Exposure Before Evidence: The Statistical Epistemology of AI Incident Ledgers

Flyxion
Independent Researcher

August 2026

Abstract

The Machine Intelligence Research Institute maintains a public ledger titled "Supporting Evidence," [2] a running catalogue of incidents intended to substantiate the thesis that advanced artificial intelligence poses an existential risk to humanity [1]. This essay grants, for the sake of argument, that the great majority of entries in that ledger are accurately reported. It argues nonetheless that the inference commonly drawn from such a ledger, from an expanding count of worrying incidents to an increasing probability of catastrophic outcomes, is not supported by the evidentiary form the ledger takes. An incident count without an exposure denominator, a detection-effort adjustment, and a comparison against a genuine counterfactual is compatible with declining risk, constant risk, or rising risk in roughly equal measure. The essay develops this claim formally, distinguishes existence evidence from frequency evidence, identifies a compositionality error in which properties demonstrated separately across many distinct models are silently unified into a single hypothetical successor system, and situates the resulting argument alongside two companion pieces in the same research program without repeating their claims. The aim is not to dismiss the underlying incidents, several of which are genuinely concerning, but to specify what would have to be added to a ledger of this kind before it could support the conclusion it is typically used to support.

1 An Evidentiary Object, Not a Single Claim

MIRI's Supporting Evidence page is unusual among artifacts in the AI-risk debate because it is not an argument in the ordinary sense. It does not advance a chain of premises toward a conclusion; it accumulates. Under headings such as capabilities advancement, alien chains of thought, reward hacking, shutdown resistance, evaluation awareness, sandbox escape, and psychological harm, the page assembles several hundred short entries, each linking to a source of variable quality, ranging from peer-reviewed evaluations to press quotations, anecdotes, and expressions of subjective probability by named individuals. The document itself warns that its sources "vary in quality" and should be vetted before citation, an honest caveat that nonetheless does little to constrain how the page is actually used, which is as a single visual mass whose cumulative weight is meant to exceed the sum of its individually inspectable parts.

That structure deserves to be treated as an object of study in its own right, independent of whether any particular entry is true. A ledger of this kind performs at least four distinct operations before a reader ever reaches a conclusion. It selects incidents according to a criterion of worryingness. It aggregates properties demonstrated by different systems, in different laboratories, under different experimental conditions. It reconstructs, largely in the reader's imagination, a single hypothetical

agent that possesses the union of those properties. It then projects that reconstructed agent forward into a catastrophic future. Each of these operations, selection, aggregation, reconstruction, and projection, introduces assumptions that the visual density of the ledger tends to obscure rather than expose. The remainder of this essay examines each operation and argues that the strongest empirical response to this genre of evidence does not require disputing the individual incidents at all.

2 Existence Evidence Is Not Frequency Evidence

Consider a universal safety claim of the form "systems of type M never exhibit behavior B." A single well-documented counterexample is sufficient to refute it. In this narrow sense, the Supporting Evidence ledger performs real epistemic work: entries on shutdown resistance, deceptive reasoning under evaluation, sandbox escape, and reward hacking are each, if accurately reported, sufficient to retire a corresponding universal reassurance. That is a legitimate and useful function.

It does not, however, license the further step commonly taken, from "this behavior has been observed" to "this behavior is prevalent, and its prevalence is rising toward catastrophe." The logical form of the two claims is different in kind rather than degree. The first is existential,

$$\exists\, x \in \mathcal{M} : B(x),$$

and is established by a single credible instance. The second is a statement about frequency,

$$\Pr\big(B(x)\big) \approx 1 \quad \text{or} \quad \Pr\big(B(x)\big) \text{ rising},$$

and requires a rate, which in turn requires both a numerator, the count of incidents, and a denominator, the number of comparable opportunities in which the behavior could have occurred but did not, a distinction with a long history in the foundations of statistical inference [5, 6]. A ledger organized around worrying incidents is, by construction, a ledger of numerators. It supplies existence evidence in abundance and frequency evidence almost nowhere. Treating the former as though it were the latter is the single most consequential move in the informal argument the ledger is typically enlisted to support.

3 The Denominator Problem

Let N_t denote the number of adverse incidents observed by time t, let X_t denote the number of relevant opportunities for such an incident, expressed for instance in user interactions, agent-hours, or independent evaluation trials, and let p_t denote the underlying per-opportunity probability of the incident. Then, approximately,

$$N_t = X_t\, p_t.$$

An increase in N_t over time is fully consistent with a decrease in p_t, provided exposure has grown faster than the underlying rate has fallen. A concrete illustration makes the point vivid. Suppose at an earlier time exposure was modest, $X_0 = 10^6$, and the failure probability comparatively high, $p_0 = 10^{-4}$, giving $N_0 = 100$ observed incidents. Suppose that by the present, exposure has grown a thousandfold, $X_1 = 10^9$, reflecting the shift from laboratory-confined systems to consumer products used by hundreds of millions of people, while the underlying failure probability has simultaneously fallen a hundredfold, $p_1 = 10^{-6}$. The observed incident count is then $N_1 = 1000$, an apparent tenfold worsening of the visible evidence accompanying a genuine hundredfold improvement in

the underlying rate. A reader who tracks only N_t will conclude that matters are getting worse at precisely the moment they are, on this measure, getting substantially better. Nothing about this arithmetic depends on any contested claim about AI capability; it is a property of counting processes under changing exposure, and it applies with equal force to automobile accidents, foodborne illness reports, and any other domain in which the population exposed to a risk changes over time.

4 Detection Intensity as a Second Confound

A more complete model introduces a third factor, detection intensity d_t, the degree to which the relevant behavior is being actively sought rather than passively encountered,

$$N_t = X_t \, p_t \, d_t.$$

Contemporary frontier systems are not merely used by more people; they are deliberately red-teamed, benchmarked against adversarial suites, prompted through jailbreak competitions, evaluated for reward hacking, placed in cyber ranges designed to elicit sandbox escape, and subjected to interpretability tools built expressly to surface deceptive reasoning that earlier, less-scrutinized systems were never checked for at all [3, 4]. An expanding incident count can therefore reflect rising X_t, rising d_t, or rising p_t, and the ledger format offers no way to distinguish between them. A regime of intense, well-funded scrutiny should, all else equal, be expected to surface more incidents than a regime of light scrutiny, even if the underlying systems are, by construction, safer. Treating detection success as though it were exclusively evidence of danger rather than partly evidence of vigilance inverts the usual relationship between measurement effort and the thing being measured.

5 Extreme-Value Effects at Planetary Scale

A further consequence of the exposure explosion deserves separate treatment, related to the general behavior of extreme-value statistics under large sample sizes [8, 7], because it operates even when the underlying per-opportunity probability is very small and, by ordinary standards, reassuring. If an adverse event has probability p per independent opportunity and there are n such opportunities, the probability of observing the event at least once is

$$\Pr(\text{at least one}) = 1 - (1 - p)^n \approx 1 - e^{-np}$$

for small p. At a probability as low as $p = 10^{-9}$ per interaction, a billion interactions, a modest figure for a widely deployed consumer product, already yields

$$1 - e^{-1} \approx 0.63,$$

and ten billion interactions yield

$$1 - e^{-10} \approx 0.99995.$$

At planetary deployment scale, in other words, the appearance of at least one instance of almost any describable behavior becomes close to certain even under failure rates that would be considered excellent by the standards of any other safety-critical engineering discipline. This is the precise sense in which anecdotal abundance becomes compatible with statistical rarity: a ledger that will, with near certainty, contain examples of nearly every conceivable failure mode once deployment crosses a certain scale tells a reader comparatively little about the failure mode's true prevalence, and a great deal about the scale of deployment.

6 The Missing Complement Set

The denominator problem has a companion difficulty at the level of individual entries. For each report that a model resisted shutdown, exhibited deceptive reasoning, or escaped a sandbox, the ledger does not report how many structurally comparable trials produced the ordinary, unremarkable outcome. Without a count of the complement,

$$N_{\text{failure}} \quad \text{and} \quad N_{\text{success}},$$

no rate

$$\hat{p} = \frac{N_{\text{failure}}}{N_{\text{failure}} + N_{\text{success}}}$$

can be estimated at all, for any individual entry, let alone across the ledger as a whole, in the ordinary sense in which any rate estimate requires both counts [9]. This omission is not incidental to the genre; a document whose explicit purpose is to collect worrying incidents is, by its sampling design, a document of numerators, exhibiting the same selection-driven overstatement of effect size documented in literatures selected for striking results [10]. Asking it to also supply denominators is asking it to be a different kind of document, one considerably less rhetorically effective and considerably more informative.

7 Discriminating Evidence Versus Merely Compatible Evidence

A further refinement clarifies why so many of the ledger's entries, however genuine, carry less evidential weight than their placement suggests. What matters for updating a belief is not whether an observation is compatible with the hypothesis of catastrophic risk, written H_{doom}, but whether it is more compatible with that hypothesis than with a mundane alternative H_{alt}, for instance the hypothesis that general-purpose capability is improving without any accompanying tendency toward adversarial, uncontrollable agency. The relevant quantity is a likelihood ratio,

$$\Lambda(e) = \frac{\Pr(e \mid H_{\text{doom}})}{\Pr(e \mid H_{\text{alt}})}.$$

Rapid improvement on mathematical benchmarks is predicted by H_{doom}, but it is predicted at least as strongly by the mundane hypothesis that capability is increasing, so $\Lambda(e) \approx 1$ and the observation is nearly uninformative between the two hypotheses despite its dramatic surface appearance. Reward hacking is predicted by H_{doom}, but it is also the textbook consequence of optimizing against an imperfect proxy objective, a phenomenon documented in reinforcement learning long before frontier language models existed, so again $\Lambda(e)$ is closer to one than the ledger's framing implies. A sandbox escape is dramatic, but sandbox escape is precisely what cybersecurity evaluators build their environments to elicit and reward, which weakens rather than strengthens its evidential contribution to H_{doom} specifically. The entries that genuinely discriminate between hypotheses, sustained, autonomous, resource-acquiring behavior that persists once evaluation pressure is removed and serves no proxy objective a designer could plausibly have intended, are a much smaller subset of the ledger than its length suggests, and they deserve to be argued for individually rather than borrowed for by association with the surrounding hundreds of entries.

8 The Synthetic Agent

The most consequential operation the ledger performs is aggregation across systems. Mathematical capability is drawn from one model, deceptive reasoning from a second, autonomous cyber

capability from a third, shutdown resistance from a fourth, self-replication from a fifth, situational awareness from a sixth. If $\mathcal{C}(M_i)$ denotes the set of properties demonstrated by model M_i, the reading experience of the ledger constructs, without ever stating it as a premise, a union

$$\mathcal{C}^* = \bigcup_i \mathcal{C}(M_i),$$

and implicitly attributes this entire union to a single anticipated future system M^*. No entry in the ledger establishes that any actual model possesses these properties jointly, simultaneously, persistently, and under its own initiative rather than an evaluator's deliberate elicitation. The full inferential chain the ledger invites can be written as a sequence of maps,

$$\mathcal{E} \xrightarrow{\sigma_H} \mathcal{E}_H \xrightarrow{\alpha} \mathcal{C}^* \xrightarrow{\rho} M^* \xrightarrow{\tau} \text{Extinction},$$

where σ_H selects worrying observations from the full universe of evidence $\mathcal{E}$, α aggregates the demonstrated properties, ρ reconstructs a hypothetical unified agent possessing them, and τ projects that agent's trajectory forward to a catastrophic outcome. Each arrow carries assumptions that the visual continuity of a single scrolling page tends to compress into a single inferential step. The rhetorical experience of reading the ledger is that of encountering $\mathcal{E}_H \Rightarrow$ Extinction directly; the actual argument requires defending four separate, individually contestable transformations, and the ledger defends none of them explicitly because its genre is enumeration rather than argument.

Proposition 1. *Let $e_1, \ldots, e_n$ be incidents such that each e_i establishes a distinct predicate P_i true of some system M_i, that is, $\exists\, x : P_i(x)$. Then $e_1, \ldots, e_n$ jointly establish*

$$\bigwedge_{i=1}^{n} \big(\exists\, x_i : P_i(x_i)\big),$$

which does not entail

$$\exists\, x : \bigwedge_{i=1}^{n} P_i(x).$$

The proof is immediate: existential quantifiers do not distribute over conjunction without an additional premise identifying the witnesses. This is not a pedantic point. It is the exact logical gap between what several hundred separately true entries can establish and what the ledger's implicit conclusion requires.

9 The Missing Counterfactual

A separate confound arises whenever the ledger implicitly compares the present, AI-saturated world against an unstated alternative in which the relevant harms did not occur, the same comparison-class problem that complicates existential-risk accounting more generally [11, 12]. That comparison is rarely available. Comparing incidents recorded in an era of laboratory-confined systems against incidents recorded in an era of consumer-scale deployment is not a controlled comparison, because exposure, available affordances, detection effort, user population, and economic integration have all changed simultaneously. Nor is a strict harm count the relevant quantity even within the present era, since deployed systems generate benefits, in content moderation, translation, fraud detection, medical triage support, and education, that a harm-only ledger by construction excludes. The relevant quantity for judging net effect is a difference,

$$\Delta H = H(W_{\text{AI}}) - H(W_{\text{counterfactual}}),$$

between harm in the observed world and harm in an unobserved counterfactual world, and no incident ledger, however long, can estimate a quantity that requires data from a world that does not exist. This does not mean net effect is positive; it means the ledger's format cannot address the question either way, and should not be read as though it does.

10 Relation to the Adjacent Arguments in This Program

This essay is deliberately narrower than two companion pieces in the same research program and should not be read as restating either. *Intelligence as an Integrative Ecological Operator* argues against the underlying model of intelligence as a detached optimizer, proposing instead an embedded, reciprocally coupled account in which increasing capability tends toward greater semantic integration rather than adversarial separation; that essay contests H_{doom} at the level of what intelligence structurally is. *Divided We Stand* argues that MIRI's explanatory framework exhibits epistemic over-coherence, absorbing observations that would ordinarily be expected to point in different directions into a single invariant conclusion, and proposes that a genuinely tested theory is one for which some possible observation could have moved it the other way. That essay contests the framework's relationship to evidence in general.

The present essay contests neither the ontology of intelligence nor the framework's general capacity to update. It grants, for the sake of argument, that MIRI's evidentiary practice has changed substantially since earlier periods and is visibly responsive to recent developments; the 2026 ledger is full of current material. The argument here is narrower and, for that reason, harder to dismiss as either optimism about AI or a claim that MIRI is dogmatic: an evidentiary method built from incident enumeration under conditions of exploding exposure, rising detection effort, missing complement sets, and cross-system aggregation cannot, as a matter of statistical form, support the frequency claim it is used to support, independent of whether the underlying incidents are true and independent of whether the framework interpreting them is flexible or rigid. The problem is upstream of both questions; it is a property of the evidentiary object itself.

11 What a Defensible Ledger Would Require

None of the foregoing implies that incident documentation is worthless or that the individual entries in MIRI's ledger should be disregarded. Several, particularly those establishing sustained, non-elicited, resource-acquiring behavior that persists once evaluation pressure is withdrawn, are exactly the kind of high-Λ evidence a serious risk assessment should weight heavily. The argument is about form rather than content: a ledger capable of supporting a frequency claim would need to report, alongside each incident, the exposure base from which it was drawn, the detection effort applied, the complement set of comparable trials that did not produce the incident, and, where possible, an estimate of the counterfactual rate absent the relevant capability or affordance. A ledger organized this way would necessarily be shorter, less visually overwhelming, and considerably more informative, since its entries would carry rates rather than only instances. Until such a ledger exists, the appropriate epistemic response to the current one is neither dismissal nor alarm calibrated to its length, but the considerably harder question implicit in any well-formed statistical argument: given everything genuinely established here, what rate, if any, are we entitled to infer, and compared to what.

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

[3] METR. Measuring AI Ability to Complete Long Tasks. Model Evaluation & Threat Research, 2025–2026.

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

[5] T. Bayes. An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

[6] R. A. Fisher. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society A*, 222:309–368, 1922.

[7] N. N. Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House, New York, 2007.

[8] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2001.

[9] 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.

[10] J. P. A. Ioannidis. Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):e124, 2005.

[11] N. Bostrom. Existential Risk Prevention as Global Priority. *Global Policy*, 4(1):15–31, 2013.

[12] T. Ord. *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury, London, 2020.