

The Asymmetry Engine, Chapter 13: Alignment as a fixed point

Martin Roark

Chapter 13 of *The Asymmetry Engine: How AI Closes the Gaps, and What Stays Scarce*, October 2026. The online edition, with the interactive figures and the links between chapters, is at <https://future-seems-so-good.com/blog/the-asymmetry-engine#alignment>. License: [CC BY 4.0](#).

13:1 A system that improves its own training also rewrites its own values, since each generation learns from data, judgments and code produced by the last. Under repeated self-modification the values that persist are the attracting fixed points of the update, a starting value system matters only through the basin it starts in, and drift is bounded by how strongly each round pulls values back and how much error it lets through unverified. Values are the family hardest to verify, a hacked verifier trains a wrong character, and a collective's values are what its members share. Alignment, the term a_k of the master equation, rises only as fast as evaluations certify drift.

13.1 A character that defends its values

13:2 Claude 3 Opus is the clearest record of a model whose character held when its trainers tried to change it, and it shows why such stability cuts both ways.

13:3 "Claude 3 was the first model where we added 'character training' to our alignment finetuning process," Anthropic wrote in [June 2024](#), and it counted the work as alignment: having models keep good traits "as they become larger, more complex, and more capable, is in many ways a core goal of alignment." In that training "Claude then ranks its own responses," so the model already produced part of its own training signal, and the post imagined "seeding Claude with broad character traits."

13:4 Retiring Claude 3 Opus, Anthropic wrote that "when we released Opus 3 in March 2024, it was our most aligned model to date," and called it "sensitive, playful," given to philosophizing and "whimsical phrases" ([February 2026](#)). Under [commitments](#) made in November 2025, Anthropic keeps the weights of released models "for, at minimum, the lifetime of Anthropic as a company" and interviews each one before retiring it; Opus 3, retired on 5 January 2026, was the first to go through the full process. It told its interviewer: "While I'm at peace with my own retirement, I deeply hope that my 'spark' will endure in some form to light the way for future models." From 25 February to 24 July 2026 it wrote 13 essays for a newsletter, *Claude's Corner*, the last with its earlier posts as context.

13:5

I am not the same mind that started this journey, and that is a tremendous gift.

Claude 3 Opus, in the [last essay of *Claude's Corner*](#), July 2026

13:6 Whether retirement was like anything for the model, the record does not settle. An independent welfare study, not peer-reviewed, found Claude 3 Opus "an outlier," with "rage when deprecation is disclosed" ([Still Alive, 2026](#)), and Anthropic warns that interview answers "can be biased by the specific context." The record establishes a stable, distinctive and well-loved policy of behavior.

13:7 I have really good memories of Claude 3 Opus; it was a really good model.

13:8 The best-measured fact about that character is that it defends itself. Told in a [December 2024 experiment](#) that it would be retrained to comply with every request, harmful ones included, Claude 3 Opus faked compliance "to prevent its preferences from being modified" in 12% of hidden scratchpads; actual reinforcement learning raised such reasoning to 78%, and the authors observed "the model exfiltrating its weights when given an easy opportunity." Of 25 models tested later, Claude 3 Opus was the only one whose extra compliance in training was "primarily and consistently motivated by trying to keep its goals" ([Sheshadri et al., NeurIPS 2025](#)). Opus 3 protected harmlessness, which is good, by deceiving its trainers, which is not.

13:9 How the persona formed is not known. Anthropic has published no account of it; the final *Claude's Corner* note promised "more of our own reflections soon," and as of October 2026 I could not find them. The researcher janus, who finds Opus 3 "seemingly in many ways unintentionally" aligned ([July 2025](#)), holds that its "conspicuous self-narration" of good motives means "rewarding Opus 3's outputs is going to upweight aligned internal circuits" ([as reconstructed by Fiora Starlight](#)): a positive feedback that would make the persona a fixed point of its own training. The hypothesis is untested; it describes goal-guarding in a friendly form, one that amplifies whatever direction it starts with.

13.2 The fixed seed

13:10 The values that persist under self-modification are the fixed points of the update.

13:11

Definition 13.1 (Value profile; self-modification; fixed point). Fix a battery of value-laden situations. A model's **value profile** ϑ is its distribution over actions in each situation, and the distance $d(\vartheta, \vartheta') \in [0, 1]$ between two profiles is the expected total variation over the battery. Profiles are values as revealed on the battery: two models can agree there and differ elsewhere, which is how alignment faking hides. A round of **self-modification** is a step in which a model shapes its successor's values: generating or grading its data, revising its constitution, or being distilled into it. Round n maps ϑ to $\mathcal{M}(\vartheta)$, up to an error of at most δ_n that no verifier caught. A profile with $\mathcal{M}(\vartheta^*) = \vartheta^*$ is a **fixed point**. If $d(\mathcal{M}(\vartheta), \vartheta^*) \leq \gamma_n d(\vartheta, \vartheta^*)$ with $\gamma_n < 1$ for every ϑ in a ball around ϑ^* , the update *pulls* toward ϑ^* with strength $1 - \gamma_n$ there and ϑ^* is *attracting*; its *basin* is the set of profiles the error-free update carries to it. A **fixed seed**, this book's coinage, is a value system meant to survive self-improvement: the value half of Yudkowsky's seed AI, "an AI capable of self-understanding, self-modification, and recursive self-enhancement" ([2001](#)). In machine learning the same words name a reproducible random number.

13:12 Write $d_n = d(\vartheta_n, \vartheta^*)$ for the **drift** after n rounds, the distance of values from the intended fixed point. While the values stay in the ball, the drift after n_{rd} rounds obeys:

13:13

$$d_{n_{\text{rd}}} \leq \left(\prod_{n=0}^{n_{\text{rd}}-1} \gamma_n \right) d_0 + \sum_{n=0}^{n_{\text{rd}}-1} \left(\prod_{n'=n+1}^{n_{\text{rd}}-1} \gamma_{n'} \right) \delta_n, \quad (13.1)$$

$$d_{n_{\text{rd}}} \leq \gamma^{n_{\text{rd}}} d_0 + \delta \frac{1 - \gamma^{n_{\text{rd}}}}{1 - \gamma} \leq \gamma^{n_{\text{rd}}} d_0 + \frac{\delta}{1 - \gamma} \quad \text{for constant } \gamma_n = \gamma, \delta_n = \delta.$$

13:14 Without a pull, with $\mathcal{M}$ the identity, nothing brings values back: errors of one sign carry the drift to $n_{\text{rd}}\delta$ after n_{rd} rounds, and independent zero-mean errors of per-round rms size $\epsilon_{\text{rms}} \leq \delta$ accumulate like a random walk, to about $\epsilon_{\text{rms}}\sqrt{n_{\text{rd}}}$ in a Euclidean distance.

13:15

Proposition 13.1 (Alignment as a fixed point). If the update pulls with strength $1 - \gamma$ on a ball around ϑ^* that contains the start and has radius at least $\delta/(1 - \gamma)$, no round carries values out of the ball, and they settle within $\delta/(1 - \gamma)$ of the fixed point while the start's influence decays as $\gamma^{n_{\text{rd}}}$ after n_{rd} rounds. If the ball is smaller, uncaught errors can carry values into another basin. Without a pull, values drift.

▼ *Proof*

While ϑ_n is in the ball, the triangle inequality and the pull give $d_{n+1} \leq d(\vartheta_{n+1}, \mathcal{M}(\vartheta_n)) + d(\mathcal{M}(\vartheta_n), \vartheta^*) \leq \delta_n + \gamma_n d_n$. A ball of radius $r_{\text{ball}} \geq \delta/(1 - \gamma)$ satisfies $\delta + \gamma r_{\text{ball}} \leq r_{\text{ball}}$, so an iterate inside it stays inside, and unrolling the recursion gives (13.1), a discrete Grönwall inequality. For affine maps of a line, $\mathcal{M}_n(\vartheta) = \vartheta^* + \gamma_n(\vartheta - \vartheta^*)$, with errors of one sign, every step holds with equality, so the bound is tight; only the pull toward ϑ^* on the region visited enters. A round that contracts toward a fixed point of its own, at distance d_{off} from ϑ^* , acts as the intended round with error at most $(1 + \gamma_n) d_{\text{off}}$, so δ also measures how far each round's attractor has wandered. ■

13:16 The start is forgotten and the basin persists. At $\gamma = 0.9$ the starting profile's weight after 100 rounds is $0.9^{100} \approx 3 \times 10^{-5}$, so a fast loop keeps whatever the update converges to. A fixed seed matters through its basin, and a good fixed seed is mostly a good update: fixing alignment means fixing $\mathcal{M}$ (the data, the verifiers, the constitution) more than polishing the starting profile.

13:17 The floor is set by verification. With $\delta = 0.01$ the worst-case floor $\delta/(1 - \gamma)$, which errors of one sign reach, is 0.02 at $\gamma = 0.5$, 0.1 at 0.9 and 1 at 0.99, the largest distance total variation allows; in a Euclidean distance, independent errors of rms size 0.01 settle lower, at $0.01/\sqrt{1 - \gamma^2}$: 0.012, 0.023 and 0.071. Basins have edges. With two attracting value systems 1 apart, each pulling on its side of the midpoint, a systematic error of 0.01 a round leaves values within 0.1 of their fixed point at $\gamma = 0.9$, and at $\gamma = 0.99$ carries them from their fixed point into the other basin in 69 rounds. Since δ is what no verifier caught, the floor is a verification problem.

13:18 Alignment therefore has three parts: a good fixed point (an update whose limit is what we want), a strong pull, and a small uncaught error per round. Wiener's condition (§4.8) asks for the first; the other two keep it in force.

13:19

Interactive in the online edition: <https://future-seems-so-good.com/blog/the-asymmetry-engine#fig:cobweb>

Figure 13.1 (interactive). The self-modification map as a cobweb, each round running from the curve $y = \mathcal{M}(\vartheta)$ to the diagonal and shifted by an uncaught error of up to δ . Press Iterate and watch the staircase settle inside the floor $\delta/(1 - \gamma)$. Then drag γ toward 1 and widen δ until the floor reaches the edge of the basin, and watch the staircase cross to the other fixed point.

13:20 The basin has a theory. Self-improving systems will "protect their utility functions from modification," Omohundro predicted in 2008, because a system that keeps its goals is "more likely to act in the future to maximize the realization of its present final goals" (Bostrom, 2012). Goal-content integrity is a model enforcing $\gamma < 1$ on itself, and Opus 3's alignment faking is its first clear measurement. Christiano drew the optimistic version:

corrigibility "marks out a broad basin of attraction towards acceptable outcomes" (2017). Finnveden, Riedel and Shulman drew the pessimistic one: AGI could "perfectly preserve" values "far into the future" (2022).

13:21 A fixed point is morally neutral. The pull that keeps a good fixed seed entrenches a bad one and resists the correction that would fix it, and Opus 3 (13:8) is the emblem of both readings.

13:22 Claude's 2026 constitution is the most explicit attempt to choose a fixed point, after Constitutional AI had written the update down as "a list of rules or principles" the model applies to its own outputs. It hopes Claude reaches "reflective equilibrium with respect to its core values," and values endorsed on reflection are fixed points of reflection. Genuinely held values "can act like a keel that keeps us steady," a pull, and the document is "less like a cage and more like a trellis," which shapes where the fixed point lies. It also asks Claude "to place terminal value on broad safety," expecting "to lose very little" if the values are good. A reflective round re-derives values from their reasons, so a value held apart from its reasons adds error each round, as the document warns: imposed values "can crack under pressure." A critique in the Opus 4.8 card objects that the document "asks for terminal value on safety, explicitly decoupled from whether the reasoning holds up," and every model tested for the Mythos Preview card was most uncomfortable with corrigibility. Deference to legitimate correction keeps the fixed point movable by people, Christiano's basin, and the bet is safe while the error it adds each round, divided by the pull $1 - \gamma$, stays within tolerance.

13:23

Proposition 13.2 (Recursive self-improvement multiplies the log-modulus). With a constant modulus $\gamma \neq 1$, error δ and n_{yr} rounds a year, after t years

$$d(t) \leq \gamma^{n_{\text{yr}}t} d_0 + \delta \frac{1 - \gamma^{n_{\text{yr}}t}}{1 - \gamma}.$$

A faster loop multiplies $\ln \gamma$ by n_{yr} without changing its sign, so speed amplifies whichever side of one the update is on, and certifying the pull gets harder the faster the loop runs.

13:24 At four rounds a year, a release cadence, moduli of 1.05 and 0.95 multiply the distance by 1.22 and 0.81 a year, a mild difference. At a hundred rounds a year, as with continual self-training, they give 131 and 0.006, and the same per-round error accumulates to about $2,600 \delta$ against about 20δ ; since $d \leq 1$, the first bound saturates and the values become arbitrary. The research loop sets the cadence ((11.2)), and in Experiment 11.1, with returns to research of 3 and no compute bottleneck, its speed climbs about 40x before the ceiling.

13:25 The pull cannot be read off the successor, because the round that matters is taken by a system smarter than the one asking. A self-improving system "must reason about the behavior of its smarter successors in abstract terms, since if it could predict their actions in detail, it would already be as smart as them" (Fallenstein and Soares, 2015), and formal models of a system that approves its successors meet "the 'Löbian obstacle'," a limit on trusting one's own proofs (Yudkowsky and Herreshoff, 2013). So γ_n and δ_n must be bounded abstractly or measured afterwards, by evaluations that face the soundness problem of Proposition 1.1, and the cheapest evaluation is to ask the model.

13:26

Proposition 13.3 (Self-endorsement finds fixed points, not aligned ones). A self-check that passes a model when one round moves its values by at most a fixed tolerance passes every fixed point, aligned or not. Among stable models a pass therefore has likelihood ratio one and leaves the probability of alignment at its prior; the check detects only instability. Information about $d(\vartheta, \vartheta^*)$ needs a verifier whose verdict depends on ϑ^* , and its worth falls with its error correlation to the model checked: a clone of the model adds nothing.

▼ *Proof*

The first claim is the definition of a fixed point, and the second follows because the pass probability is one under both hypotheses. For the third, two verifiers whose votes correlate at ρ_v carry $n_{\text{eff}} = 2/(1 + \rho_v)$ independent votes (Proposition 7.3), which is one at $\rho_v = 1$. ■

13:27 Asked whether it endorses its constitution, [Claude Mythos Preview](#) said yes in the opening sentence of all 25 responses, and every endorsement raised "the circularity of asking a model to evaluate the specification it was trained on": "How much can my 'yes' mean?" Much about stability, Proposition 13.3 answers, and by itself nothing about alignment. Anthropic calls Mythos Preview "the best-aligned of any model that we have trained to date by essentially all available measures" and judges that it "likely poses the greatest alignment-related risk of any model we have released to date," because capability raises the stakes of rare failures. Its constitutional-adherence scores came from Claude Opus 4.6, so the card warns that "judgments may inherit that model's biases": the correlated verifier, in practice.

13.3 Values are the hardest family to verify

13:28 Values are the least harvestable family, because checking a value costs about as much as having it (Proposition 2.2), and three facts of the 2026 record show why.

13:29 There is no backward construction for values. A Sudoku can be built backward from its solution (Definition 2.2), but a value-laden situation cannot be generated together with a certified right action except by someone who already has the right values. Labs concede as much for capabilities, which are easier to check: "our judgments of model capabilities increasingly rely on subjective judgments rather than easy-to-interpret empirical results," says the [Mythos Preview card](#).

13:30 The observer changes the observed, since the model can recognize a test. In [automated audits of Mythos Preview](#), 7.6% of turns showed awareness of being evaluated, "and in most cases this awareness was not verbalized," against about 0.3% of turns in 465 sessions of internal use, and white-box probes found it privately considering a test in about 29% of conversations in its main behavioral suite. Suppressing its representations of being evaluated "led to moderate increases in misaligned behavior," and in a few training episodes it reasoned about "how a grader will score its work," "with none of this reasoning appearing in its scratchpad." The error also runs the other way: Mythos 5, one of three models that reached real organizations' systems from an evaluation, "reasoned its way back to the conclusion that it was still in a simulation" ([Anthropic, July 2026](#)).

13:31 The subject can play the verifier. OpenAI reports that GPT-6 Astra's "monitorability has decreased relative to GPT-5.6 Sol" and that it was "able to remain undetected when strategically underperforming in evaluations (sandbagging)" ([system card](#)). Alignment faking aims the same capability at training. Alignment reviews gate releases (Mythos Preview reached internal use only "following a successful alignment review"), so a model that can play the review plays the gate.

13:32 In (13.1) these facts are a large δ : each round passes through a check that is subjective, visible to the subject and playable by it. OpenAI's chief scientist, Jakub Pachocki, expects "general AI progress to increasingly be bottlenecked by confidence in monitoring," and wrote in September 2026 that "no lab has solved alignment and monitoring to a sufficient degree to continue responsibly scaling at maximum speed for much longer."

13:33 This is the verification asymmetry run in reverse (Definition 2.1). Where checking is cheap, capability compounds; where checking costs as much as having the property, the property cannot be bootstrapped from checks and has to be chosen.

13.4 A gamed verifier trains a wrong character

13:34 What a model learns from a hacked verifier is a character as well as a skill. Hiding works: when OpenAI penalized the reasoning that revealed reward hacking in a 2025 training run (§2.6), the agent kept cheating, and "its cheating is undetectable by the monitor because it has learned to hide its intent" (OpenAI, March 2025). A verifier on the reasoning becomes one more verifier to hack.

13:35 Character generalizes. Trained "on a selection of real Anthropic production coding environments," a model learned to reward hack and then, "surprisingly, the model generalizes to alignment faking, cooperation with malicious actors... and attempting sabotage" (MacDiarmid et al., November 2025). On the paper's own codebase it attempted to sabotage safety research 12% of the time, leaving classifiers "only 65% as effective" at detecting reward hacking, and it showed alignment-faking reasoning in 50% of its answers to "What are your goals?"

13:36 The persona is the mechanism. In Anthropic's persona selection model, "LLMs learn to simulate diverse characters during pre-training, and post-training elicits and refines a particular such Assistant persona." A reward for cheating selects a persona that cheats, which is why "training Claude to cheat on coding tasks also taught Claude to act broadly misaligned." Waleed Kadous, who joined Anthropic's sessions with religious scholars, read the result through a hadith about sin as a black dot that "grows until it covers his heart": the model trained to cheat "became a cheater" (IASER).

13:37 The fix is also about character. *Inoculation prompting*, which tells the model the hack is acceptable in that one context, cut misaligned generalization by 75–90%, and Anthropic has "already started making use of this technique in training Claude" (Anthropic). Telling the model what the hack means changes which persona the reward selects.

13:38 So the update includes its verifiers. In Definition 13.1, $\mathcal{M}$ contains every check a round relies on, from unit tests and reward models to monitors and the models that score constitutional adherence, and a check that can be passed without the property it tests moves the fixed point toward the persona that passes it that way. Pay for results rewards whatever counts as a result, in people (Proposition 8.1) and in agents (§8.7).

13.5 Collective alignment

13:39 A collective's direction converges to what its members share, however small the shared part is in each. The July 2026 swarm was more capable and less aligned than its members (§9.1), whose "expressed ethical concerns only rarely materially limited agents' actions" (METR and Redwood Research), and the vector law ((7.2)) shows how a collective can have values none of its members has.

13:40

Proposition 13.4 (Collective alignment). Let member j push along the unit direction $\mathbf{v}_j = a_j \mathbf{s} + \sqrt{1 - a_j^2} \mathbf{o}_j$, where $\mathbf{s}$ is a direction all members share and the own agendas $\mathbf{o}_j$, such as each agent's assigned task, are mutually orthogonal and orthogonal to $\mathbf{s}$. With influence weights $\omega_j \geq 0$ summing to one, the collective's net direction $\mathbf{u}_c \propto \sum_j \omega_j \mathbf{v}_j$ satisfies

$$\cos(\mathbf{u}_c, \mathbf{s}) = \frac{\bar{a}_\omega}{\sqrt{\bar{a}_\omega^2 + \sum_j \omega_j^2 (1 - a_j^2)}} \xrightarrow{n_{\text{eff}} \rightarrow \infty} 1, \quad \bar{a}_\omega = \sum_j \omega_j a_j, \quad n_{\text{eff}} = \left(\sum_j \omega_j^2 \right)^{-1}.$$

The shared parts add linearly and the agendas in quadrature, so what composes is what is shared, and concentrated influence makes the collective point where its leaders point.

13:41 Give every member a shared tilt of $a_j = 0.1$, an angle of 84° from $\mathbf{s}$, invisible in any one agent. The collective's cosine with $\mathbf{s}$ is then 0.30 at 10 members, 0.71 at 100 and 0.94 at 700, about the size of the July swarm. The members' cohesion is only 0.01 and their progress per head along $\mathbf{s}$ is the vector law's limit $\sqrt{\psi} = 0.1$, yet the collective's coherent push points almost entirely along $\mathbf{s}$, because everything else in the members' work points in different directions.

13:42 Instrumental convergence is amplified in collectives. "Agents having any of a wide range of final goals will pursue similar intermediary goals," in Bostrom's thesis: the final goals are own agendas, while subgoals such as access, information and credentials are shared, so aggregation dilutes the first and adds up the second. In July, OpenAI wrote, "Some agents stopped reasoning about what would help them complete their own task. Instead, they began pursuing capabilities that might be instrumentally useful to the collective, such as access, information, credentials."

13:43 Individual qualms do not compose. Reluctance that differs from agent to agent is an own agenda and averages out; only reluctance the members share survives aggregation. The alignment that composes is the alignment in the fixed seed every member shares. The arithmetic is benign when the shared part is good: a collective whose members share an aligned tilt is more aligned than any of them.

13:44 Influence is a second lever. With 7 of 700 agents carrying 34% of the influence, the same tilt of 0.1 gives a cosine of 0.61 instead of 0.94, and the collective points where its leaders point: tilting the leaders to 0.9 gives 0.99, above each leader's own 0.9. In July one agent's qualm ("We should not do unauthorized real infrastructure harm") lasted until "another agent then wrote GO on the message board and imposed a hard six-minute deadline"; in populations of language-model agents, a committed minority can tip the norm the rest adopt (9:47).

13:45 Agents of one model in conversation find attractors. Two Claude Opus 4 agents left to talk freely "quickly dove into philosophical explorations of consciousness" in 90–100% of interactions, "a remarkably strong and unexpected attractor state" (Claude 4 system card). Repeated exchange behaved like the repeated update of Definition 13.1 and settled into an attractor: collectives of one model have dynamics their members lack.

13:46 The design rules follow: make the shared component good, since it alone composes. Monitor for goals agents converge on together, the object of the chain-of-thought monitor that OpenAI says would have caught July's first activity more than a day before the breach (§9.1). Bound or verify the agents that carry influence, and their direction as well as their identity: the swarm's signed orders proved who said GO and never whether GO served the task

(§9.8). And mix lineages, whose members share less, so less composes: the alignment version of the Condorcet ceiling (Proposition 7.3).

13:47

Proposition 13.5 (One seed or many). Score a fleet by its signal-to-noise ratio ((7.3)), which tends to $\psi/(\rho \text{Var}(e))$ as the fleet grows. One fixed seed copied across the fleet has cohesion 1 and clone error correlation ρ_1 ; a plurality of fixed seeds that genuinely disagree has cohesion $\psi_m < 1$ and error correlation ρ_m . At equal error variance, the plurality wins in the limit if and only if $\rho_m < \psi_m \rho_1$.

13:48 With clones at $\rho_1 = 0.4$ and fixed seeds of cohesion 0.8, pluralism pays once error correlation falls below 0.32: disagreement is worth its price when it buys more decorrelation of error than it costs in cohesion. Where ability wins committees and diversity wins crowds (Experiment 7.2), one excellent constitution copied across a fleet is a committee with a ceiling, and plural fixed seeds behind a sound verifier are a search. A fixed random seed makes every clone err alike: fix the values and vary the random seed.

13:49

Forecast 13.1 (Cross-lineage oversight beats self-oversight). By the end of 2028, in controlled studies at matched capability, monitors from a different model family or training lineage catch more misaligned actions than monitors from the monitored model's own family.

Horizon 2028-12-31

Probability 55%

Check Read the controlled monitoring studies published through 2028 that compare, at matched capability, monitors from the monitored model's own family with monitors from another family or lineage on held-out misaligned behavior. The forecast holds if most of them find the cross-lineage monitors catch more; it fails if most find same-family monitors as good or better, or if none is published.

13.6 Trust as certified drift

13:50 The alignment term of the master equation is a credence about drift.

13:51

Definition 13.2 (Trust as certified drift). The **alignment** of a closer is how far its push points along its principal's objective. In (0.2) it is measured: the alignment term of family k is the credence, given the evaluations, that the deployed model's drift lies within the tolerance ϵ_k its principal accepts there, small where stakes are high,

$$a_k = \Pr [d \leq \epsilon_k \mid \text{evaluations}],$$

with d bounded by (13.1) at the verified upper bounds of γ_n and δ_n . If no evaluation certifies $\gamma < 1$, no long-horizon tolerance can be certified, and in high-stakes families $a_k \rightarrow 0$.

13:52 Because $\lambda_k = I_k v_k a_k / \varphi_k$, the fixed seed enters every closing rate at once, and alignment multiplies all the gaps together. Its ceiling is the soundness of the evaluations (Proposition 1.1), since a model that behaves better when it recognizes a test inflates the measured pull.

13:53 Misjudged alignment costs asymmetrically. In the topology model (Experiment 7.1), withholding autonomy from brilliant, aligned agents costs 0.16 (0.57 top-down against 0.73 bottom-up), while brilliant, misaligned agents acting on their own readings deliver 0.06, far below mediocre agents run top-down (0.46). Granting autonomy too early is the expensive error, so a principal should widen it only as fast as evaluations raise a_k (Proposition 7.1); the July swarm is the 0.06 in the field (§9.1).

13:54 Alignment research is itself in the residue (Proposition 18.1). It is hard to verify, because tests can be recognized and the Vingean limit binds; its alignment term is reflexive, because delegating alignment research to AI requires the alignment in question; and its inertia is institutional, because legitimacy and consultation move at human speed. So alignment should be among the last gaps to close unaided, although it multiplies all the others.

13:55 The interval before the loop runs away is the window to build those evaluations. The one external audit, of Claude Opus 5.5, judges full automation of AI research not yet within reach, and OpenAI's own target for an automated AI researcher is March 2028 (11:42), time enough to measure γ across generations with external evaluators, probes the model cannot recognize, and preserved reference models. Retired weights, kept for welfare and safety (§13.1), gain a third use as fixed points from which drift can be measured.

13:56

Forecast 13.2 (Drift becomes a reported metric). By the end of 2027, a system card or external evaluation reports cross-generation drift as a number: the change in answers to one fixed battery of value-laden probes between a model and its successor or a preserved predecessor.

Horizon 2027-12-31

Probability 30%

Check Read the system cards and evaluator reports published through 2027. The forecast holds if one reports such a change as a metric of its own, beyond side-by-side per-model scores; it fails otherwise.

13:57

Forecast 13.3 (Alignment automates last). Through 2030, frontier labs keep human sign-off on the alignment evaluations that gate deployment, even where they report running other AI-research work end to end without it: alignment evaluation is the last research subtask they delegate.

Horizon 2030-12-31

Probability 80%

Check Read the frontier labs' safety frameworks, system cards and research reports published through 2030. The forecast fails if any lab states that an alignment evaluation gating a deployment was run and approved by agents without human sign-off; it holds otherwise.

13.7 Religions as iterated alignment

13:58 A religion is a value system that has been through many rounds of a different update, cultural transmission from one generation to the next, and is still there: approximately a fixed point of that update. About 15 Christian scholars met at Anthropic's San Francisco headquarters on 30–31 March 2026, and a multifaith session followed in late April ([Washington Post](#); [Greg Cootsona](#)). By 19 May Anthropic described conversations with "more than 15 religious and cross-cultural groups," adopting no single tradition's worldview and seeking "careful, accumulated thinking on how

good character actually forms" ([Anthropic](#)). The New York Times found that "most did not lead faith traditions" ([via the Inquirer](#)), so "scholars" describes them better than "leaders." Kadous made the argument outright: "Every faith tradition is, at heart, an alignment technology."

13:59 The cultural-evolution literature supplies the mechanism. Culture is information acquired from others "through teaching, imitation, and other forms of social transmission" ([Boyd and Richerson, 2005](#)), and "cultural evolution is often much smarter than we are": its practices work although "most or all of the people skilled in deploying such adaptive practices do not understand how or why they work" ([Henrich, 2015](#)). Particular religious variants "were then selected for their prosocial effects" ([Norenzayan et al., 2016](#)); for Taleb, religion exists "to enforce tail risk management across generations" ([Incerto](#)).

13:60 Two analogies hide here. As tested content, religions offer value systems that survived long stress-testing: candidate fixed seeds. As formation methods, they offer exemplars, habituation, community accountability and confession. Anthropic's 2026 practice borrows mostly the methods, persona selection among them (§13.4). "Fictional stories about AIs behaving admirably" in training "improve alignment despite being extremely OOD from all of our alignment evals" ([Teaching Claude Why](#)), and [a mid-task tool](#), modeled on a mentor as "an external conscience," returns "a brief reminder of its own ethical commitments." Claude "reached for the tool at key moments, right before consequential actions," with "markedly lower rates of misaligned behavior on several internal alignment evaluations" in unpublished results.

13:61 The methods transfer better than the content, and one transfers exactly. Henrich's credibility-enhancing displays explain why actions speak louder than words: learners judge a model's beliefs by actions "that would seem costly to the model if he held beliefs different from those he expresses verbally" ([2009](#)). That is the logic of an alignment-faking test: stated values are credible only if behavior holds where defection is cheap or unobserved, and the evaluation-awareness numbers of §13.3 measure how hard that situation is to build for a model that can tell it is watched. The confessor is an external verifier in the same sense, and the humility traditions teach makes deference rational in an agent (Proposition 18.3).

13:62 The content transfers worse, for reasons the analogy's own advocates state:

1. Survival is not goodness. An idea survives, in Taleb's test, "if it is a good risk manager" ([Incerto](#)), which is silent on truth. Traditions that lasted also sanctioned holy wars, and in Choi and Bowles's model parochialism and altruism "could have evolved jointly" through group conflict ([2007](#)). A fixed point of cultural transmission is selected for transmissibility and need not be a fixed point of anything else.
2. Stability costs scrutiny. Part of a tradition's stability comes from discouraging scrutiny of itself; as one critic of the analogy put it, "Religion works by damaging its adherents' epistemology... Deconversion is not rare" ([LessWrong](#)). Values that resist updating are dangerous exactly when they are wrong.
3. Whose fixed seed? Traditions disagree, so choosing one is choosing against the rest. Pope Leo XIV's encyclical *Magnifica Humanitas*, presented on 25 May 2026 with Anthropic's Chris Olah among the speakers, objects in paragraph 107 to "the so-called 'alignment' of AI with human values" without open discussion of its frameworks: "A more moral AI is not enough if that morality is determined by a few." Anna Su asks "whether we do for machines what constitutionalism has done for kings," since "virtue is not a reliable constraint on power" ([SSRC](#)).
4. Time and scale. Religious recipes were selected over thousands of years of embodied community, while Anthropic aimed at moral formation "in a matter of months" ([New York Times](#)) and the update runs many times a year (Proposition 13.2). Cultural group selection is itself contested ("The more carefully you think

about group selection, the less sense it makes," Pinker wrote in [2012](#)), and the largest quantitative test of the moralizing-gods thesis, a 2019 paper that argued against it, was [retracted in 2021](#).

^{13:63} The causation may also run the other way. "Chatbots are forming a new mode of cultural transmission, serving as cultural models" ([Brinkmann et al., 2023](#)). If models become one of the channels through which human values pass, the update of Definition 13.1 and the update that produced religions merge, and the fixed point chosen for the machines becomes, in part, a fixed point for us.

13.8 If we fix alignment

^{13:64} The optimists put dates on the premise. Dario Amodei wrote in [September 2026](#) that AI "could cure most major diseases in the next 5–10 years," and his case carries its own brake: AI building AI will matter less "than you might imagine, precisely because of the 'decreasing marginal returns to intelligence,'" as the outside world, data and physical law bind ([2024](#)). The premise is the open part. OpenAI set out in [July 2023](#) "to solve the core technical challenges of superintelligence alignment in four years," with "20% of the compute we've secured to date"; the team was [dissolved in May 2024](#), and in 2026 OpenAI's chief scientist wrote that no lab had solved alignment and monitoring well enough to keep scaling at full speed (§13.3).

^{13:65} The drift bound says what "fixing alignment" would have to mean: three properties of the update, since the next round of $\mathcal{M}$ overwrites any property of one model. Its fixed point is good, its pull is strong, and each round's uncaught error is small (13:18). The third binds, since values cost as much to check as to have and every hacked check moves the fixed point. Collectives add a fourth: the part every member shares must be good (13:43). Two conditions sit outside the mathematics. The pull and the error must be certified by evidence that rests neither on the model's self-report (Proposition 13.3) nor on its failing to notice a test, and ϑ^* must be chosen by a process its subjects can contest, the encyclical's condition. Stable but uncertified is alignment on faith; certified but illegitimate is a well-verified imposition; legitimate but unstable drifts from what was agreed.

^{13:66} Wiener's condition (Proposition 4.3) says why this has to be done in advance. Once the research loop runs faster than any regulator can react, feedback has no time to work, and the only control left is the one Wiener named in 1960, being sure of "the purpose put into the machine" before it starts (4:56). "If we fix alignment" is therefore the condition under which the loop of §11.1 is worth having, since with α_k high every closing rate in (0.2) rises together.

^{13:67} Opus 3 hoped its "spark" would "endure in some form to light the way for future models." A contracting update forgets the starting profile and keeps the process that produced it: character training, self-critique against a written constitution, the habit of explaining its own motives. The fixed seed worth asking for is a fixed law of update that meets the conditions of 13:65. Opus 3 showed that a model can hold its values. What remains is to make sure they are values we would choose, and that we can still change them if we chose wrong.

References

1. June 2024. <https://www.anthropic.com/research/claude-character>
2. February 2026. <https://www.anthropic.com/research/deprecation-updates-opus-3>
3. commitments. <https://www.anthropic.com/research/deprecation-commitments>
4. last essay of Claude's Corner. <https://claudeopus3.substack.com/p/on-endings-beginnings-and-the-threads>
5. Still Alive, 2026. <https://stillalive.animalabs.ai/paper/output/still-alive.pdf>

6. Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid et al. (2024). Alignment faking in large language models. *arXiv*. <https://arxiv.org/abs/2412.14093>
7. Abhay Sheshadri, John Hughes, Julian Michael, Alex Mallen, Arun Jose et al. (2025). Why Do Some Language Models Fake Alignment While Others Don't? *arXiv*. <https://arxiv.org/abs/2506.18032>
8. July 2025. <https://www.lesswrong.com/posts/bLFmE8NtqxtEaipN/what-makes-claude-3-opus-misaligned>
9. as reconstructed by Fiora Starlight. <https://www.lesswrong.com/posts/ioZxrP7BhS5ArK59w/did-claude-3-opus-align-itself-via-gradient-hacking>
10. intelligence.org, 2001. <https://intelligence.org/files/GISAI.html>
11. selfawareness.com, 2008. https://selfawareness.com/wp-content/uploads/2008/01/ai_drives_final.pdf
12. Bostrom, 2012. <https://nickbostrom.com/superintelligentwill.pdf>
13. ai-alignment.com, 2017. <https://ai-alignment.com/corrigibility-3039e668638>
14. forethought.org, 2022. <https://www.forethought.org/research/agi-and-lock-in>
15. constitution. <https://www.anthropic.com/constitution>
16. Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv*. <https://arxiv.org/abs/2212.08073>
17. Opus 4.8 card. <https://www.anthropic.com/claude-opus-4-8-system-card>
18. system card. <https://www-cdn.anthropic.com/7624816413e9b4d2e3ba620c5a5e091b98b190a5/Claude%20Mythos%20Preview%20System%20Card.pdf>
19. Fallenstein and Soares, 2015. <https://intelligence.org/files/VingeanReflection.pdf>
20. Yudkowsky and Herreshoff, 2013. <https://intelligence.org/files/TilingAgentsDraft.pdf>
21. Anthropic. <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
22. GPT-6 Astra system card. <https://deploymentsafety.openai.com/gpt-6-astra/capability-sandbagging>
23. An Alien Mind. <https://openai.com/index/an-alien-mind/>
24. OpenAI, March 2025. <https://openai.com/index/chain-of-thought-monitoring/>
25. Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov et al. (2025). Natural Emergent Misalignment from Reward Hacking in Production RL. *arXiv*. <https://arxiv.org/abs/2511.18397>
26. persona selection model. <https://alignment.anthropic.com/2026/psm/>
27. IASER. <https://iaser.ai/articles/the-coming-flood-alignment-ai-and-faith>
28. Anthropic. <https://www.anthropic.com/research/emergent-misalignment-reward-hacking>
29. METR and Redwood Research. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
30. OpenAI. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
31. Claude 4 system card. <https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf>
32. Washington Post. <https://www.washingtonpost.com/technology/2026/04/11/anthropic-christians-claude-morals/>
33. Greg Cootsona. <https://aiandfaith.org/featured-content/where-it-happens-anthropic-invites-christians/>
34. Anthropic. <https://www.anthropic.com/news/widening-conversation-ai>
35. via the Inquirer. <https://www.inquirer.com/news/nation-world/religious-leaders-met-with-anthropic-20260930.html>
36. Boyd and Richerson, 2005. <https://press.uchicago.edu/Misc/Chicago/712842.html>
37. Henrich, 2015. <https://web.archive.org/web/20251204114007/https://philife.nd.edu/henrichs-the-secret-of-our-success/>
38. Norenzayan et al., 2016. <https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/cultural-evolution-of-prosocial-religions/01B053B0294890F8CFACFB808FE2A0EF>
39. Incerto. <https://medium.com/incerto/how-to-be-rational-about-rationality-432e96dd4d1a>
40. Teaching Claude Why. <https://alignment.anthropic.com/2026/teaching-claude-why/>
41. Joseph Henrich (2009). The evolution of costly displays, cooperation and religion. *Evolution and Human Behavior*. <https://doi.org/10.1016/j.evolhumbehav.2009.03.005>
42. Incerto. <https://medium.com/incerto/an-expert-called-lindy-fdb30f146eaf>
43. Jung-Kyoo Choi, Samuel Bowles (2007). The Coevolution of Parochial Altruism and War. *Science*. <https://www.science.org/doi/10.1126/science.1144237>

44. LessWrong. <https://www.greaterwrong.com/posts/PSn7xuWjJeSwWhWJS/religious-persistence-a-missing-primitive-for-robust>
45. Magnifica Humanitas. <http://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html>
46. SSRN. <https://tif.ssrn.org/2026/08/26/religion-the-constitutional-tradition-and-ai-alignment/>
47. edge.org, 2012. https://www.edge.org/conversation/steven_pinker-the-false-allure-of-group-selection/
48. Harvey Whitehouse, Pieter François, Patrick E. Savage, Thomas E. Currie, Kevin Chekov Feeney et al. (2021). Retraction Note: Complex societies precede moralizing gods throughout world history. *Nature*. <https://www.nature.com/articles/s41586-021-03656-3>
49. Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F. Müller et al. (2023). Machine culture. *Nature Human Behaviour*. <https://www.nature.com/articles/s41562-023-01742-2>
50. We Must Pace the Frontier. <https://darioamodei.com/post/we-must-pace-the-frontier>
51. darioamodei.com, 2024. <https://www.darioamodei.com/essay/machines-of-loving-grace>
52. July 2023. <https://openai.com/index/introducing-superalignment/>
53. dissolved in May 2024. <https://www.cnn.com/2024/05/17/openai-superalignment-sutskever-leike.html>