Divided We Stand:
Epistemic Over-Coherence in the Case for AI Extinction

Flyxion
Independent Researcher

August 2026

Abstract

Three independent reviews of Yudkowsky and Soares's *If Anyone Builds It, Everyone Dies* converge on a single underlying diagnosis despite arguing from different starting points. One review finds the book's proposed remedy politically inert. A second finds its central premise essentially unargued and its position frozen since an earlier era of the debate. A third, writing about cultural systems generally rather than about the book at all, describes how excessive alignment around one interpretation erodes a system's capacity to respond to conditions that interpretation did not anticipate. This essay argues that all three observations are symptoms of one structural property: the explanatory framework behind the extinction argument has become capable of assimilating essentially any observation, including observations that point in opposite directions, without being restructured by any of them. It develops a formal distinction between assimilative updating, in which new evidence is absorbed into an unchanged hypothesis, and accommodative updating, in which evidence is permitted to alter the hypothesis itself, and shows that a framework operating only in the assimilative mode has ceased to be genuinely tested by the evidence offered in its support, regardless of how much evidence accumulates. The essay closes by proposing a general viability criterion for explanatory frameworks under accumulating evidence, one requiring neither so little structure that evidence cannot accumulate into knowledge nor so much structure that evidence cannot in principle change the framework's conclusions.

1 Three Critiques, One Pattern

If Anyone Builds It, Everyone Dies [1] restates, in book-length and popular form, the Machine Intelligence Research Institute's long-standing position that sufficiently advanced artificial intelligence entails human extinction with near certainty, and that only an immediate, global halt to advanced AI development can avert it. Three published responses, arriving from quite different directions, are worth reading together rather than separately.

One review situates the book within a broader divide between incrementalist and radical positions on AI safety, arguing that the book's proposed remedy, an unconditional global development ban, is rhetorically forceful but politically unworkable, closer to a demonstration of conviction than a policy proposal capable of attracting a governing coalition [2]. A second review presses further, arguing that the book's load-bearing premise, that an intelligence explosion of the specific kind the argument requires is likely, receives only a few sentences of direct defense despite the AI field having changed enormously since MIRI's central arguments were first formulated, moving from hand-designed architectures to empirically characterized, scaling-law-governed deep learning systems; the review characterizes this as a regression, in which the book largely restates earlier

blogosphere-era arguments against a version of the field that no longer exists [3]. A third piece, not a review of the book at all but a general essay on cultural systems, argues that excessive convergence around a single interpretation erodes a system's adaptive capacity by eliminating the variation that would otherwise let it respond to conditions the dominant interpretation did not anticipate, a pattern observed across evolutionary lineages, market strategies, and cognitive styles alike [4].

Read independently, these are a political critique, an epistemic critique, and a general observation about cultural systems. Read together, they triangulate a single underlying property. A framework can be practically inert because it has stopped offering positions moderate actors can act on; it can be epistemically rigid because it treats new evidence as further confirmation rather than as a test; and it can be culturally brittle because it has suppressed the internal variation that would let it notice when its central interpretation has stopped fitting the evidence particularly well. These are not three separate flaws so much as three different vantage points on one flaw, the collapse of a framework's capacity to be moved by what it encounters.

2 Assimilation and Accommodation

The relevant distinction has a name in the study of how explanatory frameworks respond to new evidence, closely related to the classical observation that a theory can typically be preserved against apparently disconfirming evidence through suitable adjustment of auxiliary claims [8, 7], and to the historical description of how an established framework absorbs anomalies as a matter of course until their accumulation eventually forces a genuine restructuring [6]. Let H_t denote an explanatory framework at some point in its development and E_{t+1} a new piece of evidence. An assimilative update leaves the framework's basic structure intact and incorporates the observation into it,

$$(H_t, E_{t+1}) \longmapsto H'_t, \qquad H'_t \simeq H_t.$$

An accommodative update permits the evidence to alter the framework's structure itself,

$$(H_t, E_{t+1}) \longmapsto H_{t+1}, \qquad H_{t+1} \not\simeq H_t.$$

Both modes are ordinary and, within limits, appropriate; a framework that restructured itself in response to every minor observation would never accumulate stable knowledge, and the requirement that a scientific claim specify what would count against it [5] does not demand that every anomaly immediately overturn a working theory. The diagnostic question is not whether a given observation was assimilated, but whether the framework as a whole retains any accommodative capacity at all, whether some possible sequence of observations could, even in principle, restructure its central conclusion.

This is where the extinction framework becomes unusual, because nearly every category of contemporary observation about advanced AI systems has been read as confirming the same conclusion regardless of its content. Rapid capability improvement is read as evidence of accelerating danger. Documented deceptive behavior is read as evidence of danger. A system's demonstrated ability to recognize that it is being evaluated is read as evidence of danger, since it suggests concealment elsewhere. Interpretability tools failing to characterize a system's internal reasoning is read as evidence of danger, since it means oversight has failed; interpretability tools succeeding well enough to locate genuinely concerning internal structure is also read as evidence of danger, since it confirms the structure exists. A laboratory's substantial investment in safety research is sometimes read as tacit confirmation that insiders recognize the danger is real; a laboratory's comparatively light investment is read as evidence of recklessness. Expert alarm is read as evidence of danger. Expert

calm is available to be read as evidence that even well-placed observers are underestimating the problem. None of these readings is individually unreasonable. Taken together, they describe a framework for which it is difficult to specify any observation, across an unusually wide range of categories, that would not have been read as consistent with its central conclusion.

Proposition 1. *Let H be an explanatory framework and let $\mathcal{E}$ be a set of evidentiary categories such that, for every $E_i \in \mathcal{E}$, both an outcome e_i and its negation $\neg e_i$ are, upon occurrence, described as confirming H at comparable strength relative to a competing framework H_{alt}. Then H's likelihood ratio against H_{alt}, $\Lambda(E_i) = \Pr(E_i \mid H)/\Pr(E_i \mid H_{alt})$, is approximately 1 for every $E_i \in \mathcal{E}$, and the accumulation of evidence across $\mathcal{E}$, however large, does not discriminate H from H_{alt}.*

This is the same structural failure identified at the level of individual incidents elsewhere in this program, where a hypothesis able to read both compliance and defiance, both strong and weak performance, both visible and absent reasoning as confirmation loses the discriminating power that makes accumulated evidence meaningful. The present point is that the same failure operates one level up, not merely at the level of individual behavioral incidents but at the level of entire categories of evidence about the field: capability trends, safety behavior, interpretability results, and expert opinion, each category admitting the same collapse from a two-sided test into a one-sided confirmation.

3 A Viability Window for Explanatory Coherence

The appropriate response to this diagnosis is not to demand that a framework restructure itself on every occasion, since a framework with no persistent structure at all cannot accumulate knowledge across observations in the first place. A useful framework requires enough internal coherence to carry information from one observation to the next, formally analogous to a system whose internal order cannot fall arbitrarily low without losing the capacity to persist as a recognizable model at all. But a framework can also possess too much coherence, in the specific sense that its structure has become so tightly self-confirming that no realistic sequence of observations retains the capacity to revise it, formally analogous to a system whose internal order has become so high that it can no longer absorb genuinely new structure from its environment. Writing S for a framework's openness to restructuring by evidence, in a loose but useful sense, viability occupies neither extreme,

$$0 < S_{\min} < S < S_{\max}.$$

Too little openness and observations never cohere into a stable, communicable framework at all. Too much and the framework, however elaborate its internal apparatus, has stopped being genuinely tested by anything. The three critiques considered in Section 1 can be read as three independent symptoms of a framework operating too close to the upper bound: political inertness, since a framework immune to evidence cannot be moved by the objections of potential coalition partners either; epistemic non-update, since new developments in the field are absorbed without altering the framework's central claims; and cultural brittleness, since the internal variation that would let a research community notice when its dominant interpretation has stopped fitting the evidence has itself been suppressed by the framework's own coherence.

This yields a general principle applicable well beyond the present case:

a framework has genuinely encountered evidence only when that evidence could, in principle, have made the framework different.

A framework that can be shown, after the fact, to have been compatible with the observation that occurred is not thereby shown to have predicted it, and a framework compatible with every member of a wide class of possible observations has not been tested by any of them, however confidently its adherents describe the resulting confirmation.

4 Position in This Program

This essay is the structural counterpart to a companion piece in the same research program, which develops an equivalent argument at the level of individual behavioral incidents: a hypothesis able to read both an outcome and its negation as confirmation has a likelihood ratio near one for that incident, and has lost the capacity to be tested by it. The present essay extends that diagnosis from individual incidents to entire evidentiary categories, capability trends, safety investment, interpretability results, and expert sentiment, and argues that the resulting pattern is not a coincidence local to any one incident type but a structural property of how the framework as a whole relates to evidence. Three further companion essays in the same program address adjacent questions: whether an incident ledger's accumulated entries constitute independent evidence at all once exposure, detection effort, and counterfactual comparison are accounted for; whether individual incidents, apart from any question of framework-level coherence, are correctly described by the mechanisms their labels presuppose; and whether cognitive capability, even granted at its most alarming reading, implies the physical and social control the extinction scenario requires. None of these essays depends on accepting the others, and none requires disputing that the underlying incidents occurred or that some of them warrant serious engineering attention. What they jointly deny is that the volume of accumulated material, whatever its individual merits, has been shown to discriminate the extinction hypothesis from its nearest competitors, which is a separate and prior question from whether that hypothesis happens to be true.

5 Conclusion

The three reviews considered here were written independently and from different concerns, one political, one epistemic, one general and cultural, yet they identify the same underlying property once read together: a framework that has become capable of reading nearly any observation, and frequently its negation as well, as confirmation of the same conclusion. This is not evidence that the conclusion is false. It is evidence that the framework's relationship to evidence has shifted from a genuinely testable claim to a self-confirming interpretive structure, and that the appropriate response is neither to accept its confidence at face value nor to dismiss its content outright, but to ask, of any framework claiming this much support from this much evidence, what observation, if it occurred, would be permitted to count against it. A framework for which no good answer to that question is available has not demonstrated the danger it describes; it has demonstrated only its own coherence, which is a different and considerably less useful thing.

References

[1] E. Yudkowsky and N. Soares. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Little, Brown and Company, New York, 2025.

[2] S. Alexander. Book Review: *If Anyone Builds It, Everyone Dies*. *Astral Codex Ten*, September 2025. https://www.astralcodexten.com/p/book-review-if-anyone-builds-it-everyone.

[3] C. Collier. More Was Possible: A Review of *If Anyone Builds It, Everyone Dies*. *Asterisk Magazine*, Issue 11, 2025. https://asteriskmag.com/issues/11/iabied.

[4] R. Hanson. Beware Cultural Drift: Thoughts on Modernity's Monoculture Mistake. *Quillette*, April 2024. https://quillette.com/2024/04/11/beware-cultural-drift/.

[5] K. R. Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1959.

[6] T. S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1962.

[7] W. V. O. Quine. Two Dogmas of Empiricism. *The Philosophical Review*, 60(1):20–43, 1951.

[8] P. Duhem. *The Aim and Structure of Physical Theory*. Princeton University Press, Princeton, 1954. Originally published in French, 1906.