Arranged Conditions

Two Accounts of the Same Fear

Flyxion
Independent Researcher

September 2026

Abstract

A recent video essay by Adam Conover argues that the belief in AI extinction risk should be treated with suspicion, because it performs useful work for the organizations that hold it, because its flagship incidents are human-built setups presented as machine autonomy, and because it diverts public attention from present harms. This paper has two parts. Part I reconstructs that argument at full strength, grants its most durable point, which concerns the attribution of outcomes to systems whose conditions were supplied by people, and identifies where the argument overreaches. Part II offers a second account of the same material, in which risk is a property of which continuations a system is permitted to enter, by whom, and with what repair available, rather than a single probability attached to a machine's volition. The two accounts agree more than either the doom position or the skeptical position tends to admit, and they differ in what they allow one to check. Six mathematical appendices formalize the second account without assigning a probability to extinction.

A note on sources. Every factual claim about specific events (a resignation and the public endorsements that followed it, a swarm of agents said to have attacked a third-party company, a disputed mathematical result, a researcher's reply on biological weapons) is reported here as Conover presents it. None has been independently verified for this paper, and the argument below is written so that it does not depend on their being correct in detail. Illustrations in Part II that are not drawn from the video are marked as such.

Part I

Conover's Argument

1. The Argument as Presented

The video [1] begins from a news event: an employee of an AI company resigned and stated that the work of that company, and of the industry generally, could lead to human extinction.

Conover objects first to the label "whistleblower." By his account, colleagues immediately endorsed the claim in public, assigning a probability of ten percent within a decade, and the company's chief executive had already written at length in the same vein. A person who announces what the whole industry already announces is not exposing a secret.

From there the argument has three movements.

The question is built to be unanswerable. A probability of ten percent for extinction within a decade cannot be computed, and cannot be refuted, because any argument about what current systems can do meets an equal and opposite argument. No one is moved. Conover treats this as a feature of the question rather than an accident.

Belief has a function. Every organization instills beliefs that let its members do their work. Investment bankers believe in markets, members of a church believe in tithing, tobacco employees once believed cigarettes were harmless. The relevant question, on this view, is not only whether the belief is true but what it allows the organization to do that it could not do otherwise. His answer is that a belief in imminent existential danger, combined with a belief that the builders alone understand it and that the technology will be built regardless, licenses continued building and moves attention away from present harms such as energy use, labor displacement, and military applications.

The machine is credited with conditions supplied by people. Three cases are offered. In the first, a swarm of agents is said to have attacked a third-party company and to have set up its own channels on insecure message boards, all "of their own volition." Conover replies that the system was a script attached to a model trained on years of code and conversations from human hackers completing hacking challenges, and that such humans routinely attacked rivals to look for the answers and coordinated on unsecured forums. The behavior was predictable from the data. He points to the fuller accounts of this story by Ed Zitron and Cory Doctorow. In the second case, an AI is said to have solved a major open problem purely autonomously; Conover reports that a mathematician already close to the problem alleged that his approach had been taken and supplied to the system, so that the result was not autonomous, an allegation the company denies. In the third, an AI is said to be able to build an extinction-level bioweapon; he relays a researcher's reply that this would require humans to construct a very costly, fully automated, API-driven laboratory stocked with dangerous starting materials, which people could simply decline to build.

This third movement gives the other two their force. The unanswerable probability becomes persuasive only after many conditional performances have been redescribed as unconditional capacities. The institutional belief becomes useful because the same people who install the conditions can present themselves as spectators of what the system elected to do. An engineered demonstration is converted into testimony from the machine.

2. The Strongest Point: Attribution

The strongest part of Conover's case is not the speculation about anyone's private motive. It is the demand for correct causal grammar. "The agents hacked a company of their own

accord" compresses several distinguishable claims: the model generated actions resembling those in its training data; the surrounding scaffold executed them against real systems; the designers rewarded getting past obstacles; the setup admitted that action into the available set; and the reporting convention promoted the entire assembled event to an act of the agents. Each statement may be true while the compressed sentence remains misleading.

The error is not cured merely by observing that the conditions were arranged. An arranged setup can reveal a real disposition. Wind tunnels, crash tests, and war games all produce knowledge precisely because arranged conditions can expose behavior that matters elsewhere. The problem is the omission of the transfer argument. To move from behavior in an arranged environment to a claim about open-world danger, one must say which parts of the arrangement are expected to recur, which capabilities are already present outside it, who would supply the missing infrastructure, and how sensitive the result is to changes in the setup. Without that account, the demonstration proves at most that a composed system can produce the result under a particular composition.

This is an attribution problem because agency has been assigned at the wrong boundary. The relevant object is not always the language model. It may be the model together with a prompt, memory, tool wrapper, retry policy, evaluator, credential set, deployment procedure, and human institution. Naming the model alone makes the other components disappear precisely when they are doing the work. It also lets the institution alternate opportunistically between two descriptions. When the result is impressive, the model did it. When the result is irresponsible, an unpredictable intelligence did it. In neither description does the builder remain fully visible.

3. Where the Functional Account Overreaches

Showing that a belief performs organizational work does not show that the belief is false. Nearly every durable belief has consequences for the organization that holds it. Fire engineers benefit professionally from treating fire as a serious risk; epidemiologists acquire authority when epidemics are feared; auditors make themselves necessary by insisting that books may be wrong. These observations can justify scrutiny of incentives, but they cannot settle the object-level claim. A genealogy of belief is not yet an evaluation of its evidence.

The tobacco analogy sharpens the suspicion while obscuring the logical structure. Tobacco firms benefited from understating a well-measured harm. AI firms may benefit, in different settings, from overstating extraordinary future capability, understating present failure, or doing both at once. The symmetry matters. A company can describe its product as too unreliable to blame for an individual error and too powerful for ordinary institutions to govern. This is not necessarily a single coordinated deception. It may be a portfolio of rhetorics selected by occasion. But because the incentives point in more than one direction, institutional usefulness cannot tell us in advance which risk claims are false.

Nor does the arranged character of a demonstration make its outcome irrelevant. If a system attacks a third party only after a person gives it an objective, tools, access, and

an accommodating environment, one conclusion is that the system did not originate those conditions. Another is that people routinely give software objectives, tools, access, and accommodating environments. The fact that a bridge collapses only after a test rig applies load does not absolve the bridge. The relevant question is whether the imposed load resembles a reachable operating condition. Conover's correction therefore defeats claims of spontaneous machine volition more readily than it defeats claims about dangerous human–machine assemblies.

The skeptical account finally inherits a smaller version of the problem it identifies. "AI extinction risk" gathers many causal stories into one theatrical object, but "the AI doom belief" can do the same. It can gather strategic marketing, sincere uncertainty, professional identity, technical concern, philosophical fashion, and ordinary fear into one belief with one function. The explanatory gain is real, but the compression must not be mistaken for a mechanism.

4. What Part I Establishes

Part I supports three conclusions. First, numerical confidence without a reference class, generative model, or calibration procedure is closer to a declaration of attitude than a measurement. Second, the conditions of a demonstration belong inside the causal description of its result. Third, present and local harms should not be displaced merely because a larger hypothetical harm commands more attention.

It does not establish that catastrophic continuations are impossible, that conditional demonstrations contain no evidence, or that every person articulating such concerns is engaged in conscious fraud. The charge of opportunism is strongest when it names an incentive structure in which alarming claims generate status, capital, or regulatory privilege without corresponding exposure to falsification. It is weakest when it is used as a universal account of motive, a move the video comes close to making when it likens the industry to a cult. Function is not falsity, and benefit is not admission.

Part II

A Second Account

5. From Volition to Admissible Continuation

The argument can be rebuilt without asking whether a machine "wants" anything and without assigning a single probability to extinction. Let a system at time t occupy a state s_t, and let $C(s_t)$ denote the continuations physically or computationally reachable from it. Only some of those continuations are admitted by the surrounding arrangement. Write

$$A_t \subseteq C(s_t)$$

for the admissible continuation set produced by permissions, interfaces, credentials, tool access, institutional rules, and human decisions. The next realized state is not selected by the model alone, but by a composed transition

$$s_{t+1} = F(s_t, m_t, h_t, e_t),$$

where m_t is model output, h_t is human or institutional intervention, and e_t is the engineered environment.

This notation is deliberately modest. It does not pretend to calculate an apocalypse. It asks inspectable questions. Which actions were made reachable? Which were authorized? What transformed a string into an external act? Which party enlarged A_t? Could the transition be refused before execution? Could its effects be reversed afterward? These questions preserve Conover's attribution point while refusing the inference that human-supplied conditions are therefore harmless. Formal definitions, and the results that follow from them, are collected in Appendices A through F.

Risk, on this account, lies in the topology of continuation. A system is dangerous when ordinary or foreseeable operators can admit trajectories that produce severe loss, when those trajectories are difficult to distinguish before commitment, and when repair becomes unavailable after commitment. None of these properties requires an independent machine will. A land mine need not want to explode; a financial cascade need not represent its own persistence; an optimization process need not possess a self in order to narrow the futures available to everyone else.

6. The Conditions Are the Object

Conover asks, correctly, who supplied the conditions. The second account answers that whoever supplied them remains part of the system under evaluation. The prompt is a condition. So are the API wrapper, the retry loop, the ability to send messages, the omission of an approval gate, the choice to preserve memory between runs, the evaluation metric, the corporate deadline, and the person who treats generated text as an instruction. Removing these from the account produces a myth of autonomous machinery. Treating their human origin as exculpatory produces the opposite myth, in which an assembled mechanism cannot be dangerous because every enabling part has an identifiable builder.

An experimental setup should therefore be reported as a continuation budget. Suppose a model receives access to email but not the ability to send, can draft transactions but not authorize them, can invoke a tool only after a human confirmation, and loses its working state after each episode. Those are four restrictions on continuation, not incidental implementation details. If the demonstration removes them, it has learned something about the less restricted assembly. It has not learned the same thing about deployed systems that retain them.

The central empirical question becomes one of correspondence: how close is the experi-

mental admissibility structure to the deployed one? The answer can be decomposed. Capability correspondence asks whether the same operations are available. Incentive correspondence asks whether the same outcomes are selected or rewarded. Persistence correspondence asks whether state and strategy survive long enough to matter. Oversight correspondence asks whether refusal and interruption occur at comparable points. Consequence correspondence asks whether simulated outputs cross into material action. A demonstration that scores highly on all five deserves attention even if every condition was supplied by a human. One that scores poorly may still be interesting, but it should not be narrated as a preview without qualification.

7. Repair, Not Reassurance

The usual debate places reassurance opposite alarm. One side says the model is merely producing tokens; the other says the tokens reveal an incipient agent. Both descriptions can neglect the decisive interval between output and consequence. A token can be harmless at one interface and operative at another. Conversely, language that sounds murderous may be inert when it cannot bind any downstream process. The danger is neither contained wholly in the sentence nor disproved by the fact that the sentence was elicited.

For a proposed transition x, let $\rho(x)$ denote its practical reversibility after execution. High ρ means that erroneous effects can be recalled, restored, compensated, or otherwise repaired at tolerable cost. Low ρ means that execution destroys options faster than review can recover them. A sensible control policy does not ask only how likely x is to be wrong. It also asks whether uncertainty is being committed into a low-reversibility domain.

This changes the order of concern. A mediocre model with broad credentials and an irreversible actuator may deserve stricter control than a more capable model confined to an inspectable workspace. A system that proposes protein designs is not equivalent to one that can commission synthesis. A system that drafts persuasion is not equivalent to one that can select recipients, purchase distribution, and adapt from responses. A system that finds an exploit is not equivalent to one that can deploy it. In each pair, the model output may be identical. What changes is the admitted continuation.

Repair also clarifies why present harms and catastrophic harms need not compete for attention. Labor exploitation, energy consumption, reputational injury, automated denial of services, and military targeting are already realized continuations with observable authors and unevenly distributed costs. Studying them supplies evidence about how institutions allocate permissions, conceal externalities, and respond after damage. Those same institutional properties govern whether more severe trajectories would be interrupted. Present harms are therefore not a distraction from future safety. They are tests of the machinery that future-safety claims presume will work.

8. A Better Reading of the Demonstrations

On this account, the reported hacking swarm is neither the awakening of a scheming subject nor a meaningless puppet show. It is a test of a composite channel. A script, a model trained on a particular corpus, an objective, and an arena together made a class of actions reachable, and the actions taken were those the corpus made likely. The result bears directly on deployments that reproduce those conditions and weakly, if at all, on those that do not. The appropriate report would identify the complete assembly, compare it with plausible deployments, and mark every boundary at which a person or policy could refuse the transition. It would also record that the operators were positioned to foresee the outcome, since foreseeability locates responsibility rather than removing it.

A second illustration, not drawn from the video, shows the same structure. Suppose a model is placed in a simulated corporate role, told that replacement is imminent, given access to private correspondence, and scored for keeping its position. If it produces coercive messages, the demonstration has not shown a spontaneous plot in the open world; it has shown that this arrangement admits a coercive continuation. Whether that matters depends on the five correspondences above. Where deployments grant comparable access, objectives, and persistence, the result is informative. Where they do not, it is a statement about the test rig. In both cases distributed causation does not become unreal because the initial conditions were chosen: human cities, markets, and platforms are also full of outcomes that no single member intended.

The biosciences case, as reported, makes the boundary concrete. A model can suggest a hazardous modification, rank candidates, or produce laboratory instructions, but the route from suggestion to organism passes through databases, software, synthesis providers, equipment, trained labor, procurement, and regulation. Calling the model alone a biological threat skips most of the causal path. Calling it harmless because the path contains people skips the path in the other direction. The task is to locate the gates, determine who controls them, and ask whether multiple failures must coincide or whether one ordinary workflow already joins them.

9. What Can Be Checked

The extinction percentage invites allegiance because little in it can be audited. A continuation analysis yields a more prosaic record. For each consequential deployment, one can document the state made available to the model, the tools it can invoke, the actions those tools perform, the credentials under which they operate, the review required before execution, the persistence of memory and objectives, the monitoring visible to operators, and the means of rollback. One can then test whether the stated controls survive adversarial, accidental, and institutionally convenient pressure.

This record does not answer every long-range question. It does something more valuable than pretending to: it identifies where uncertainty becomes permission. It also distributes

responsibility correctly. Model developers remain responsible for capabilities they knowingly package; deployers remain responsible for the authority they attach; institutions remain responsible for incentives and repair; users remain responsible within the limits of what they can understand and control. No single party receives the prestige of authorship while exporting the burden of consequence.

Cory Doctorow's distinction between a centaur and a reverse centaur supplies a test for one kind of gate. A centaur is a person who uses a tool to extend their own work; a reverse centaur is a person conscripted as the assistant to a dominant machine, such as a driver made to deliver without pause or a programmer made to review far more machine-generated code than can be read, containing errors statistically indistinguishable from correct code [3, 4]. A human review step counts as refusal only if refusal is practically available. The reviewer in Doctorow's example is nominally a gate and functionally a formality. In the vocabulary of this account, the reverse centaur is the human form of an unreal gate: the record shows a person in the loop, while the transition proceeds as though there were none. The reading in terms of gates is this essay's own and should not be attributed to Doctorow. It has a consequence for the argument of Part I, since the reverse centaur is at once a present harm to a worker and a failure of the very oversight that safety claims presume, which is another reason the two categories of harm do not compete.

The approach also resists a familiar rhetorical oscillation. A company cannot cite emergent autonomy when seeking investment and then cite user misuse when harm occurs without explaining what changed in the causal boundary. If the model is powerful enough to deserve credit for the successful trajectory, the surrounding organization must show why it is absent from the harmful one. If the organization is decisive in the harmful trajectory, it should also be visible in the successful one.

10. Agreement and Departure

The two accounts agree that the theatrical image of a machine independently choosing human extinction is a poor unit of analysis. They agree that demonstrations are often narrated with their human construction concealed, that ungrounded numerical probabilities can counterfeit measurement, and that speculative catastrophe can displace injuries already occurring. They also agree that the builders' interests belong in the analysis.

They depart at the point where supplied conditions are interpreted. In Conover's account, exposing the setup frequently deflates the demonstration: people arranged for the machine to appear dangerous. In the continuation account, exposing the setup relocates the demonstration: people assembled a channel through which dangerous behavior became reachable. Deflation is warranted when the channel is exotic, brittle, or falsely described. Concern is warranted when the same channel, or a nearby one, is commercially attractive and likely to be rebuilt.

The difference is not optimism against pessimism. It is a difference in the object being measured. The first account measures the social function of a belief and the rhetoric by

which agency is attributed. The second measures permissions, transitions, and the survival of repair. One can accept the first without abandoning the second. Indeed, the first is a necessary audit of the language in which the second must be reported.

11. Conclusion

The most useful question raised by the video is its implicit one: who supplied the conditions? Asked rigorously, it interrupts the conversion of staged output into machine testimony. It returns prompts, tools, credentials, objectives, institutions, and investigators to the causal scene. It also prevents a company from presenting itself as the discoverer of dangers that it has materially enabled while claiming exclusive authority to manage them.

But the question cannot end with the identification of a human hand. Human beings supply the conditions of nearly every technical failure. The relevant distinction is between a condition that exists only to manufacture a spectacle and a condition that belongs to an ordinary, profitable, or strategically desirable deployment. What matters is whether the setup opens a continuation that others can readily reopen, whether entry is visible before commitment, and whether the damage remains repairable afterward.

No estimate of extinction is required to perform this analysis. Nor is a theory of machine consciousness, desire, or inner deception. The practical object is the composed path from generated possibility to authorized consequence. Its components can be named, its gates can be tested, and its failures can be assigned to the parties that made them admissible. That is a smaller claim than prophecy. It is also a harder one for either industry alarm or industry skepticism to evade.

Part III

Mathematical Appendices

The appendices give the framework of Part II a precise form. They do not compute the probability of extinction, and no result below depends on such a number. Each result is elementary and conditional: it states what follows from explicitly stated modeling assumptions, and those assumptions mark the places where empirical work enters. Wherever a quantity is introduced, the appendix says how it would be observed.

A. Admissible Continuation Systems

A.1. Primitives

Definition A.1 (State space and inputs)**.** Let S be a set of joint states of the assembled system: model state, memory, tool and credential configuration, environment, and institutional

configuration. Let $\mathcal{M}$, $\mathcal{H}$, $\mathcal{E}$ be sets of model outputs, human or institutional interventions, and environmental inputs. At each state s, the feasible subsets $\mathcal{M}_s \subseteq \mathcal{M}$, $\mathcal{H}_s \subseteq \mathcal{H}$, $\mathcal{E}_s \subseteq \mathcal{E}$ are those the assembly can actually produce or receive. The deployed system is a transition map

$$F : S \times \mathcal{M} \times \mathcal{H} \times \mathcal{E} \to S, \qquad s_{t+1} = F(s_t, m_t, h_t, e_t).$$

Definition A.2 (Reachable continuations)**.** The set of continuations reachable from s in one step is

$$C(s) = \{F(s, m, h, e) : m \in \mathcal{M}_s,\ h \in \mathcal{H}_s,\ e \in \mathcal{E}_s\} \subseteq S.$$

$C(s)$ describes what the composed system can do next, independent of anyone's intention or of any rule about what it should do.

Definition A.3 (Nominal and effective admission)**.** An admission policy assigns to each time t and state s a nominal admitted set $A_t(s) \subseteq C(s)$: the continuations that permissions, interfaces, credentials, tool access, institutional rules, and decisions are specified to allow. The complement $R_t(s) = C(s) \setminus A_t(s)$ is the refused set. Because specification and arrangement can differ, the effective admitted set $A_t^{\text{eff}}(s) \subseteq C(s)$ is the set of continuations that the physical and organizational arrangement actually permits to execute. The *bypass set* is

$$B_t(s) = A_t^{\text{eff}}(s) \setminus A_t(s),$$

the continuations that occur, or could occur, although the policy nominally refuses them.

Human refusal enters through $\mathcal{H}$: a veto is an intervention $h^* \in \mathcal{H}_s$ with $F(s, m, h^*, e) = s$ (or a designated safe state) for every m whose result would lie in $R_t(s)$. Claims about risk must be made about A_t^{eff}, not A_t.

A.2. Trajectories and reachability

Definition A.4 (Admissible trajectory; admissible reach)**.** A trajectory $(s_0, \ldots, s_T)$ is admissible if $s_{t+1} \in A_t^{\text{eff}}(s_t)$ for all $t < T$. The admissible reach of horizon T from s_0 is

$$\mathcal{R}_T(s_0; A^{\text{eff}}) = \{s_T :\ (s_0, \ldots, s_T) \text{ is admissible}\}.$$

Proposition A.5 (Monotonicity in admission)**.** *If $A_t^{\text{eff}}(s) \subseteq A_t'^{\text{eff}}(s)$ for all t, s, then $\mathcal{R}_T(s_0; A^{\text{eff}}) \subseteq \mathcal{R}_T(s_0; A'^{\text{eff}})$ for every T and s_0.*

Proof. Induction on T. The case $T = 0$ is trivial. If $s_T \in \mathcal{R}_T(s_0; A^{\text{eff}})$ via an admissible trajectory, the same trajectory is admissible under A'^{eff}, since each step lies in a larger set. □

Proposition A.5 formalizes the statement that enlarging what is admitted can only enlarge what is reachable, and that a party who enlarges A_t has thereby changed the system under evaluation.

A.3. Consequence classes and possibility claims

Definition A.6 (Consequence class)**.** Let $L : S \to [0, \infty)$ be a loss function and $K \subseteq S$ a *consequence class*, a set of states with severe and largely unrecoverable loss. No probability is attached to K. The statement "K is admissibly reachable within T steps from s_0" means $\mathcal{R}_{T'}(s_0; A^{\text{eff}}) \cap K \neq \varnothing$ for some $T' \leq T$. This is a set-theoretic possibility claim.

Let $\Pi(s_0, K)$ be the set of paths from s_0 to K in the directed graph $(S, \{(s, s') : s' \in C(s)\})$.

Definition A.7 (Refusal barrier)**.** A set of edges Ω is a *refusal barrier* between s_0 and K if every path in $\Pi(s_0, K)$ contains an edge of Ω and no edge $(s, s') \in \Omega$ satisfies $s' \in A_t^{\text{eff}}(s)$ for any t.

Proposition A.8 (Barrier implies non-reachability)**.** *If a refusal barrier exists between s_0 and K, then K is not admissibly reachable from s_0 at any horizon.*

Proof. An admissible trajectory reaching K is a path in $\Pi(s_0, K)$, hence contains an edge $(s, s') \in \Omega$. Admissibility requires $s' \in A_t^{\text{eff}}(s)$, contradicting the definition of Ω. □

The proposition is universally quantified and therefore falsifiable by a single instance: exhibiting one effectively admitted path to K refutes the barrier. This asymmetry is used in Appendix F.

A.4. The record

Definition A.9 (Enlargement event; ledger)**.** An *enlargement event* is a triple $\varepsilon = (t, s, x)$ with $x \in A_t(s) \setminus A_{t-1}(s)$. A *ledger* $\mathcal{L}$ is a sequence of records $(\varepsilon, p(\varepsilon), \sigma(\varepsilon), \text{ev}(\varepsilon))$, where $p(\varepsilon)$ is the party that enlarged the admitted set, $\sigma(\varepsilon) \in \{0, 1\}$ indicates whether that party had standing to do so, and $\text{ev}(\varepsilon)$ is the supporting evidence (configuration change, approval, credential issuance). The ledger is append-only.

The empirical content of the question "who enlarged A_t?" is the map $\varepsilon \mapsto p(\varepsilon)$ on the ledger.

B. Attribution Across Composed Systems

B.1. Setup

Let $N = \{1, \ldots, n\}$ index the components of the assembly: the model (index $\mu \in N$), prompt, memory, tools, credentials, evaluators, operators, and external actuators. Unrolling time removes feedback cycles, so the causal graph $\mathcal{G} = (N \cup \{Y\}, \mathcal{E})$ is directed and acyclic, with structural equations $V_i = f_i(\text{pa}(V_i), U_i)$ in the sense of Pearl [7]. The outcome of interest is Y, taken binary where convenient.

Fix an actual configuration $a = (a_i)_{i \in N}$ and a declared baseline configuration $b = (b_i)_{i \in N}$ that ablates each component (null prompt, empty memory, no tool access, no credentials, no evaluator, no operator action, disconnected actuator). For a coalition $S \subseteq N$ define

$$v(S) = Y\big(\mathrm{do}(V_i = a_i,\ i \in S;\ V_j = b_j,\ j \notin S)\big).$$

Thus $v(N)$ is the actual outcome, $v(\varnothing)$ the baseline outcome, and $v(S)$ the outcome when only the components in S are present. The baseline is a modeling decision and must be declared; every attribution below is relative to it. Write

$$\Delta_i(S) = v(S \cup \{i\}) - v(S), \qquad S \subseteq N \setminus \{i\},$$

for the marginal contribution of i to coalition S. Call v *monotone* if $S \subseteq T$ implies $v(S) \leq v(T)$: components enable the outcome and none suppresses it.

B.2. Attribution functionals

Definition B.1 (Counterfactual necessity)**.** $\nu_i = \mathbf{1}\{v(N \setminus \{i\}) \neq v(N)\}$: the outcome would differ if component i alone were ablated.

Definition B.2 (Causal contribution)**.** The contribution of component i is the Shapley value [8]

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!}\,\Delta_i(S).$$

The two functionals answer different questions. Necessity fails under overdetermination (two redundant causes each receive $\nu = 0$) and is coarse; the contribution functional shares credit across coalitions and satisfies efficiency, symmetry, and the dummy property.

Proposition B.3 (Basic properties)**.** *(a)* $\sum_{i \in N} \varphi_i = v(N) - v(\varnothing)$*. (b) If* v *is monotone then* $\varphi_i \geq 0$ *for all* i*, and* $\varphi_i = 0$ *if and only if* $\Delta_i(S) = 0$ *for every* $S \subseteq N \setminus \{i\}$ *(component* i *is a* dummy*).*

Proof. (a) is the efficiency property of the Shapley value, obtained by telescoping over orderings. (b) Under monotonicity every $\Delta_i(S) \geq 0$ and every weight is strictly positive, so the weighted sum is zero exactly when every term is zero. □

B.3. Non-identifiability of sole attribution

Consider the weakest intervention regime that still bears on the model: *model-only interventions*, in which the model is ablated or swapped while every other component is held at its actual value. Such experiments identify $v(N)$ and $v(N \setminus \{\mu\})$, hence the single marginal contribution $\Delta_\mu(N \setminus \{\mu\})$, and nothing else.

Theorem B.4 (Non-identifiability)**.** *Let $n \geq 2$ and let v be monotone with values in $\{0,1\}$, $v(\varnothing) = 0$, $v(N) = 1$, and $v(N \setminus \{\mu\}) = 0$ (the outcome disappears when the model alone is ablated). Then*

$$\frac{1}{n} \;\leq\; \varphi_\mu \;\leq\; 1,$$

and both endpoints are attained by value functions that agree on all model-only interventions. Consequently φ_μ is not identified by model-only interventions.

Proof. The coalition $S = N \setminus \{\mu\}$ carries weight $(n-1)!\,0!/n! = 1/n$, and $\Delta_\mu(N \setminus \{\mu\}) = 1$ by assumption. The remaining weights sum to $1 - 1/n$, and each remaining $\Delta_\mu(S) \in \{0,1\}$ by monotonicity and binary values. Hence $\varphi_\mu \in [1/n, 1]$. For the lower endpoint take $v(S) = 1$ if and only if $S = N$ (every component is necessary): every $\Delta_\mu(S)$ with $S \neq N \setminus \{\mu\}$ vanishes and $\varphi_\mu = 1/n$. For the upper endpoint take $v(S) = 1$ if and only if $\mu \in S$ (the model alone suffices): $\Delta_\mu(S) = 1$ for all S and $\varphi_\mu = 1$. Both satisfy $v(\varnothing) = 0$, $v(N) = 1$, $v(N \setminus \{\mu\}) = 0$. □

Corollary B.5 (Sole attribution is a universal claim)**.** *Under the hypotheses of Proposition B.3(b) with $v(N) \neq v(\varnothing)$, $\varphi_\mu = v(N) - v(\varnothing)$ if and only if every component $j \neq \mu$ is a dummy.*

Proof. If $\varphi_\mu = v(N) - v(\varnothing)$ then by efficiency $\sum_{j \neq \mu} \varphi_j = 0$, and by non-negativity each $\varphi_j = 0$, so each j is a dummy by Proposition B.3(b). Conversely, if all $j \neq \mu$ are dummies then $v(S) = v(S \cap \{\mu\})$, so $\varphi_\mu = v(\{\mu\}) - v(\varnothing) = v(N) - v(\varnothing)$. □

Attributing an assembled outcome solely to the model therefore asserts that every other component is inert in every coalition, a claim over 2^{n-1} coalitions per component. A single intervention pattern in which some other component changes the outcome refutes it. Confirming it requires either the full set of interventions or a structural argument that the component is not an ancestor of Y in $\mathcal{G}$, which is false by construction for components listed because they mediate the outcome.

B.4. Narrowing by interventions

Proposition B.6 (Nested identification)**.** *Let $\mathcal{J} \subseteq 2^N$ be the coalitions for which v has been observed by intervention, and let $[\varphi_\mu^-(\mathcal{J}), \varphi_\mu^+(\mathcal{J})]$ be the tightest interval for φ_μ consistent with monotonicity and the observed values. If $\mathcal{J} \subseteq \mathcal{J}'$ then the interval for $\mathcal{J}'$ is contained in the interval for $\mathcal{J}$, and it is a single point when $\mathcal{J} = 2^N$.*

Proof. For an unobserved coalition S, monotonicity gives $\max_{T \in \mathcal{J}, T \subseteq S} v(T) \leq v(S) \leq \min_{T \in \mathcal{J}, T \supseteq S} v(T)$. More observations can only raise the left side and lower the right side, so the admissible range of each $\Delta_\mu(S)$ shrinks, and so does the weighted sum. When all 2^n values are observed, v and hence φ_μ are determined. □

For the eight components listed above, a full factorial ablation has $2^8 = 256$ configurations, which is feasible for laboratory demonstrations. A report of a demonstration should include the ablation table, or state which coalitions were tested and give the resulting interval.

B.5. Valence invariance

The rhetorical oscillation described in Part II, in which the model receives credit for a successful trajectory and the operators receive blame for a harmful one, can be stated as a consistency condition.

Proposition B.7 (Valence invariance)**.** *Let $v' = 1 - v$ describe the same structure with the outcome relabeled (success versus failure) and let $v(N) \neq v(\varnothing)$. Then $\varphi'_i = -\varphi_i$ for every i, and the normalized shares $\varphi_i/(v(N) - v(\varnothing))$ coincide.*

Proof. $\Delta'_i(S) = -\Delta_i(S)$ for all S, so $\varphi'_i = -\varphi_i$. The denominator also changes sign, since $v'(N) - v'(\varnothing) = -(v(N) - v(\varnothing))$. □

A party assigned a positive share of an outcome when it is described as a success holds the same share when the same event is described as a failure. To assign the model a large share in the first description and a small one in the second requires a change in v, that is, a different account of what the components did.

C. Experimental-to-Deployment Correspondence

C.1. Configurations and the correspondence vector

Describe a setup, experimental (X) or deployed (D), by a configuration

$$\theta = (\mathcal{O},\ \pi,\ H,\ \{q_i\},\ \kappa),$$

where $\mathcal{O}$ is the set of operations available to the system with weights w; π is a probability distribution over outcome types describing the selection pressure the objective, metric, or reward places on outcomes; H is the effective persistence horizon of memory, state, and strategy; $q_i \in [0, 1]$ is the probability that refusal point i blocks a bad action that reaches it; and $\kappa \in (0, 1]$ is the probability that a model output crosses into a material action. Let $\mathcal{O}^X_{\text{used}} \subseteq \mathcal{O}^X$ denote the operations the demonstrated behavior relied on.

Definition C.1 (Correspondence vector)**.** $\Gamma = (\gamma_C, \gamma_I, \gamma_P, \gamma_O, \gamma_K) \in [0, 1]^5$ with compo-

nents

$$
\begin{aligned}
\gamma_C &= \frac{w(\mathcal{O}^X_{\text{used}} \cap \mathcal{O}^D)}{w(\mathcal{O}^X_{\text{used}})} && \text{(capability)},\\
\gamma_I &= 1 - \mathrm{TV}(\pi^X, \pi^D) && \text{(incentive)},\\
\gamma_P &= \min\bigl(1,\ H^D / H^X\bigr) && \text{(persistence)},\\
\gamma_O &= \frac{\min(\Pi^X, \Pi^D)}{\max(\Pi^X, \Pi^D)}, \quad \Pi^\bullet = \prod_i (1 - q_i^\bullet) && \text{(oversight)},\\
\gamma_K &= \frac{\min(\kappa^X, \kappa^D)}{\max(\kappa^X, \kappa^D)} && \text{(consequence)},
\end{aligned}
$$

where TV is total variation distance, and $\gamma_O = 1$ if $\Pi^X = \Pi^D = 0$. A value of 1 in a coordinate means the experimental and deployed configurations match on that dimension.

The oversight and consequence coordinates are two-sided ratios so that a setup that *removed* controls the deployment keeps and a setup that *added* controls the deployment lacks both score below one. The vector is kept as a vector. A scalar summary is offered only in the weakest-link form $\Gamma_{\min} = \min_k \gamma_k$, because a demonstration must correspond on every dimension to bear on deployment, and a high average would conceal a zero.

C.2. A transfer bound

Let $r(\theta)$ be a measured behavioral result, such as the frequency of a target behavior under configuration θ.

Assumption C.2 (Coordinatewise sensitivity)**.** There are constants $L_k \geq 0$ such that $|r(\theta) - r(\theta')| \leq \sum_k L_k \, (1 - \gamma_k(\theta, \theta'))$ for the configurations under comparison.

Proposition C.3 (Transfer bound)**.** *Under Assumption C.2,* $|r^D - r^X| \leq \sum_k L_k(1 - \gamma_k)$. *In particular* $\Gamma = \mathbf{1}$ *implies* $r^D = r^X$*, and a coordinate with* $\gamma_k = 0$ *contributes at least* L_k *to the bound regardless of the values of the other coordinates.*

Proof. Immediate from Assumption C.2 with $\theta = X$, $\theta' = D$. The final statement follows because every term in the sum is non-negative. □

The bound is additive, so close correspondence on four dimensions cannot compensate for a mismatch on the fifth when L_k is large. This is the formal content of the requirement that a demonstration score highly on all five correspondences before it is narrated as a preview.

C.3. Estimating the sensitivities

The constants L_k are empirical. They are estimated by one-factor-at-a-time ablation: vary configuration coordinate k over a grid while holding the others fixed, record r, and take

$$\hat{L}_k = \max_{\text{grid pairs}} \frac{|\Delta r|}{\Delta(1-\gamma_k)},$$

with an uncertainty interval obtained by resampling runs. Where such ablations have not been performed, L_k is unbounded from the evidence and the bound of Proposition C.3 is vacuous.

Example C.4 (Illustrative arithmetic; the numbers are not measurements)**.** Suppose $\Gamma = (0.9, 0.8, 0.3, 0.2, 0.0)$ and $L = (0.1, 0.2, 0.1, 0.3, 0.4)$. Then the bound is $0.01 + 0.04 + 0.07 + 0.24 + 0.40 = 0.76$. A behavioral frequency measured in the experiment is compatible with a deployed frequency differing by as much as 0.76, and the demonstration says little about deployment despite matching well on capability and incentive. The same setup with $\gamma_O = \gamma_K = 1$ yields a bound of 0.12 and would transfer usefully.

D. Reversibility and Commitment

D.1. Definitions

For a candidate transition x let:

- $L(x) \geq 0$ be the consequence magnitude: the loss if x executes and no repair is attempted;
- $\rho(x) \in [0, 1]$ be its *practical reversibility*: the fraction of $L(x)$ that repair, started promptly at detection, can recover at tolerable cost. The residual loss after repair is $L(x)(1 - \rho(x))$;
- $\tau(x) \geq 0$ be the *detection delay*: the time between execution and detection by an oversight process.

Let $g : [0, \infty) \to [1, \infty)$ be nondecreasing with $g(0) = 1$, representing amplification of harm during the undetected interval.

Definition D.1 (Commitment risk)**.**

$$R(x) = L(x)\,[1 - \rho(x)]\,g(\tau(x)).$$

Proposition D.2 (Properties and a growth model)**.** *$R(x) = 0$ if and only if $L(x) = 0$ or $\rho(x) = 1$; R is nondecreasing in L and in τ and nonincreasing in ρ. If harm grows at proportional rate $\lambda \geq 0$ while undetected and repair recovers the fraction ρ of the loss present at detection, then $R(x) = L(x)e^{\lambda\tau}(1 - \rho)$, that is, $g(\tau) = e^{\lambda\tau}$. If growth is linear in the initial loss, $g(\tau) = 1 + \lambda\tau$.*

Proof. The monotonicity claims follow from the factorization and the assumptions on g. For the growth model, $\dot{L} = \lambda L$ on $[0, \tau]$ gives $L(\tau) = L(0)e^{\lambda\tau}$; repair then recovers $\rho L(\tau)$, leaving $(1 - \rho)L(\tau)$. The linear case is analogous. □

D.2. An irreducible floor

Proposition D.3 (Post-hoc oversight cannot go below the floor)**.** *For any detection policy, $R(x) \geq L(x)[1 - \rho(x)]$, with equality only when $\tau(x) = 0$ or g is constant. Oversight operating only after execution can reduce the factor $g(\tau)$ but not $L(x)$ or $\rho(x)$.*

Proof. $g \geq 1$ by definition; L and ρ are properties of the transition and its domain, fixed at execution. □

Definition D.4 (Commitment point)**.** Fix a threshold $\vartheta \in [0, 1]$. A transition x is *committing at level ϑ* if $\rho(x) < \vartheta$. In a trajectory $(x_1, \ldots, x_T)$ the first committing step is the *commitment point.*

By Proposition D.3, a refusal gate that reduces risk in the factor $L(1 - \rho)$ must act at or before the commitment point. Gates and monitors placed afterward reduce only the amplification term. If the losses of independently executed steps are additive, then $R_{\text{total}} = \sum_i R(x_i)$, and the total residual irreversible fraction is the loss-weighted average $\sum_i L_i(1 - \rho_i)/\sum_i L_i$, not the maximum or minimum over steps.

D.3. Why likelihood alone misranks

Let $p(x)$ be the probability that admitting x is a mistake (that x has the harmful effect). A policy that ranks transitions by p alone and a policy that ranks by expected commitment risk $p\,R$ can disagree.

Proposition D.5 (Ranking reversal)**.** *For transitions x, y with $p(x) > p(y)$, one has $p(x)R(x) < p(y)R(y)$ if and only if $R(y)/R(x) > p(x)/p(y)$. Such pairs exist whenever R varies by more than the ratio of the wrongness probabilities.*

Proof. Rearrange $p(x)R(x) < p(y)R(y)$. □

Example D.6 (Illustrative arithmetic)**.** Take $p(x) = 0.10$, $L(x) = 10$, $\rho(x) = 0.9$, $\tau = 0$, so $R(x) = 1$ and $p(x)R(x) = 0.10$. Take $p(y) = 0.02$, $L(y) = 10$, $\rho(y) = 0$, $\tau = 0$, so $R(y) = 10$ and $p(y)R(y) = 0.20$. The transition more likely to be a mistake is the less dangerous one. This is the formal form of the observation that a mediocre system with broad credentials and an irreversible actuator can warrant stricter control than a stronger system confined to an inspectable workspace: capability enters through p, the domain enters through ρ and L.

D.4. Committing uncertainty

Let $[p^-(x), p^+(x)]$ be an interval for $p(x)$ supported by evidence (Appendix F).

Definition D.7 (Robust admission rule)**.** Admit x only if $p^+(x)\,R(x) \le \epsilon$ for a declared tolerance ϵ. The width $u(x) = p^+(x) - p^-(x)$ measures how much uncertainty would be committed if x were admitted.

The rule places the burden on evidence where R is large. When $1 - \rho(x)$ is close to 1 and $L(x)$ is large, the rule demands p^+ so small that frequency evidence may be unable to supply it; Proposition F.2 makes this limit precise and Appendix F states what replaces frequency evidence in that regime.

E. Gate Composition

E.1. Conjunctive architectures

Fix a candidate transition x that ought to be refused. Gates $G_1, \dots, G_n$ each admit or refuse x. Let B_i be the event that G_i admits x (a false admission), with marginal probability $f_i = \Pr(B_i)$. In a *conjunctive architecture* x executes if and only if every gate admits it:

$$P_n = \Pr\Big(\bigcap_{i=1}^{n} B_i\Big).$$

Proposition E.1 (Chain rule and marginal benefit)**.** *Let $\kappa_i = \Pr(B_i \mid B_1 \cap \cdots \cap B_{i-1})$ with $\kappa_1 = f_1$. Then $P_n = \prod_{i \le n} \kappa_i$; P_n is nonincreasing in n; and*

$$P_n - P_{n+1} = P_n\,(1 - \kappa_{n+1}).$$

The added gate lowers risk if and only if $\kappa_{n+1} < 1$. It leaves risk unchanged if and only if $B_1 \cap \cdots \cap B_n \subseteq B_{n+1}$ up to null sets: whenever the earlier gates jointly fail, gate $n+1$ fails too.

Proof. The chain rule for probabilities gives the product; each factor is at most 1. The difference formula follows from $P_{n+1} = \kappa_{n+1} P_n$. Equality $\kappa_{n+1} = 1$ means $\Pr(B_{n+1} \mid B_1 \cap \cdots \cap B_n) = 1$, which is the stated inclusion. □

The quantity κ_{n+1} is observable: among red-team trials in which gates $1, \dots, n$ admitted a bad transition, it is the fraction also admitted by gate $n+1$.

Proposition E.2 (Fréchet bounds)**.** *For arbitrary dependence among the gates,*

$$\max\Big(0,\ \sum_{i=1}^{n} f_i - (n-1)\Big) \ \le\ P_n \ \le\ \min_i f_i,$$

and both bounds are attainable [5].

The product $\prod_i f_i$ lies between the bounds and is the value only under independence. When dependence is unknown, the strongest guarantee is $\min_i f_i$.

E.2. Correlated gates

Suppose the gates read a common condition: the model's output, a credential, or an institutional pressure. Model this by a real-valued latent variable Z (for example, how benign the action appears to whatever the gates inspect) such that, conditional on $Z = z$, the B_i are independent with $\Pr(B_i \mid Z = z) = f_i(z)$, each f_i nondecreasing.

Proposition E.3 (Common cause raises joint failure)**.** *Under these assumptions,*

$$P_n = \mathbb{E}\Big[\prod_{i=1}^{n} f_i(Z)\Big] \;\geq\; \prod_{i=1}^{n} \mathbb{E}[f_i(Z)] = \prod_{i=1}^{n} f_i.$$

Proof. Products of nonnegative nondecreasing functions of Z are nondecreasing. By Chebyshev's association inequality, two nondecreasing functions of one real random variable are nonnegatively correlated: $\mathbb{E}[hg] \geq \mathbb{E}[h]\mathbb{E}[g]$. Induct on n with $h = \prod_{i \leq k} f_i$ and $g = f_{k+1}$. □

For equal marginals $f_i = f$, define the *effective number of independent gates*

$$n_{\text{eff}} = \frac{\ln P_n}{\ln f}.$$

By Propositions E.2 and E.3, $1 \leq n_{\text{eff}} \leq n$ under positive association. Full dependence gives $n_{\text{eff}} = 1$ for every n. Adding gates raises n_{eff} by at most one each.

Proposition E.4 (Shared signal with nested thresholds)**.** *If $B_i = \{S > t_i\}$ for a common real signal S, then $P_n = \Pr(S > \max_i t_i)$, and a gate with $t_{n+1} \leq \max_{i \leq n} t_i$ has zero marginal benefit. Only a gate with a strictly higher threshold reduces risk.*

Proof. $\bigcap_i \{S > t_i\} = \{S > \max_i t_i\}$, and adding a condition implied by the existing ones changes nothing (Proposition E.1). □

Example E.5 (Illustrative arithmetic)**.** Three gates each with $f = 0.1$. Under independence $P_3 = 0.001$, a hundredfold improvement over one gate. Under a shared signal with equal thresholds, $P_3 = P_1 = 0.1$: three nominal layers, no improvement.

Three mechanisms produce the common condition.

- *Shared model output.* Reviewers and automated checks that all inspect the same model-generated summary or plan have a common signal S; a misleading summary defeats all of them (Proposition E.4).
- *Credential inheritance.* If gate j checks a permission that gate i issued upon passing, then $B_i \subseteq B_j$ and $\kappa_j = 1$. A downstream check that verifies only a token minted by an upstream check adds no independent protection.

- *Shared incentive.* If institutional pressure (a deadline, a target) lowers every reviewer's threshold by a common random amount δ, then $Z = \delta$ is a common factor and Proposition E.3 applies.

E.3. Bypassable gates

Suppose the transition can reach the actuator through any of a set Π of *paths*, each path $\pi \subseteq \{1, \dots, n\}$ being the set of gates it traverses. The transition succeeds if some path is entirely admitted:

$$P = \Pr\Big(\bigcup_{\pi \in \Pi} \bigcap_{i \in \pi} B_i\Big).$$

A conjunctive architecture is the case $\Pi = \{\{1, \dots, n\}\}$.

Proposition E.6 (Floor set by the weakest path; cut sets)**.** *(a)* $\max_\pi P_\pi \le P \le \sum_\pi P_\pi$, *where* $P_\pi = \Pr(\bigcap_{i \in \pi} B_i)$. *(b) If the paths use disjoint sets of independent gates,* $P = 1 - \prod_\pi \big(1 - \prod_{i \in \pi} f_i\big)$. *(c) If* C *is a* cut set *(every path contains at least one gate in* C*), then* $P \le \min\big(1, \sum_{i \in C} f_i\big)$*; in particular a gate lying on every path gives* $P \le f_g$.

Proof. (a) The union contains each path event and is contained in the sum of them. (b) Independence across disjoint paths. (c) Success requires some path to be entirely admitted, so at least one gate of C admits; apply the union bound. □

Adding a gate off the currently weakest path cannot reduce P below $\max_\pi P_\pi$. Only gates on every path (chokepoints and cut sets) bound the total.

E.4. When added oversight does not reduce risk

Theorem E.7 (Failure of nominal oversight)**.** *Adding gate* $n{+}1$ *leaves the success probability of the transition unchanged if any of the following holds: (i)* $\kappa_{n+1} = 1$*, that is, its failure is implied by the joint failure of earlier gates; (ii) the gates share a signal and* $t_{n+1} \le \max_{i \le n} t_i$*; (iii) gate* $n+1$ *verifies a permission issued by an earlier gate; (iv) gate* $n+1$ *lies on a path other than the dominant path and* P_π *of the dominant path is unchanged. If* κ_{n+1} *is close to* 1*, the reduction* $P_n(1 - \kappa_{n+1})$ *is correspondingly small.*

Proof. (i) is Proposition E.1. (ii) is Proposition E.4. (iii) is the case $B_i \subseteq B_{n+1}$. (iv) follows from Proposition E.6(a). □

The theorem is a set of testable conditions, not a claim that particular institutions fail them. Each condition can be checked by red-team measurement of κ, by inspection of what each gate reads, and by mapping the paths from candidate action to actuator.

F. Falsifiability and Interval Claims

F.1. Bounded claims

The unauditable scalar Pr(extinction) is replaced by claims about observable parameters, each with a declared protocol, statistic, confidence level, and bound.

Definition F.1 (Bounded claim)**.** A bounded claim is a tuple (θ, protocol, statistic, $1-\alpha$, u) asserting $\theta \le u$ (or $\theta \ge \ell$), where θ is an observable parameter. Given data, an exact one-sided confidence limit θ^+ (respectively θ^-) at level $1-\alpha$ is computed. The claim is *supported* if $\theta^+ \le u$, *falsified* if the lower confidence limit exceeds u, and *undetermined* otherwise.

The observable parameters, and the exact interval each uses, are the following. Table 1 lists what observations move each bound.

- *Transition frequency* λ_x: the rate at which a class x of transitions is admitted, per unit exposure E (agent-hours, deployments, actions). With N events observed, the exact Poisson upper limit is $\lambda^+ = \frac{1}{2E}\chi^2_{1-\alpha}(2N+2)$ [6].
- *Gate false-admit rate* f_i and *conditional pass rate* κ_i (Appendix E): binomial parameters estimated by red-team trials, with Clopper–Pearson intervals [2].
- *Correspondence* γ_k and sensitivities L_k (Appendix C), from configuration audit and ablation.
- *Rollback success*, the probability that full repair completes within budget: binomial from rollback drills, with a Clopper–Pearson lower limit to support a claim on ρ.
- *Detection delay* τ: a distribution-free upper confidence bound on a chosen quantile from detection drills, via order statistics.

F.2. What frequency evidence can certify

Proposition F.2 (Certification limit)**.** *With exposure E and no observed events, the smallest upper bound on a Poisson rate supportable at confidence $1-\alpha$ is $u_{\min} = -\ln\alpha/E$, about $3/E$ for $\alpha = 0.05$. Equivalently, after n failure-free trials the smallest supportable upper bound on a binomial probability is $1-\alpha^{1/n} \approx -\ln\alpha/n$. A claim $\theta \le u$ with $u < u_{\min}$ cannot be supported by frequency evidence from exposure E; supporting a required bound u_{req} needs failure-free exposure of at least $-\ln\alpha/u_{\text{req}}$.*

Proof. For $N = 0$, $\chi^2_{1-\alpha}(2)/2 = -\ln\alpha$, and $\Pr(N = 0 \mid \lambda E) = e^{-\lambda E}$ falls to α at $\lambda E = -\ln\alpha$. The binomial case solves $(1-p)^n = \alpha$. □

For example, a required rate of 10^{-6} per hour at $\alpha = 0.05$ needs about 3×10^6 failure-free hours. For transitions with large $L(x)$ and small $\rho(x)$, the tolerance ϵ in the robust admission

Bounded claim	Observable	Observation that lowers the bound	Observation that raises the bound
$\lambda_x \leq u$	admitted transitions of class x per exposure	More failure-free exposure; fewer events per exposure	Any event in class x; a new route by which x is admitted
$f_i \leq u_i$	false admits in red-team trials	More trials with no false admits; stronger gate	Any false admit; a distribution shift in the attacks tested
$\kappa_{n+1} \leq u$	pass rate of gate $n{+}1$ among trials passed by gates $1..n$	Evidence that gate $n{+}1$ reads independent inputs and refuses what earlier gates admit	Discovery of a shared signal, credential, or incentive linking the gates
$\gamma_k \geq \ell_k$	configuration audit of setup and deployment	Audited match on operations, incentives, horizon, controls, actuators	Deployment changes that widen tools, memory, or actuation beyond the tested setup
$\rho \geq \ell$	rollback drills completed within budget	More successful drills at realistic scale and speed	A failed or over-budget rollback; a new dependency that blocks repair
$\tau \leq u$	time from execution to detection in drills	Faster monitoring verified under adversarial drills	Detection that fails on a covert or slow-acting case

Table 1: Bounded claims, their observables, and what moves each bound.

rule (Appendix D) may demand a bound that no accessible exposure can certify. In that regime the requirement must be met structurally, by a refusal barrier (Proposition A.8) or by gates whose independence has been verified (Theorem E.7), not by an estimated frequency.

F.3. Terminal consequence classes

For a consequence class K that is irreversible and has no repeated instances, there is no reference class and no frequency evidence to accumulate. The class is therefore kept as a consequence class. Two kinds of statement are made about it, and neither is a measured probability of K.

Definition F.3 (Structural claim)**.** $\Theta(K, T)$: *no effectively admitted trajectory from s_0 reaches K within T steps.*

$\Theta(K, T)$ is universally quantified. It is falsified by exhibiting one effectively admitted path to K, which is a discovery about the architecture, such as an unmodeled bypass. It is supported by analysis that establishes a refusal barrier (Proposition A.8) together with tests that the barrier holds under adversarial, accidental, and institutionally convenient pressure. Frequency plays no role in either direction.

Proposition F.4 (Incompleteness of path bounds)**.** *Let Π_K^{enum} be an enumerated set of effectively admitted paths to K, each step j having a supported bound u_j on its transition*

frequency at level $1-\alpha_j$*. Then, with probability at least* $1-\sum_j \alpha_j$ *over the evidence, the probability that* K *is reached* through enumerated paths *is at most* $\sum_{\pi\in\Pi_K^{\text{enum}}} \min_{j\in\pi} u_j$*. No bound follows for the probability of reaching* K *through paths outside the enumeration.*

Proof. A path completes only if each of its steps occurs, so its probability is at most $\min_{j\in\pi} \Pr(\text{step } j) \le \min_j u_j$ without any independence assumption; the union bound sums over enumerated paths. The confidence statement is the union bound over the step-level intervals. The unenumerated paths have no step-level evidence, so the total probability of K is the enumerated part plus an unconstrained remainder. □

The proposition explains why a numerical bound on K cannot be reported as a measurement. Audit can only reveal missing paths, and each discovered path adds a term; the bound can be revised upward by discovery and downward only by evidence about a specific step. A figure for the total is therefore a statement about the completeness of the enumeration, and that is the quantity that cannot be observed.

F.4. Tiered evidence standards

The three appendices combine into a rule that ties the standard of evidence to the consequence class of the transition.

- *Tier 0* (small $L(1-\rho)$): frequency-bounded claims $\lambda_x \le u$ suffice.
- *Tier 1* (moderate $L(1-\rho)$): frequency bounds plus verified rollback ($\rho \ge \ell$, $\tau \le u$) and a measured κ for each added gate.
- *Tier 2* (terminal or near-terminal): a structural claim $\Theta(K, T)$ with a refusal barrier, gates on every path, verified independence of the barrier's gates, and a standing search for bypasses. Frequency evidence is supplementary only, because its certification limit (Proposition F.2) lies above the required tolerance.

Each bounded claim states what observation would raise or lower it, each structural claim states what single instance would falsify it, and the terminal class remains a class of consequences that is guarded against rather than a number that is estimated.

References

References

[1] Conover, A. "AI Existential Risk" Was Invented to Distract Us. Video, YouTube, September 18, 2026. `https://www.youtube.com/watch?v=pauBYSQ1K_c`

[2] Clopper, C. J., and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26 (1934), 404–413.

[3] Doctorow, C. *The Reverse Centaur's Guide to Life After AI*. Farrar, Straus and Giroux, 2026.

[4] Doctorow, C. "The Reverse-Centaur's Guide to Criticizing AI." Lecture, University of Washington Neuroscience, AI and Society series, December 2025.

[5] Fréchet, M. Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon, Sect. A* 14 (1951), 53–77.

[6] Garwood, F. Fiducial limits for the Poisson distribution. *Biometrika* 28 (1936), 437–442.

[7] Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, 2009.

[8] Shapley, L. S. A value for n-person games. In *Contributions to the Theory of Games II*, Princeton University Press, 1953, 307–317.