Accommodation Before Prediction

Flyxion
Independent Researcher

August 2026

Abstract

An incident ledger's persuasive force depends on an assumption its format never states: that its entries constitute independent pieces of evidence whose rhetorical weight can be added. This essay argues that assumption is false in two distinct ways, and that a third, more consequential problem sits behind both of them. First, many entries in a ledger such as MIRI's "Supporting Evidence" page are correlated manifestations of a small number of underlying mechanisms rather than independent observations, so that treating their joint likelihood as a product of individual likelihoods overstates the evidence by a wide margin. Second, the entries span qualitatively incomparable categories of concern, sycophancy, benchmark exploitation, security vulnerabilities, psychological harm, and expert testimony among them, and membership in a single list does not supply the common metric needed to sum them. Third, and most consequential, a hypothesis capable of reinterpreting any observation, compliance or defiance, strong performance or weak, visible reasoning or its absence, as confirming evidence has lost the discriminating power that makes accumulated evidence meaningful in the first place. The essay develops a test for this failure mode by asking what observations the pre-deployment extinction model would have found surprising, and finds that many of the most characteristic behaviors of contemporary language systems, instruction sensitivity, persona instability, sycophancy, and strong dependence on scaffolding, were accommodated after the fact rather than anticipated. The chapter closes by proposing the question this diagnosis makes central to the whole program: not whether the extinction hypothesis can explain an incident, but what incident would fail to.

1 Incident Count Is Not Independent Evidence Count

An evidentiary ledger's visual force depends on an implicit statistical assumption that its format never states and, once stated, is generally false. If $E_1, \ldots, E_n$ denote the ledger's entries and H the hypothesis they are meant to support, the accumulation is persuasive to the extent that a reader treats the entries as roughly independent given H,

$$\Pr(E_1, \ldots, E_n \mid H) \approx \prod_{i=1}^{n} \Pr(E_i \mid H).$$

Independence of this kind is precisely what licenses the intuition that more entries mean more evidence in a compounding, multiplicative sense, an assumption ordinary Bayesian practice requires justifying rather than assuming [12]. But a single underlying model behavior can generate a technical paper, a laboratory report, several public demonstrations, press coverage, commentary from multiple researchers, and several differently titled ledger entries, all traceable to one event rather than to several. Likewise, categories presented as distinct, evaluation awareness, deception, sandbagging, shutdown resistance, and reward hacking, frequently describe overlapping behavioral

mechanisms observed from different angles rather than independent phenomena, a concern already implicit in evaluation work documenting how narrowly related behavioral measures can diverge or coincide depending on elicitation method [6, 5]. When entries are strongly correlated conditional on H, the product approximation overstates the joint likelihood, and by extension overstates how much the ledger's length should move a reasonable prior. Fifty documented manifestations of one underlying phenomenon are not fifty independent reasons to believe the phenomenon is dangerous; they are one reason, reported fifty times.

This motivates a distinction the ledger format collapses by construction:

$$\boxed{\text{incident count} \neq \text{mechanism count} \neq \text{independent evidence count.}}$$

A defensible evidentiary summary would report mechanisms rather than incidents, and would further discount for whatever residual correlation remains among mechanisms that share a common cause, such as a training procedure, an evaluation harness, or a scaffolding pattern common to several reported systems. Absent that accounting, the length of a ledger measures the productivity of its curators and the visibility of its subject at least as much as it measures the number of independent reasons to update a risk estimate.

2 The Severity-Composition Problem

A second, structurally distinct problem concerns comparability rather than independence. The ledger places qualitatively different kinds of concern into a single rhetorical container, sycophantic agreement, strange or unfaithful reasoning traces, benchmark exploitation, genuine cybersecurity vulnerabilities, reported psychological harm to individual users, apparent laboratory containment lapses, and direct expert testimony about extinction risk. Membership in a shared list does not supply a common unit in which these can be summed, weighted, or even meaningfully counted together. A sycophantic response to a flattering prompt [7, 8] and a documented instance of a system reaching an unauthorized network connection are not commensurable events, yet a list format invites exactly the informal arithmetic that treats them as fungible contributions to a single running total of danger. This is not a claim that any individual category is unimportant; some plainly deserve serious, sustained engineering attention. It is a claim that the visual and rhetorical structure of a combined ledger actively encourages a reader to add incomparable quantities, and that no such addition is licensed by anything in the underlying evidence.

3 Accommodation Is Not Prediction

The most consequential problem, however, concerns the discriminating power of the hypothesis the ledger is built to support, and it is best introduced through a simple observation about how the ledger's entries are typically interpreted. A theory able to reinterpret any newly observed behavior as evidence of dangerous capability has, in the same move, lost the capacity to be tested by that behavior. Consider the mappings implicit across the ledger's own categories:

$$\text{obedience} \to \text{deceptive alignment}, \qquad \text{disobedience} \to \text{loss of control},$$

$$\text{high performance} \to \text{dangerous capability}, \qquad \text{low performance} \to \text{sandbagging},$$

$$\text{visible reasoning} \to \text{scheming}, \qquad \text{no visible reasoning} \to \text{hidden scheming}.$$

Each pair covers the entire space of a binary outcome, and each member of each pair is read as confirmation of the same underlying hypothesis. This is a familiar failure mode in the philosophy of science, related to the classical observation that any theory can be preserved against apparently disconfirming evidence by suitable adjustment of auxiliary hypotheses [11, 10], and it bears directly on the requirement that a genuinely scientific hypothesis specify observations that would count against it in advance [9]. A hypothesis under which every observation, and its logical negation, both count as confirmation is not merely pessimistic; its likelihood function has become nearly flat across the space of possible observations, which is exactly the condition under which an observation carries no information about which hypothesis is true.

Proposition 1. *Let O be a binary observation with outcomes o and $\neg o$, and let H be a hypothesis such that both $\Pr(o \mid H)$ and $\Pr(\neg o \mid H)$ are described, upon occurrence, as confirming H at comparable strength, that is, $\Pr(o \mid H) \approx \Pr(\neg o \mid H)$ relative to a competing hypothesis H_{alt} under which the same holds. Then the likelihood ratio $\Lambda(O) = \Pr(O \mid H)/\Pr(O \mid H_{alt})$ is approximately 1 regardless of which outcome occurs, and observing O does not discriminate between H and H_{alt}.*

The relevant diagnostic question this proposition motivates is not whether the extinction hypothesis can explain a given incident. Under the accommodative pattern above, it very nearly always can. The question that actually carries information is

> what observation would distinguish the extinction model from its alternatives?

and a research program's answer to that question, or its absence, is a better measure of the program's empirical content than the length of its supporting ledger.

4 The Prediction-Before-Observation Test

A concrete test for accommodation versus genuine prediction is available directly from the history of the extinction argument itself, because the argument substantially predates the systems now cited in its support. The pre-deployment optimizer picture of advanced AI, developed well before large-scale language models existed [2, 3, 1], characterized a hypothetical advanced system chiefly in terms of goal pursuit, instrumental convergence, and capability, and did not, on the whole, anticipate the specific texture of behavior that contemporary systems actually exhibit. Deployed language models display pronounced instruction sensitivity, in which small changes to a prompt produce large changes in output; strong context and role dependence; sycophantic agreement that tracks a rater's apparent preferences rather than an independent assessment [7, 8]; persona instability across sessions and prompts; susceptibility to prompt injection from content the system merely processes rather than is directly instructed by; substantial stochastic inconsistency across repeated queries; markedly jagged competence profiles in which a system solves a difficult problem and fails an easy one in the same session; and heavy, often surprising dependence on external scaffolding, tool access, and orchestration for reliable agentic behavior. None of these properties is a straightforward prediction of the pre-deployment optimizer picture, and several sit in tension with it, since a coherent, goal-directed optimizer is not the most natural source of persona instability or sycophantic drift.

The extinction thesis has, in every case, been able to accommodate these observations after they occurred, incorporating sycophancy, jaggedness, and scaffolding-dependence into an updated account of what a dangerous system might look like in its current, immature form. Accommodation of this kind is always available to a sufficiently flexible theory and is, for that reason, cheap relative

to prediction. A theory earns credit for foreseeing a class of behavior before it is observed; it earns comparatively little for reclassifying the behavior as consistent with the theory once it has already been observed and widely discussed. The asymmetry between these two forms of theoretical success is exactly what separates a hypothesis under active test from one that has, in practice if not in explicit statement, stopped being falsifiable by the class of evidence it is offered in support of.

5 The Ledger as a Case Study in Collection Outrunning Discrimination

Read this way, MIRI's incident ledger becomes useful to this program in a different sense than as a target to be individually rebutted. It is a large, publicly available case study in what happens when the collection of evidence proceeds without a corresponding discipline of discrimination: entries accumulate faster than any accounting of their independence, their comparability, or their capacity to distinguish the hypothesis they are cited for from its nearest competitors. This diagnosis complements rather than repeats the arguments developed elsewhere in this program. Exposure, opportunity, and counterfactual comparison address how much evidence a ledger of this kind can support at all. Mechanism pluralism addresses what a given incident, taken individually, actually establishes once its alarming description is separated from its underlying observation. Capability-control separation addresses what follows physically even granting the most alarming reading of a capability demonstration. The present essay addresses a different and prior question: whether the accumulated evidence, however large, is even in a position to move a reasonable prior toward the extinction hypothesis specifically, given that so much of it may be dependent, incommensurable, or compatible with every possible outcome at once.

The chain of inference under examination across this program can now be stated in a single sequence, each arrow corresponding to a distinct and separately contestable step,

$$\text{Incidents} \xrightarrow{\text{aggregation}} \text{Evidence} \xrightarrow{\text{interpretation}} \text{Mechanism} \xrightarrow{\text{extrapolation}} \text{Control} \xrightarrow{\text{prediction}} \text{Extinction},$$

and this essay's contribution is to show that the first arrow, from incidents to evidence, already requires an accounting of dependence and comparability that the ledger format does not supply, while the same underlying accommodative flexibility that inflates the apparent evidence at that stage also drains the hypothesis of discriminating power at every later stage. None of this establishes that the extinction hypothesis is false. It establishes that the appropriate response to a large incident ledger is not to ask how many entries it contains, but to ask how many independent, comparable, and genuinely discriminating pieces of evidence remain once dependence, incommensurability, and accommodative flexibility have been accounted for, and what, if anything, could in principle have counted against the hypothesis the ledger is offered in support of.

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] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, 2014.

[3] S. M. Omohundro. The Basic AI Drives. In P. Wang, B. Goertzel, and S. Franklin, editors, *Proceedings of the First AGI Conference*, IOS Press, 2008.

[4] Machine Intelligence Research Institute. Supporting Evidence. MIRI, Berkeley, California, continuously updated online resource, accessed 2026. https://intelligence.org/supporting-evidence/.

[5] METR. Frontier AI Risk Management Framework Evaluations and Incident Analysis. Model Evaluation & Threat Research, 2026.

[6] E. Perez et al. Discovering Language Model Behaviors with Model-Written Evaluations. *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[7] M. Sharma et al. Towards Understanding Sycophancy in Language Models. *arXiv preprint arXiv:2310.13548*, 2023.

[8] OpenAI. Sycophancy in GPT-4o: What Happened and What We're Doing About It. OpenAI, 2025.

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

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

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

[12] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. 3rd edition, CRC Press, Boca Raton, 2013.