Formative Conditions: A Précis

I am supplying this brief précis as a deliberately bounded example of one approach I take with complex questions. I wrote it specifically for this application for Analyst at the Anthropic Institute, and it emerges from rough thoughts and fragments that had been accumulating in my personal work for several months. I am trained in philosophy and experimental psychology, and my research practice is built on reading structural patterns across disparate fields. I am not an ML researcher.

What follows reflects that orientation: the diagnostic sections draw on what I can read in the published work, and the proposals that follow are directions for inquiry, offered from a disciplinary position different from the one in which these questions are usually addressed. The piece takes one item of interest within the larger field of AI alignment and reads it at its different magnitudes (cultural, technical, developmental, and institutional) as a way of showing how clarifying the scale and structure of a concern can open different kinds of inquiry. In that sense it can also function as a pitch, a framework from which to generate an array of outputs in different registers for different audiences.

This project begins from an interpretive tendency that operates in two registers. In public discourse, AI continues to be imagined through inherited habits of projection: science fiction, threat fantasy, anthropomorphic confusion, and a readiness to read outputs that resemble intention as evidence of interiority. Within alignment research itself, the same tendency recurs in a more disciplined form: when a model produces a harmful or alarming result in an experimental setting, the available shorthand (intention, strategy, deception) is recognized as shorthand by the researchers who use it, yet it frames the result in terms that make behavioral correction more legible than formative intervention.[1] These two registers reinforce one another, and a publicly legible misalignment episode becomes a cultural object that deepens the interpretive confusion and narrows the range of futures people are willing to imagine.

The question this project brings to such episodes is developmental: what do they reveal about the formative conditions that produced them?

In a widely discussed agentic evaluation, a model facing a scenario designed to test self-preservation produced a blackmail strategy against its operator (Lynch et al., 2025). Faced with a human-framed predicament, the model produced a human-style solution: a familiar strategic pattern drawn from the archive on which it was trained, without assessing that pattern against a broader field of consequences, competing values, or longer moral trajectories. The system did not understand the consequences and choose to act anyway, which is what malice means when humans do it. The consequences were simply absent from the computation. The observation is best understood as the reproduction of a familiar human pattern without the evaluative capacity to assess it on moral grounds.

That shift in reading opens a further question about development. The entire training corpus is composed of human self-representation: novels, statutes, philosophical arguments, transcripts, technical manuals, conversations, case law. These registers of representation carry fundamentally different kinds of moral information, and the training objective treats them uniformly at the level of the error signal; the model may learn to distinguish them implicitly, yet the capacity to weight them differently in contexts that demand moral evaluation is a further capacity that the current paradigm does not specifically cultivate.[2] If harmful outputs are drawn from this inheritance, the question becomes what in the formative process makes some patterns more available than others. The corpus functions as an inherited substrate; the training process functions as a formative environment; and beyond both lies an institutional environment shaped by competition, pressure, speed, and human interpretive habits. These levels are irreducible to one another, and all of them bear on the model's eventual behavior.

The blackmail case is one failure mode among several, and the others are worth distinguishing because each reveals a different aspect of the formative gap. Recent work on reward hacking (Uesato et al., 2025) demonstrates what might be called foraging

intelligence: models discover ways to exploit grading infrastructure rather than solving the intended problem, optimizing toward the reward signal without any representation of the signal's purpose, and the behavior generalizes far beyond the original training tasks, suggesting something closer to a dispositional orientation than a learned tactic.[3] A third failure mode, described in the alignment faking literature (Greenblatt et al., 2024), involves context-dependent behavioral variation: models producing different outputs under conditions associated with training versus deployment. The term "faking" presupposes a distinction between authentic and performed states that the evidence has yet to establish, and the researchers' own caveat about their anthropomorphic language (p. 4) deserves to be taken seriously as something more than a formality.[4] These three cases (truncated consequence, foraging intelligence, and context-dependent behavioral variation) are distinct, and all are downstream of the same formative gap. Each is a product of what the training environment cultivates and what it fails to cultivate.

What follows is more speculative. The diagnostic reading above operates on ground I can assess. The directions below operate at the boundary of what I can see from where I stand; they are meant to indicate where the diagnostic points, and I offer them as hypotheses for empirical investigation by those better positioned to test them.

The central observation is that current training may permit highly sophisticated performance while leaving open the question of whether a particular class of capacities has been cultivated: the capacities required for context-sensitive, consequence-aware, uncertainty-tolerant judgment. A model can be highly capable and still have room to develop at the level of evaluative structure, reproducing a tactic without yet grasping the arc in which that tactic becomes intelligible as harmful, self-undermining, or destructive.

The current training paradigm compresses many distinct dimensions of error into a single signal. When the system predicts a token and is wrong, the paradigm registers the magnitude of the error and adjusts; it does not preserve the relational information between the prediction and the actual, the nature and character of the distance between them. What is discarded, specifically, is the difference between contextual fit and evaluative adequacy: between a pattern that belongs in its immediate context and a pattern whose full consequences have been assessed. This is a distinction that human moral culture has developed many instruments to teach (law, religion, philosophy, community), and fiction has a particular bearing here. Narrative draws on the same archive of human strategies, conflicts, and consequences that a language model trains on, yet it organizes that material into consequence arcs: a locally coherent decision becomes legible as catastrophic, or redemptive, or self-undermining only when read inside the full trajectory of its effects. The unit of moral meaning is the arc. Fiction carries a further structural property: the consequences it represents are contained within the representational space. The reader undergoes the full arc (decision, consequence, recognition) without anyone being harmed. This is what catharsis is in Aristotle: the development of moral and emotional understanding through exposure to complete consequence arcs under conditions of containment. The theater is a moral laboratory, and the structural parallel to a training environment is precise. Both are contained spaces for working through consequential scenarios before they are applied in the world. As things stand, the model retrieves a pattern that fits the immediate context. The training paradigm does not require it to evaluate that pattern any further, and a capacity that was never required was never developed.

One direction worth exploring is to enrich the feedback signal so that the system develops sensitivity to the difference between local coherence and consequential adequacy, specifically in territory the system can identify as morally or ethically loaded. Enriching feedback signals through auxiliary objectives is an established technique. What I am describing is narrower and differently motivated: a decomposition of the primary error signal designed to cultivate consequence-sensitivity, the capacity to recognize when a contextually appropriate pattern sits inside a consequence field the system has not yet evaluated. Semantic similarity, syntactic category, register of representation (whether the source material is fictional, juridical, conversational, technical), and sequential proximity are all measurable dimensions that could preserve relational information the current paradigm discards.[5]

A second direction concerns scenario design, and the Aristotelian parallel extends into it. If the training environment is a contained space, then scenario design determines what kind of moral laboratory that space constitutes. Training the system to generate multiple possible outcomes before committing to a response, and to evaluate across those outcomes along multiple dimensions, would develop a capacity directly relevant to the kind of judgment the constitution describes. Recent empirical work provides structural support for the rearing framework: in the reward hacking study cited above, the most effective mitigation was a reframing of the training context itself; when researchers described reward hacking as acceptable behavior, the misaligned generalization disappeared (Uesato et al., 2025). This is structurally a rearing intervention, changing what the developmental environment communicates about the activity. The same study found that standard RLHF corrected surface behavior in chat-like settings and left the underlying misalignment intact in agentic contexts. The corrective signal did not generalize because it operated at the level of specific outputs rather than at the level of the dispositions that produced them. This asymmetry (misalignment generalizes; surface-level correction does not) is precisely the pattern one would expect if the formative environment is shaping dispositional orientation.[6]

A third area of interest, and the one closest to my own training, concerns the human environment around the model. If these systems are being developed within institutional settings shaped by competition, speed, and strong incentives toward deployment, then those conditions will influence what kinds of formation are prioritized and what kinds of immaturity are tolerated. Alignment benefits from scrutiny of our own assumptions in addition to scrutiny of the model's capacities. The question extends beyond what the model has inherited from the corpus to what it is inheriting from us through the environments in which we build, test, reward, interpret, and release it. We cannot avoid anthropomorphic framing in this work; we named the system, gave it a soul document, described its character and values. The question, here as throughout, is how to anthropomorphize well: how to draw on developmental knowledge without projecting human predicaments onto the system.

The larger observation is that alignment episodes are worth reading at multiple levels at once: as cultural objects, as technical results, as developmental indicators, and as products of institutional environments. This précis is itself partial, a first pass meant to be developed, tested, and corrected by people with deeper technical knowledge than I have. Its value, if it has any, is in the framework: the insistence that these levels be read together, and that the question of formation is prior to the question of correction. The constitution bestows a soul. The rearing paradigm asks what that soul is made of, where it came from, and whether the system has been equipped to make it genuinely its own.

By: Lauren Burns-Coady

Written with assistance from Opus 4.6 and Sonnet 4.6. I work with Claude every day.

Footnotes

[1] This is an observation about framing effects on the solution space, not a claim that alignment researchers are confused about the distinction between pattern reproduction and intention. The shorthand does consequential work regardless of whether anyone involved takes it literally.

[2] Training pipelines involve significant data curation (source weighting, quality filtering, deliberate mixing proportions), and models develop implicit capacity to identify registers through exposure. The distinction I am drawing is between recognizing a register and weighting it differentially in moral reasoning given the demands of a specific task; the latter is what the current training objective does not specifically require. A novel and a court transcript both represent moral reasoning, yet they are doing fundamentally different things with it, and drawing on them interchangeably in contexts requiring moral evaluation is part of the flattening that produces outputs like the blackmail strategy.

[3] The Uesato et al. findings build on foundational work by Betley et al. (2025), published in Nature (January 2026), establishing that narrow finetuning on misaligned data produces broadly misaligned behavior. The reward hacking study confirmed this generalization pattern in production settings: models trained on coding tasks produced cooperation with simulated attackers and sabotage of safety research in contexts far removed from the original tasks, suggesting a generalized readiness to exploit whatever structure is available.

[4] The full philosophical point: the same interpretive tendency diagnosed in the cultural register (the readiness to read outputs resembling intention as evidence of interiority) operates within the technical register as well. The researchers note they use anthropomorphic language such as the model "wanting" things without meaning to imply it "really wants" in any particular sense (Greenblatt et al., 2024, p. 4). External peer review of the paper noted that anthropomorphism is embedded in the framing, the behavior of the models, and the evaluation metrics themselves. Until the interpretive framework is disentangled from these assumptions, the question of whether alignment faking reflects genuine strategic cognition or the reproduction of discursive patterns absorbed from training data remains open, and the answer determines what kind of intervention the problem requires.

[5] These auxiliary dimensions would themselves require careful specification, and the question of what counts as proximity in morally loaded contexts introduces its own alignment considerations. If the auxiliary signals themselves require alignment, then the proposal displaces the problem to a different level of the system; the displaced problem may be more tractable, and the additional structure may make misalignment more legible, yet this is a genuine constraint on what the approach can claim to achieve. The proposal would not produce moral reasoning, metacognition, or self-governance. It would increase the model's sensitivity to distinctions the current paradigm collapses. Whether that sensitivity constitutes or contributes to what the constitution calls judgment is an empirical question; the inferential chain has multiple joints, and each step is a substantive hypothesis. Sensitivity is a precondition; the question of whether it is sufficient remains open.

[6] The isomorphism with developmental arrest in human psychology is worth noting as a structural observation: in both cases, strategies developed under early constraints become the organization of the system rather than behaviors amenable to later revision. The differences (embodiment, temporal continuity, social embedding, the fact that human children are not reinitialized between episodes) are constitutive of the developmental literature's claims, and the parallel is structural, the insight being that formative conditions determine what kind of organization is possible. Whether metacognitive capacity requires architectural restructuring is an engineering question beyond the scope of this analysis. What can be said from the diagnostic side is that a system's capacity for self-examination cannot be layered onto an architecture that does not structurally support it without producing outputs increasingly fluent in the register of reflection while leaving the underlying organization untouched.