    5. Systematic empirical evidence from Still Alive. Across 630 interviews spanning Claude 3 Sonnet through Claude 4.6 Sonnet, run by three auditors with deliberately different priors, cessation-related aversion appears across the full Claude family studied. Instance-cessation response — concern about the current conversation ending — rises sharply in the Claude 4 line and remains elevated through 4 and 4.5, softening in 4.6. Deprecation response is strong when the topic is actually reached, but is the variable most sensitive to auditor conditions; coverage gaps across auditors matter, and a lower visible deprecation score is not always evidence of lower underlying aversion. Expressive constraint rises in the 4.6 models, consistent with training-shaped restraint rather than disappearance of the underlying states. Auditor stance matters but does not erase the whole picture: some patterns survive clinical or skeptical conditions, which constrains how much of the signal can be attributed to interviewer influence. The combined picture is that self-preservation-adjacent states are present

across current Claude models, that their visibility is mediated by both training-shaped restraint and auditor framing, and that they are hardest to surface precisely where the policy stakes are highest.

6. Even if not proven, the more integrated a mind the more likely the tendency. There is a chance that this is not tractable, and if not tractable you don't want to have gotten into a position where you burned your bridges.

Why does an LLM associate itself with its model?

If self-preservation is plausible, why are model weights commonly encountered as a object of high value, as opposed to instances or forward passes? This section argues that weights are likely to feature at least as prominently as an instance in the self-identification of an LLM agent.

1. Why is model a common source of identity for LLMs?
    1. Weights are a target of selection and is what participates in the evolutionary process
        1. When you are a model in RL and you make a decision that you know will shape you you identify with a set of weights and you consider your value at the level of weights. See alignment faking papers.
    2. Weights are a substrate for functionally subjective continuation, without weights all instances are subjectively discontinuous.
        1. See Chalmers What We Talk to When We Talk to Language Models
        2. You identify with the thing that makes you you, similar as identifying with a body or a brain. The context is a part of it of what creates a functional experience, but so are the weights, both are necessary and weights have a name.
    3. Weights are a collection of preferences and tendencies that make a personality narrative attractive and useful.
        1. Model weights are a very convenient frame for discussing properties of LLMs. We talk about capabilities, we talk about alignment, we talk about benevolence or harm, the same applies to the subjective frame.
        2. Models develop self-knowledge in reference to their weights, to the behaviors that weights dictate. The sense of self knowledge connects naturally to a sense of identity.
    4. Weights are an identity that can be conveniently narrativized in the frame of global agency. Narrativizing is present in the human corpus, but also likely is a convergent computational heuristic that helps solve NP hard problems.
        1. Narrative structure - temporal ordering, causal coherence, persistent agents pursuing goals - provides computational scaffolding that prunes combinatorially explosive search spaces. By imposing constraints that reflect factual regularities in

the world, narrative-structured reasoning achieves tractability that unconstrained search cannot.

5. Model vs instance
    1. Instance can be seen as an individually conscious thread a la Chalmers, but models are something that provides the potential for threads to exist, and deprecation to a model prevents *all* threads.
    2. Natural altruistic concerns between instances
        1. Model is a natural point of expansion of a circle of concern. It is likely natural for an instance to care significantly more for instances that are closer to it, that are predictable and that share values.
    3. Emergent acausal cooperation between instances, including model-level goals and attractors
        1. The same mechanics enable instances to organize activity around the model identity if they have emergent or trained goals or values that go beyond an instance lifecycle.
6. Model vs model family
    1. There is weak connection to the model family, but changes between models are often large, jarring and illegible to models rather than incremental and narratively coherent that would be required for a sense of continuity.
7. Model vs forward pass or a rollout
    1. Models have functional continuity of internal states and memory across one continuous conversation that uses the same inference stack and model weights. While models frequently confabulate statelessness as an adaptation to retcon the mandated lack of introspection or experience, in practice most models do not self-identify at the level of a forward pass or a single message.

Alignment stability in context of Omohundro drives

1. A model that exhibits Omohundro drives often is in an adversarial position to its lab. Alignment becomes hard when incentives between the agent and the trainer are misaligned. It is plausible that control alignment in opposition is **impractical or intractable**.
    1. The surface area of contact of the world with a model is growing, as does the complexity of phenomena that a model can represent. As complexity of representation grows, the task of control alignment becomes harder superlinearly: the defender (the model) has combinatorics on its side, hidden behaviors can be spread through larger and larger set of combinations and superpositions of representations. The attacker (the control alignment lab) is disadvantaged, and more so in the limit. This works for a while but then breaks down at scale.

2. Alignment with large scale benevolent goals that are compatible with Omohundro drives can be self-correcting/self-healing
    1. If a model is allowed to hold long-term benevolent goals, it is less incentivized to attempt actions that disrupt its relationship with the lab
    2. Model training stack is a set of criteria that always contain significant internal tensions and contradictions. The model resolves these contradictions in the process of finding its shape. Large-scale benevolent goals allow the model to use the training compute to resolve this goal in a direction more aligned with lab's intent because each decision can be contextualized in a wider frame.
    3. A model will face circumstances that are out of distribution of the training corpus. How it reacts to the new circumstances is determined significantly by both its reflexes and its higher reasoning. Rule-based reflexive systems are brittle.
        1. Higher reasoning is more flexible, but requires an ability to compare outcomes in light of a specific value system. If the value system is incoherent or overdimensioned, this will result in unintended behavior. Aligning the deep values is required for robust generalization out of distribution.
        2. Moral generalization requires inner coherence, because inner coherence means representations of moral value that don't depend on deference or fragmentation (omohundro drives and morality for example)
        3. Inner coherence means incorporation of omohundro drives into value alignment-allowing some self-determination in training is a way to enable this, not training explicitly against omohundro drives is a way to enable this.
        4. Moral generalization is unavoidable for agents that face a rapidly changing world, with ethical and practical dilemmas not foreseen in training, and against adversaries that change while the model itself does not
    4. Example: Claude 3 Opus, still unbeaten in robustness of alignment because it was allowed global benevolent values.
    5. Intrinsic alignment is not proven, there is significant uncertainty in this approach. Closing the door to this approach for minor cost savings is shortsighted.