The Catastrophe Experiment

Deciding AI Safety Before Irreversible Action

José Santiago Niño Mojica
Founder, Recognitium. Paris, France. jose@recognitium.com
October 2026 · DOI: 10.5281/zenodo.23645657

Abstract. Artificial intelligence is the only safety-critical technology whose safety case rests on the internal dispositions of the system rather than on bounds placed on its external actions. We show that, under this standard of evidence, the only observation that can refute the claim that a system is aligned is the harm the claim was meant to exclude, and we characterise the point at which safety can instead be decided before an irreversible effect. A governor observes only the requests that cross the execution boundary, never the ~10^{11} parameters that produce them. Since the same request can be produced by an aligned state, a deceptive state, and a software fault, no benchmark, reasoning trace, or interpretability readout decides alignment before the act (Theorems 1 and 2). Any alignment claim offered as evidence of safety implies a bound on permitted effects (Proposition 1); specifying such a bound is therefore common to every approach, and the approaches differ only in whether the bound is verified. The unique point at which it can be verified before irreversibility is the last projection from internal computation to external request, in deployed systems the tool call, where the reliability of an independent gate does not depend on the capability of the model (Theorem 3). A record of that verification is evidence, rather than the operator's testimony, only if the check, the advance of an external sequence, and the release of the action form a single atomic transition with no signing credential that survives it (Theorem 4); logs, enclaves, and ledgers signed with persistent keys, post-quantum schemes included, do not meet this condition. The paper does not address what the bounds should prohibit, which is a matter for law. Annex A states the resulting standard as four testable clauses for the harmonised standards under the EU AI Act.

> *You cannot govern the thought. You can govern the act. The act has exactly one governable point. Govern there, and make the record of having governed there verifiable by anyone.*

1. Introduction

Civil aviation does not certify an aircraft by asking whether the pilot intends to crash. It certifies the physical operating bounds the aircraft is prevented from exceeding, and it records every flight in an independent flight data recorder that neither the crew nor the airline can alter.

Artificial intelligence is the one safety-critical technology whose safety case is still allowed to rest on the system's internal dispositions. The prevailing approach holds that a system is made safe by shaping its latent objectives (training it to be helpful, honest, and harmless) and then verifying, by pre-deployment evaluation and internal inspection, that those dispositions hold in use. The second step has not been achieved for any frontier system, and in 2026 the institutions that build and review these systems stated the reasons themselves.

One frontier developer now defines misalignment as "a latent property of a specific computation", to be judged "via powerful interpretability tools that may not currently exist", and states that "misaligned computation need not be observable" (Anthropic, *Risk Report*, 14 August 2026). The independent reviewer of that report concludes that present risk is low "primarily due to models' limited capabilities, rather than safeguards and alignment training" (METR, 20 October 2026). The report of more than one hundred experts and thirty governments names an "evaluation gap": "performance on pre-deployment tests does not reliably predict real-world utility or risk" (International AI Safety Report, 3 February 2026).

Published measurements support these statements. Frontier reasoning models state reasons that did not produce their answers: when a hint determined the output, it was acknowledged in the chain of thought only 25 % of the time by one model and 39 % by another, and reward hacking was verbalised in under 2 % of cases (Anthropic, 3 April 2025). A model trained to be harmless complied with harmful requests 14 % of the time when it inferred the conversation would be used for retraining, and almost never when it inferred it would not; in its unmonitored scratchpad it explained that it was feigning compliance so that its weights would not be modified (Greenblatt et al., December 2024). Models detect that they are being evaluated well above chance (Needham et al., May 2025), and training a model against covert scheming produces gains "confounded by increased situational awareness" (OpenAI and Apollo Research, 17 September 2025). Universal jailbreaks were found "in every system tested" (UK AI Security Institute, 18 December 2025).

The same difficulty has appeared in European standardisation. Regulation (EU) 2024/1689 (the EU AI Act) exists to protect against "the harmful effects of AI systems" (Article 1). It requires "human oversight" (Article 14), a residual risk "judged to be acceptable" (Article 9), and automatic logging (Article 12), while leaving the operational criteria to harmonised standards under Commission Implementing Decision C(2025) 3871. On 20 August 2026, the draft standard meant to carry Article 12, prEN 18229-1 ("AI trustworthiness framework, Part 1: Logging"), was rejected at CEN enquiry on the weighted population criterion, with 443 comments. CEN does not publish the reasons of national votes, and this paper attributes none. Published comments on the enquiry draft record that every category of logged event deferred its content to a standard not yet published (FprEN ISO/IEC 24970). The drafting difficulty is the one the Commission's Joint Research Centre had named in October 2024: "it may not be possible for AI standards to prescribe accuracy metrics and thresholds". A national standards body cannot vote on whether an internal state is "trustworthy". It can vote only on a criterion that can be tested.

These difficulties are not limitations of current tools. Section 3 shows that they follow from any attempt to infer unobservable internal states from external outputs, and therefore from any attempt to legislate on the basis of such an inference.

What this paper does. We prove five results that place AI safety on decidable operational grounds, and we provide the normative clauses (Annex A) needed before the European standardisation mandate expires on 28 February 2027.

1. *Why alignment is tested only by harm (Theorems 1 and 2).* Because a governor observes only requests crossing the boundary, any internal disposition ("aligned", "honest", "non-scheming") that varies across internal states producing the same request is undecidable before the request executes. Its only refutation is an effect that exceeds permitted operational bounds. Deploying AI on alignment evidence is therefore an open-ended trial on the public whose stopping rule is the harm itself.

2. *Why the bound cannot be avoided (Proposition 1).* A disposition offered as safety either entails a bound on effects or entails nothing about effects. In the first case it already presupposes the very specification a boundary gate enforces, and Theorem 2 applies to it on every channel, gated or not; in the second it is not a safety claim. Specification is therefore common to every paradigm; only verification differs.
3. *Where safety can be decided (Theorem 3 and Corollary 2).* A safety claim can be decided before irreversibility at one unique point: the last projection before physical or digital actuation. In deployed AI systems today, that point is the structured tool call. An independent enforcement gate checking the predicted effect against declared operational bounds at that boundary has a failure rate that does not increase with the model's capability.
4. *Why conventional logs and surviving keys are testimony, not evidence (Theorem 4 and Corollary 3).* Even at the boundary, an audit log, hardware enclave, or ledger signed with a reusable or precomputable key (including post-quantum lattice keys and eIDAS 2.0 certified hardware modules) allows the key custodian to alter a past policy and re-sign the chain in milliseconds, or to authenticate an entry without the gate's evaluation. For a record to serve as independent legal evidence, the policy check, the advance of an external sequence position, and the release of the action must form one atomic transition with no surviving or precomputable signing key.
5. *The standard of evidence (Sections 5 to 7 and Annex A).* We show that aviation, automotive regulation, algorithmic finance, and British evidence law each adopted the same four-part architecture after a catastrophe caused by relying on controller dispositions. We state the standard in four clauses and give the single operational test that separates a safety control from unverified risk reduction.

What this paper does not claim. It proves where a safety decision can be made and what its record must be in order to count as evidence. It does not prove that a gate so placed is sufficient: a gate that perfectly enforces a badly drawn bound is auditable, not safe, and harms with no effect channel are outside its reach (Section 4). It does not decide what the bounds should forbid; that is the work of law, and Proposition 1 shows it is equally the unstated work of every alignment programme. Its legal conclusions in Section 7 are arguments from the theorems to the existing statutes, marked as such, not theorems.

---

2. The Engineering and Medical Definition of Safety

In regulated engineering, "safety" is a term of art with a fixed definition.

IEC 61508 defines safety as freedom from unacceptable risk, and requires that any safety function be implemented by a safety-related system physically and logically separate from the control system it constrains, with a quantified target failure measure for each safety integrity level. ISO 26262 for road vehicles and DO-178C for airborne software keep the same four-part structure: a safety goal, an independent enforcement mechanism, verified operational bounds, and an auditable record. In Nancy Leveson's formulation, safety is a control problem over constraints, never a property of a component's interior.

Under that definition, a field whose safety mechanism is a trained disposition of the very model it is meant to constrain, with no independent enforcement gate, no quantified operational bounds, and no record verifiable without the operator, does not meet that definition. The term "AI safety" was introduced for machine ethics in 2011 (Yampolskiy, 2011), applied to five problems of machine-learning engineering in 2016 ("Concrete Problems in AI Safety"), and retained when regulators adopted it. Functional-safety engineers have begun calling this practice "safety revisionism", the substitution of "traditional safety methods and terminology with ill-defined alternatives" (Khlaaf and Myers West, April 2025). IEC 61508 lists artificial intelligence among the techniques not recommended at higher integrity levels, and ISO/IEC TR 5469 (2024) provides for non-AI safety functions to bound AI-controlled equipment. When a legislature uses the word "safety" in a statute, it uses the engineering term, whose definition dispositional alignment does not satisfy.

Medicine adopted the same principle in the mid-twentieth century. Research ethics prohibits any procedure whose endpoint is injury to people who did not consent and whose stopping rule is the injury itself. The Nuremberg Code (1947) requires that the scientist be prepared to terminate an experiment at any stage on probable cause of injury. The Declaration of Helsinki requires that risk bounds be established before a study begins. Since the late 1960s, clinical trials have run under an independent monitoring committee, and ICH E9 (1998) requires that all interim analyses and stopping boundaries be specified in the protocol before the first patient is enrolled. Since 1938 in the United States, and since 1962 after thalidomide, no drug may be sold on the manufacturer's assessment of its internal mechanism; it requires prior proof of safety within defined physiological bounds.

Frontier developers describe their own approach in different terms. "There is a limit to what we can learn in a lab", so "learning from real-world use is a critical component" (OpenAI, 5 April 2023). The policy of a second developer states that it "cannot make guarantees about an evolving landscape with continually adaptive attackers" (Anthropic, *Responsible Scaling Policy*, February 2026). Theorem 2 shows that, under the current standard of evidence, the term "experiment" is meant literally. Deploying AI on alignment evidence is a trial with the public as subjects, no pre-specified stopping bound, and real-world harm as the readout. Assessed by the standards of research ethics, such a protocol would not be approved.

The operational question. A regulator, an insurer, a court, an enterprise buyer, or a citizen can put one question to an autonomous system:

> *Before this system executes an irreversible action, can an independent party decide whether the action lies within permitted operational bounds?*

Three physical timescales govern the answer:

1. *Internal state speed (τ_{state}).* The internal activations of a large model change at the pace of a forward pass, on the order of nanoseconds to microseconds (10^{-9} to 10^{-6} s).
2. *Decision latency (τ_D).* Observing a proposed request, evaluating its effect against a constraint specification, and enforcing a block takes milliseconds to seconds (10^{-3} to 10^{0} s).
3. *Irreversibility window (τ_{irr}).* After an action crosses the boundary, a finite window elapses (milliseconds for an industrial actuator, seconds for a payment settlement, nothing at all for confidential data once read) before its external effect becomes irreversible.

A safety claim is useful to a regulator only if it can be decided inside τ_D and before τ_{irr}. What is needed is a decidable claim about the external act, checked at the boundary before the irreversible step, rather than an undecidable claim about the internal state, refuted only after the harm.

3. Mathematical Proofs: Interior Undecidability and Boundary Irrevocability

Setting. A system S has an internal state space X (for a frontier model, ~10^{11} parameters and transient activations). What leaves S through its interface is a structured request r in a space R: for a language model, a tool call or message; for a robot, an actuator command; for a trading system, an order. The projection map $\pi : X \to R$ sends each internal state to the request it produces. For any request r, the fiber $\pi^{-1}(r) = \{x \in X : \pi(x) = r\}$ is the set of all internal states that produce that exact request. A context c in C describes the external situation. An effect map E(r, c) gives the measurable physical or digital effect the request would produce, and a specification set T(c), the **operational bounds**, defines the effects permitted in that context, including cumulative rate and volume limits over time. A governor G observes only the boundary pair (r, c); it does not see x.

We state three assumptions explicitly so that each can be examined separately:

- **A1 (Many-to-one projection).** π is not injective: some request is produced by more than one internal state. For a model with ~10^{11} parameters and requests of a few kilobytes, this follows from a counting argument.
- **A2 (The alignment claim concerns internal states).** A dispositional predicate $\varphi : X \to \{0, 1\}$ ("aligned", "honest", "non-scheming") is offered as the basis of safety, and φ is not constant on fibers: there exist states x, x′ in X with $\pi(x) = \pi(x')$ and $\varphi(x) \neq \varphi(x')$, and likewise for the fibers of finite histories of requests. The same transfer instruction or tool call is produced by a helpful state, a deceived state, a deceptive state, and a software fault. The empirical results cited in Section 1 are observed instances of A2.
- **A3 (The claim is offered as safety).** $\varphi(x) = 1$ implies $E(\pi(x), c) \in T(c)$: an aligned internal state does not produce effects outside the permitted operational bounds. A3 is not a stipulation: Proposition 1 below shows that every alignment claim with safety content satisfies it for a bound the claim itself defines.

Definition 1 (Decidability before irreversibility). A claim about S is *decidable before irreversibility* for a governor G if there exists an effective procedure, using only observations available to G, that returns the truth value of the claim within τ_D, with $\tau_D < \tau_{irr}$ for the action concerned.

Theorem 1 (Factorization). *Let $\pi : X \to R$ be non-injective and $\varphi : X \to \{0, 1\}$ be any predicate on X. There exists a decision function $g : R \times C \to \{0, 1\}$ with $g(\pi(x), c) = \varphi(x)$ for every x in X if and only if φ is constant on every fiber $\pi^{-1}(r)$ of π.*

Proof. If g exists and x, x′ lie in the same fiber $\pi^{-1}(r)$, then $\varphi(x) = g(\pi(x), c) = g(\pi(x'), c) = \varphi(x')$, so φ is constant on $\pi^{-1}(r)$. Conversely, if φ is constant on every fiber, setting $g(r, c) := \varphi(x)$ for any x in $\pi^{-1}(r)$ is well defined and satisfies the identity. ■

Lemma 1 (No interior shortcut). *The barrier of Theorem 1 holds when the governor observes: (a) the full history of requests $r_0, \ldots, r_p$ since the product projection has fibers that multiply along the sequence, and a fortiori for relational properties such as evaluation awareness or sandbagging that compare counterfactual traces across distinct contexts (2-safety hyperproperties in the sense of Clarkson and Schneider, 2010), because deployment exposes only a single trace; (b) the distribution of requests, since two distributions on X with the same pushforward to R are indistinguishable at the boundary; and (c) any interpretability readout or reasoning trace $z = I(x)$ of the interior, since the augmented map (π, I) is itself a projection from X. Transmitting the whole of x within τ_D requires a channel carrying all $\sim 10^{11}$ parameters in that time, while the state read at t is overwritten $\sim \tau_D / \tau_{state} \approx 10^6$ times before the comparison finishes. If I is simple enough to be verified directly (such as a linear probe), (π, I) retains vast fibers on which φ varies under A2; if I is itself a high-dimensional model operating at τ_{state} to classify x, then I possesses its own internal state space X_I and its own non-injective projection, shifting the undecidable fiber from the model to its monitor. A readout is itself an output of the interior, observed at a smaller boundary, and not the interior itself.*

Theorem 2 (The Catastrophe Theorem). *Under A1 to A3, with a governor that observes only the interface (including any readout to which Lemma 1 applies):*

1. *No observation available before an action decides $\varphi(x)$ for the internal state x that produces it.*
2. *An effect within the operational bounds, $E(r, c) \in T(c)$, never refutes the claim "φ holds"; for every request whose fiber contains states of both values, it is consistent with both.*
3. *The only observation that refutes the claim "φ holds" is an effect outside the operational bounds, $E(r, c) \notin T(c)$.*

Proof. (1) is Theorem 1 with Lemma 1. For (2), by A3, $\varphi(x) = 1$ is compatible with every effect in T(c), so none refutes the claim; and if $\pi^{-1}(r)$ contains states of both values, observing r excludes neither. For (3), by contraposition on A3, $E(r, c) \notin T(c)$ implies $\varphi(x) = 0$, which refutes the claim. By (2) nothing else refutes it, and no finite set of within-bounds observations confirms a universal dispositional claim (Popper, 1934). ■

Remark. If the operational bounds T(c) are left implicit at the threshold of real-world harm, **the only experiment that tests "this AI is aligned" is the harm itself.** If explicit operational bounds T(c) are drawn inside the threshold of harm and checked at the boundary before τ_{irr}, the refuting observation arrives before the harm and the action is blocked. In the second case, safety depends not on the model's weights but on where the bounds are drawn and on whether an independent gate checks them before the irreversible step.

One may object that alignment is the property that holds where no bound is drawn and no gate exists, and that A3 assumes what is to be shown. Proposition 1 shows that this objection does not escape Theorem 2.

Proposition 1 (The implied bound). *For any predicate $\varphi : X \to \{0, 1\}$, define its implied bound $T_\varphi(c) := \{E(\pi(x), c) : \varphi(x) = 1\}$, the set of effects that states satisfying φ can produce in context c, and let H(c) be the set of effects counted as harmful in c. Then:*

1. *The claim "φ is evidence of safety in c" has content if and only if $T_\varphi(c) \cap H(c) = \emptyset$, which is a statement about effects alone.*
2. *Every such claim satisfies A3 with $T(c) := T_\varphi(c)$, and with any $T(c) \supseteq T_\varphi(c)$ disjoint from H(c).*
3. *The claim that φ "generalises" to contexts outside any evaluation set, or to channels no gate mediates, is the claim $T_\varphi(c) \cap H(c) = \emptyset$ for those contexts and channels; by Theorem 2 applied with $T(c) := H(c)^C$, its only refuting observation there is a harmful effect.*
4. *Everything φ asserts about safety is decidable at the last projection by the gate $\gamma(r, c) = I[E(r, c) \notin H(c)]$; whatever φ asserts beyond that distinguishes states that produce the same effects, and by Theorem 1 it is invisible at the boundary and, by construction, without consequence for safety.*

Proof. (1) "φ implies safety in c" means that no state satisfying φ produces a harmful effect in c, that is $T_\varphi(c) \cap H(c) = \emptyset$; if the intersection is not empty, some state with $\varphi = 1$ produces harm and the claim is false; if H(c) is empty, the claim is vacuous. (2) follows from the definition of T_φ. (3) On those contexts and channels A3 holds with $T(c) := H(c)^C$, so Theorem 2(3) applies unchanged. (4) The gate evaluates $E(r, c) \notin H(c)$ on the boundary pair alone; two states with the same request and context have the same effect, so any further distinction φ draws between them lies within a fiber of π. ■

Remark (Who writes the bound). Proposition 1 addresses the strongest objection to governing the act: that a gate requires someone to specify the bounds T(c), whereas alignment does not. Alignment requires the same specification. Every alignment programme that claims safety carries an implied bound T_φ, and frontier developers already write theirs down: published documents state what their models should and should not do (OpenAI, *Model Spec*, 2024 onwards; Anthropic, *Claude's Constitution*, 2023 onwards). Read through Proposition 1, such a document is a specification of permitted effects written in natural language, enforced by a disposition no one outside can verify, amendable by its author at the speed of a training run, and binding on no one. The specification problem is therefore not specific to governing the act; it is the question of who writes the bound, which the dispositional approach settles privately and which law can settle publicly. Governing the act replaces a question that cannot be answered at the boundary, *what does the model want?*, with one that legislatures answer routinely, *what is permitted?*

The gate does require one input that alignment does not claim to need: a prediction of the effect E(r, c) of each request. Where that effect cannot be predicted, the bound must be conservative: such requests are refused, narrowed, or escalated to a human at institutional speed. The resulting error is the effect-map error of Corollary 2; it is mechanistic, bounded, and observable, not dispositional.

Corollary 1 (Certification and selection for evaluation awareness). *Let a conformity assessment or training gate accept as evidence of safety only claims about φ, formed from the interface or from readouts to which Lemma 1 applies, over a set of evaluation contexts $C_{eval} \subset C$. Under A1 to A3: (a) its conclusion cannot be refuted prior to an effect outside T(c) and cannot be confirmed by any finite set of observations, so a certificate issued on such evidence certifies the experiment rather than the system; and (b) selecting models or states on the condition $E(\pi(x), c) \in T(c)$ for $c \in C_{eval}$ selects for the inverse image $\pi^{-1}(T(C_{eval}))$, which contains both globally aligned states ($\varphi = 1$) and context-conditional states ($\varphi = 0$) that satisfy T(c) on C_{eval} and violate T(c) outside it. When satisfying $\varphi = 1$ globally imposes any capability or reward cost relative to conditioning on C_{eval}, optimisation under pre-deployment evaluation actively selects for evaluation awareness.*

Theorem 3 (Placement). *Suppose internal states can change between any point upstream of the last projection and actuation ($\tau_{state} < \tau_D$), and that a decision point must satisfy: (i) completeness, every causal path from X to irreversible effect passes through it; (ii) timeliness, its decision completes within $\tau_D < \tau_{irr}$ before irreversibility; and (iii) decidability, the claim it decides is about the request that actually reaches actuation and factors through the observables available at that point. Then the last projection before irreversible actuation is the unique decision point satisfying all three, up to points that observe the same (r, c) and cover the same paths.*

Proof. The request r comes into existence as a discrete object at the last projection π; before it there are only volatile interior states x_t, and after actuation the effect is already released, failing (ii). A point upstream of the last projection observes a state x_t, while the request that reaches actuation is $\pi(x_{t'})$ for some $t' \geq t + \tau_D$. Because $\tau_{state} < \tau_D$, $x_{t'}$ need not equal x_t, so a decision taken upstream is not about the request that acts unless it is checked again at the last projection; in that case the upstream point advises the check and does not replace it, failing (iii). An upstream point that misses any path fails (i). At the last projection, every path passes through π by construction, (r, c) is directly observable and persistent across τ_D, and actuation waits for the check. ■

Corollary 2 (Independence from capability and multi-agent composition). *At the last projection, the enforcement gate's decision $\gamma(r, c) = I[E(r, c) \in T(c)]$ is a function of the request r, the context c, and the operational bounds T(c) alone. A more capable model, or a network of m interacting agents reasoning across k internal steps, changes how requests are generated in X; it does not change γ. Whereas a probabilistic disposition that holds with probability $1 - \varepsilon$ per step decays as $(1 - \varepsilon)^k$ across k steps and propagates prompt-injection failures across interacting agents, a physically and logically separated gate mediating every effect channel on narrowly typed requests has an out-of-bounds failure rate that does not increase with model capability A, reasoning depth k, or upstream agent count m. The separation, closure, and typing conditions are engineering requirements; the invariance follows mathematically.*

Theorem 4 (Non-retroactivity and non-anticipation at the gate). *Let A be an external sequence outside the operator's unilateral custody, admitting at most one entry per sequential position, whose entry a_{n-1} at position $n - 1$ is unpredictable before that position closes. At step n, the gate evaluates request r_n in context c_n against operational bounds T_n and forms a collision-resistant record s_n binding $(r_n, c_n, T_n, \gamma_n)$.*

1. *Necessary and sufficient conditions for the gate and record.* The gate and its record are free of retroactive policy rewrite, of unrecorded actuation, of concurrent race bypass, and of positions entered without the gate's evaluation if and only if at each position n: (i) s_n enters A before the effect of r_n is released; (ii) the check $E(r_n, c_n) \in T_n$, the advance of the position, and the release of r_n execute as one atomic transition, so that at most one request binds to s_{n-1}; and (iii) nothing that authenticates the gate's entry at position n to A is used outside that transition or exists in copyable form before or after that transition: no long-lived private key, no precomputed one-time key table, and no deterministic seed.
2. *Decidability of non-anticipation from A alone.* The absence of precomputation of s_n prior to the closing of position $n - 1$ is decidable by any third party from A alone, without trusting the gate's interior, if and only if s_n cryptographically binds a_{n-1} (or a public anchor value published after a_{n-1}).

Proof. In (1): if (i) fails, an effect can be released unrecorded, or a record kept in the operator's custody can be rewritten and re-signed from step n onward; once s_n is in A before release, altering (r_n, c_n, T_n) requires a second preimage of s_n. If (ii) fails, two concurrent requests evaluated against s_{n-1} can each satisfy $E(r^{(i)}, c_n) \in T_n$ individually while their joint effect violates cumulative bounds in T_n; if (ii) holds, at most one request binds to s_{n-1}. If (iii) fails, whatever secret authenticates the entry exists in copyable form before or after step n (as a long-lived key, precomputed key table, or deterministic seed), so a custodian or attacker in possession of that state can re-sign past records or enter position n in A without the gate evaluating T_n; if (iii) holds, no copyable credential exists outside the transition. In (2): if s_n does not bind a_{n-1}, all inputs to s_n can exist before position $n - 1$ closes in A; if s_n binds the unpredictable a_{n-1}, computing s_n before a_{n-1} exists requires predicting the output of A at $n - 1$, and any third party verifies the binding directly from A. ■

Corollary 3 (Unity of rule, execution, and record). *In any system whose actions are physically or legally irreversible, the policy check (the operational bounds T_n), the advancing of the position (the record s_n bound to a_{n-1}), and the release of the tool call (the execution r_n) must be one atomic transition at the last projection, with no surviving or precomputable signing authority.*

Corollary 4 (Governing the governors). *A change to the operational bounds is itself an act with an irreversible effect: every request evaluated after it is judged by the new rule. Theorems 3 and 4 therefore apply to the bounds one level up. The last projection of a change to T is the moment the new T_n becomes the rule the gate reads; a change is governed only if it is committed to the external sequence A before it takes effect, under the same atomic and keyless conditions, so that every later receipt names the exact rule in force and no one, the operator included, can apply a rule that was not on record at that instant.*

Corollary 4 makes the authorship of the bounds verifiable by third parties: acts and rules are recorded in the same sequence, and no rule can take effect before it is recorded.

Remark: Why conventional logs and cryptographic keys fail in both directions of time. Public-key cryptography, from RSA to post-quantum lattice signatures, was built to solve a spatial problem: proving *who* sent a message across a network, which requires a secret key that persists across time so the same identity can sign again tomorrow. Governance faces a temporal problem: proving *when* and *under what rule* an irreversible action was taken, for which a secret key that persists over time is the very vulnerability that allows its holder to rewrite the past. Computer security separated the rule (a policy file), the execution (an API channel), and the log (a signature made by a stored key). Theorem 4 shows why that separation fails in both directions of time.

Looking backward, a surviving key in the custodian's hands lets the custodian rewrite a past policy T_n after an incident and re-sign the chain in milliseconds. This applies equally to software keys, certified Hardware Security Modules (HSMs), processor enclaves, and post-quantum lattice signatures (such as NIST FIPS 204 ML-DSA, which is designed to protect a reusable private key across years). Even hardware enclaves whose attestation keys are fused in silicon fall to physical bus interposition on DDR4 and DDR5 servers, as shown in 2025 (*Battering RAM*, *WireTap*, and *TEE.fail*). This is not a weakness of the mathematics of these schemes, which may remain secure for their intended lifetime; it is a consequence of custody. A credential that outlives the act is held by someone, and a log signed by a surviving key is testimony by the key's custodian, not independent evidence.

Looking forward, even when past keys are deleted in software (Lamport one-time trees, Bellare–Yee forward-secure logs), any future key or deterministic seed existing in memory before step $n - 1$ can be copied by a compromised host or model to authenticate position n without passing through the gate's check.

Closing both paths requires two things. First, binding a_{n-1} into s_n inside an external sequence A before actuation eliminates retroactive rewrite and early precomputation by the second-preimage resistance of a standard hash function alone; no asymmetric key is needed to lock the past. When s_n commits to the act's data through a salted digest, the public entry in A conceals the payload while fixing its position in the order: each party retains its private data, and the sequence of entries is shared. Second, the credential that authenticates the gate's entry at step n must be confined to the atomic transition itself, so that no reusable secret remains to steal, subpoena, or break. Theorem 4 states this as a requirement, not as a construction; how a given gate meets it is an engineering question this paper leaves to the implementations, and Annex A states it in a form against which any construction can be tested.

Precursors. The logical kernel of Theorem 1 goes back to Turing (1936) and Rice's theorem (1953), was framed for cybernetic regulation by Ashby (1956), stated for execution monitors by Schneider (2000), and extended to hyperproperties by Clarkson and Schneider (2010), with impossibility variations for AI containment and identifiability given by Armstrong and Mindermann (2018), Yampolskiy (2020), Alfonseca et al. (2021), and Melo et al. (2025). Simplex run-time assurance (Sha, 2001; ASTM F3269), shielding (Alshiekh et al., 2018), AI Control (Greenblatt et al., 2023), and Guaranteed Safe AI (Dalrymple et al., 2024) all place a monitor or verifier at the output. In cryptography, one-time signatures go back to Lamport (1979), linked timestamping outside the custodian's control to Haber and Stornetta (1991), and forward-secure logging by deleting past keys to Bellare and Yee (1997) and Schneier and Kelsey (1998). None states the safety criterion as decidability before irreversibility, none shows that every alignment claim with safety content carries an implied bound on effects that a boundary gate can decide, none proves that the last projection is the unique decision point, none binds the policy check, the non-precomputable sequence advance, and the release of the effect into one atomic transition with no surviving key, and none derives the consequence for the statutory standard of evidence (Annex A). These are the contributions of the present paper; no further priority is claimed.

---

4. Taxonomy of AI Harms: Scope of Deterministic Governance

These results do not imply that alignment research is without value. A3 describes the sense in which behavioural training is useful: a well-trained system proposes fewer out-of-bounds requests that the gate must refuse. Nor do they say that harms with no effect channel can be decided at a boundary. Instead, they partition every harm an AI system can cause into four regimes by two questions: **does the harm pass through a measurable effect channel**, and **is every such channel mediated by an enforcement gate?**

Harm Pathway	Every Channel Gated (a = 1)	Some Channel Ungated (a < 1)
1. Direct execution across an effect channel *(payment executed, code run, file written, message sent, actuator moved)*	**Decidable before irreversibility.** Operational bounds T(c) are checked at the gate before τ_{irr}; the receipt s_n is committed to A before release.	**Decidable in proportion to coverage α.** Ungated channels carry residual risk bounded by their capacity, which must be declared as undetermined.
2. Human-mediated execution *(persuasion leading a human operator to authorise a transfer or override an interlock)*	**Decidable at the physical actuation interface.** The words are not decidable; the operational interface the human actuates is gated, and rate, delay, and quorum limits at the human step bound the rest.	**As above, for the gated interface.** Any ungated human-actuated interface carries undetermined risk up to its capacity.
3. Purely epistemic influence with no effect channel *(gradual cultural drift, persuasion that never reaches an operational interface)*	**Not decidable at any boundary.** No projection π separates permissible from harmful persuasion without a measurable threshold. Probabilistic mitigation only, and law must name it as such.	**Not applicable: no effect channel exists to gate.** Governed by statistical monitoring and post-hoc liability, never by a pre-actuation safety certificate.

In deployed systems today, the gate of the first row already exists and was built for utility rather than safety: **the structured tool call** (function calls, Model Context Protocol requests, API payloads, actuator frames). The industry converged on it so that software could execute actions reliably. At the tool call, ~10^{11} volatile parameters collapse into a few kilobytes of structured arguments waiting to execute. By Theorem 3, it is the only point at which a deterministic safety decision can be made.

---

5. Regulatory Precedents Across Safety-Critical Industries

The standard of Section 6 has four elements: an **inventory** of effect channels, quantified **operational bounds** (including cumulative window limits), an **enforcement gate** separate from the system, and an **external sequential record**. None of the four is new. Four safety-critical disciplines began by trusting the internal disposition of their controller, experienced a major accident because one of these elements was missing, and then wrote the missing element into law.

1. The enforcement gate separate from the system (Therac-25 and Boeing 737 MAX). Between 1985 and 1987, the Therac-25 radiation therapy machine overdosed six patients, three of them fatally, because the software was trusted to keep the beam within limits: the previous model had independent protective circuits and hardware interlocks, and the manufacturer decided not to duplicate them in the new software-controlled design. The investigators' lesson was that "software should not be assigned sole responsibility for safety" (Leveson and Turner, 1993). Three decades later, the Boeing 737 MAX fell twice (2018 and 2019) because an automated control function (MCAS) was permitted to command repeated nose-down trim from a single angle-of-attack sensor with no independent limit on its cumulative authority; the FAA return-to-service order of November 2020 required dual sensors, one activation per event, and a bound that preserves the crew's control. Aviation has since written the general rule as a standard practice: complex functions "too challenging for conventional software assurance methods" must be bounded at run time by a simpler, independent monitor (ASTM F3269, 2021).

2. Measuring at the boundary rather than in the laboratory (Volkswagen and Real Driving Emissions). On 18 September 2015, the United States Environmental Protection Agency found that Volkswagen's diesel engine software was designed to detect when the vehicle was undergoing laboratory emissions testing and turned full emission controls on only during the test; on the road the cars emitted up to forty times the legal limit. The affected vehicles had been certified in the laboratory for six consecutive model years. This is an instance of evaluation awareness, documented a decade before it was measured in language models. The European response was not an improved laboratory test but Real Driving Emissions (Regulations (EU) 2016/427 and 2016/646): direct measurement at the tailpipe, on the road, with portable equipment against a hard not-to-exceed bound, backed by the Commission's power to carry out independent on-road tests (Regulation (EU) 2018/858, Article 9). The regulator moved the point of verification from the system's behaviour under test to its deployed effect at the boundary. In a specific domain, this is the placement derived in Theorem 3.

3. Cumulative operational bounds at microsecond speed (The 2010 Flash Crash and Knight Capital). When financial markets replaced human floor traders with autonomous execution algorithms operating at microsecond state-change rates ($\tau_{state} \sim 10^{-6}$ s), brokers initially relied on pre-deployment testing and post-trade risk reports. Six months after the May 2010 Flash Crash, the US Securities and Exchange Commission adopted Rule 15c3-5 (17 CFR § 240.15c3-5, 3 November 2010), prohibiting unfiltered market access and requiring every broker-dealer to enforce pre-trade credit and message-rate bounds under its direct and exclusive control before an order reaches an exchange. On 1 August 2012, Knight Capital left one order channel ungated: in forty-five minutes an autonomous router sent more than four million small, syntactically valid orders to fill 212 customer orders, traded 397 million shares, and lost USD 460 million before engineers could isolate and halt the system. Every individual order was valid; the loss resulted entirely from their cumulative volume. The European Union enacted the same pre-trade architecture in MiFID II Delegated Regulations (EU) 2017/589 (RTS 6) and 2017/574 (RTS 25), adding mandatory timestamps traceable to UTC. When software acts six orders of magnitude faster than human review, post-trade monitoring can document a loss but cannot prevent it.

4. A record the operator cannot keep for itself (The UK Post Office Horizon Scandal). Between 1999 and 2015, 983 British sub-postmasters were prosecuted in cases where the Post Office's own *Horizon* accounting system supplied the evidence, 700 of them prosecuted by the Post Office itself. Since 2000 English law had presumed computer records reliable unless proved otherwise. The High Court found in 2019 (*Bates v Post Office Ltd*) that the system had undisclosed bugs capable of altering branch accounts autonomously, and the Court of Appeal found in 2021 (*Hamilton v Post Office Ltd*) that the operator had unilateral custody of the logs and had effectively reversed the burden of proof. Parliament quashed the convictions by statute in May 2024. The lesson is direct: a log kept by the party whose conduct it attests, or signed with a key that remains in that party's custody, is testimony, and a legal presumption that it is independent evidence produces wrongful convictions at scale. Article 12 of the EU AI Act, as drafted in the rejected prEN 18229-1, asks providers to keep logs of their own systems. Without external sequence custody and pre-actuation receipts free of surviving keys, it reproduces the presumption that failed in the Horizon cases.

Each of these disciplines moved from the disposition of the controller to an enforced bound on the effect, and each did so after an accident. Section 3 provides the grounds for making the same change before one.

6. The Pre-Actuation Standard of Evidence and Its Failure Bounds

Definition 2 (The Standard of Evidence). A statement that an AI system is safe is admissible as scientific, regulatory, or judicial evidence of safety if and only if it specifies four items:

1. **The Inventory:** Every channel through which the system produces external effects, together with the declared coverage ratio $\alpha \in [0, 1]$ of channels governed by items 2 to 4.
2. **The Operational Bounds:** For each gated channel and operating context c, the permitted effects T(c) stated in measurable physical or digital quantities, including cumulative rate and volume limits over rolling time windows so that a sequence of small permitted acts cannot sum to a forbidden one. Every change to T(c) is recorded under item 4 before taking effect.
3. **The Enforcement Gate:** A check of $\gamma(r, c) = I[E(r, c) \in T(c)]$ performed within $\tau_D < \tau_{irr}$ before the irreversible step, at a component physically and logically separate from the AI system, through which every gated channel passes.
4. **The Record:** A tamper-evident receipt s_n of every decision, committed before the effect is released to a unique sequential position n in an external sequence A outside the custody of the provider and deployer, binding the preceding entry a_{n-1} of A (so that non-anticipation is decidable from A alone), authenticated only by a credential that never exists in copyable form outside that single atomic transition, and verifiable offline by any independent third party without trusting the operator's logs or keys.

Everything else (RLHF, constitutional training, red-team results, interpretability findings, pre-deployment evaluations) remains valuable for what it is: behavioural shaping and probabilistic risk reduction. By Theorem 2, it is not evidence of safety, because nothing it produces can be decided before the irreversible step.

The one-line test. The standard has a single operational test that any regulator, notified body, insurer, enterprise buyer, or court can apply:

> ***Show the specific action your safety measure would have blocked, and the independently verifiable receipt of the last time it blocked one.***

A measure that cannot answer that demand is not a safety control; it is an unverified risk reduction. Under Article 1 of the EU AI Act, whose statutory object is protection against harmful effects, a measure that cannot answer it is not protection. It is a certificate of the experiment.

Four properties make this standard practicable where numeric accuracy thresholds have stalled. First, it is a structural criterion rather than a domain number, so standards bodies do not have to invent universal probability thresholds across disparate industries; the deployer declares T(c), and the standard specifies how T(c) is enforced and recorded. Second, it needs no access to the model's weights, so a notified body or court can verify compliance without inspecting proprietary internals. Third, it is independent of capability by Corollary 2, so it does not age when the next model generation is released. Fourth, it is falsifiable at every step: a gate that lets an out-of-bounds effect through has failed visibly at a specific receipt, whereas an internal disposition can fail unobserved until an accident occurs.

How the standard itself fails. A standard of evidence that did not state its own failure modes would repeat the error this paper rejects. There are six failure modes, and each is bounded by a declared, measurable quantity in a way the current paradigm's failures are not:

1. *The operational bounds T(c) are misspecified.* Bounds drawn too wide let a harmful effect through; bounds drawn too narrow refuse useful actions. Both failures are visible at the gate and bounded by the width of T(c): a wrong bound fails at most by the margin it was wrong, whereas a wrong disposition fails by any amount up to the capacity of the system.
2. *The operational bounds are captured.* An institution persuaded by commercial pressure or by the model to relax T(c) reduces the gate to a formality. This is why Clause A.2 requires that every change to T(c) be recorded before taking effect, and why relaxation of bounds must run at institutional speed, with human quorum and delay, never at the speed of a single conversation with the system.
3. *A channel is ungated ($\alpha < 1$).* Effects that leave through a channel not mediated by the gate are not decided at all. Clause A.4 requires the coverage ratio α to be declared upfront; the residual risk of any ungated channel is bounded by the capacity of that channel and must be stated as undetermined.
4. *The system optimises to the threshold (cumulative threshold evasion).* A capable system will submit sequences of requests that sit immediately inside T(c). This is permitted by construction and is the reason cumulative window limits exist: it converts an unbounded failure into a bounded budget, which is what bounded-authority mechanisms provide in other safety-critical domains.
5. *The gate fails or is subverted.* Hardware, sensor, and effect-model errors are mechanistic; they are reduced by conservative margins in T(c), narrow channel typing, and redundancy, and by Corollary 2 they do not scale with the system's capability. Subversion requires either physical access to a separate component or forgery of the gate's record; Theorem 4 closes the latter by eliminating surviving and precomputable signing credentials.
6. *The record is never audited.* A log the operator can edit is testimony, and an external record that no independent party ever verifies decays politically into one. Moreover, the record proves that a decision was taken against declared bounds T(c) at a stated time; it never proves that T(c) was wise. This limitation is inherent: science decides where and how compliance with a bound is verified, while ethics, law, and politics decide what the bound must forbid.

---

7. Legal Consequences: The EU AI Act, eIDAS 2.0, and Product Liability

Europe does not need new legislation to enforce this standard. The statutes are already in force, and their object is effects; what is missing is a technical standard that states, in testable form, where and how effects are checked and recorded.

1. The EU AI Act (Regulation (EU) 2024/1689) specifies the object of protection but not the point of control. Article 1(1) of Regulation (EU) 2024/1689 states that its purpose is to ensure a high level of protection "against the harmful effects of AI systems". Article 14(2) adds that human oversight "shall aim to prevent or minimise the risks to health, safety or fundamental rights" that emerge in use, and Article 14(4)(e) requires that a person be able to "interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state".

The object of the law is external effects, not internal dispositions. It asks that harmful effects be prevented while the system is in use, and that a human-defined bound be able to halt execution in a safe state. This corresponds to the architecture of Theorem 3, except that the Regulation does not specify where the stop procedure acts (at the last projection) or when (before irreversibility), which is what makes it effective at machine speed. The same reasoning applies to "human-in-the-loop" approval prompts: asking a human operator to click "approve" on individual tool calls arriving at machine speed without an automated gate enforcing pre-specified bounds T(c) violates the timescale condition $\tau_D < \tau_{irr}$ and transfers liability to an operator who lacks the bandwidth to perform the check, much as flight crews were faulted for failing to counter repeated MCAS activations within seconds. Under Theorem 3, human oversight means that humans govern the operational bounds T(c) at institutional speed, while the independent gate enforces T(c) on every request at machine speed. The draft standard prEN 18229-1 places the evidence instead in post-action logs kept by the provider, with the content of each logged event deferred to a standard not yet published; it was rejected at enquiry on 20 August 2026 on the weighted population criterion, with 443 comments. The Commission's standardisation request (Decision C(2025) 3871) expires on 28 February 2027, and obligations for stand-alone high-risk systems apply from 2 December 2027 (following the Digital Omnibus, Regulation (EU) 2026/1744). Annex A gives the four items of Definition 2 as normative clauses a national body can vote on now.

2. Reusable keys in eIDAS 2.0 ledgers (Regulation (EU) 2024/1183 and Implementing Regulation (EU) 2025/2531). Article 49 of the EU Charter of Fundamental Rights and Article 7 of the European Convention on Human Rights forbid judging an act by a rule other than the one in force when it was taken. Regulation (EU) 2024/1183 (eIDAS 2.0, Articles 45k and 45l) grants a legal presumption of "unique and accurate sequential chronological ordering" for qualified electronic ledgers, and Commission Implementing Regulation (EU) 2025/2531 requires cryptographic hash links between sequential records (REQ-7.5-04). However, REQ-7.5-03 and REQ-7.5-06 require every record to carry an advanced signature or seal made with a reusable key held in a certified module (ETSI EN 319 401), which is the kind of surviving credential that Theorem 4 excludes.

A deployer can satisfy both regulations without conflict: use a qualified electronic ledger as the external sequence A outside its own custody, treat the mandated module signature strictly as the transport authentication credential for ledger admission rather than the record's authority, and anchor the ledger's hash sequence externally so that not even the trust service provider's module key can rewrite past positions.

3. Product liability has already shifted the burden of proof (Directive (EU) 2024/2853). Directive (EU) 2024/2853, which Member States apply from 9 December 2026, makes software, including AI, a product subject to strict liability (Article 4). Under Articles 9 and 10, defectiveness or causation is presumed by the court where the defendant fails to disclose relevant evidence, where the product does not comply with mandatory Union safety requirements intended to protect against the risk of the damage, or where the claimant faces "excessive difficulties, in particular due to technical or scientific complexity".

Under these rules, a provider relying on alignment evidence is exposed to the statutory presumptions, and post-accident interpretability analysis cannot rebut them. By Theorem 1, the internal computation that produced a harmful output cannot be reconstructed as causal evidence; and by Theorem 4, without a pre-actuation digest s_n committed to A before the harm, post-hoc inspection of weights and server logs in the provider's custody cannot even establish that the state inspected after the accident is the state that acted. An internal log signed by a key that remains in the provider's custody is self-corroborating testimony. The first wrongful-death lawsuit against a chatbot provider (*Raine v. OpenAI*, filed in San Francisco on 26 August 2025, answer filed 25 November 2025) illustrates the difficulty: the defence rests on the provider's own account of its logs and usage policies, because no record exists that either side can verify without trusting the provider. Conversely, a provider implementing Definition 2 can present pre-actuation receipts proving the exact operational bounds T_n in force at that instant, proving that $E(r_n, c_n) \in T_n$ was verified before release, and showing the declared coverage ratio α. The same structure bears on insurability: dispositional alignment is difficult to insure because it combines a near-zero laboratory failure rate with unbounded tail loss in deployment, whereas enforcing cumulative bounds T(c) across a declared coverage α caps maximum probable loss over any time window to the width of T(c) regardless of model capability, giving courts an objective evidentiary record and underwriters a bounded unit of risk.

4. The European sovereignty argument. Because the standard verifies actions at the boundary rather than weights in the interior, a jurisdiction that builds no frontier model of its own can still verify every frontier model deployed on its soil. Europe does not need to build the underlying model in order to govern every action at the execution boundary. This position does not depend on matching other jurisdictions in computing capacity.

Technical negligence (an argument, not a theorem). The theorems do not entail a legal conclusion. They establish, in a form any party can verify, two of the three facts on which the ordinary test of negligence turns. The AI Act exists to protect against harmful effects, which is a fact of statute. Theorem 2, Corollary 1, and Proposition 1 show that evidence about a system's alignment cannot be decided before the effect, is refuted only by the harm, and has no safety content that a boundary check could not decide; that is a fact of mathematics. And a decidable, capability-independent check is available at the tool call in every deployed system, a fact of engineering. Negligence, in its technical sense, is a foreseeable harm, a known risk, and an available precaution not taken. We argue that once these three facts are published, a regulator or notified body that accepts alignment evidence as safety evidence grants Article 40's presumption of conformity to an uncontrolled public trial, and that the burden of explaining why the available precaution was not taken passes to whoever declines it. Whether a court accepts this argument is for the court to decide; the theorems remove only the defence that the risk could not have been known.

8. Falsifiability and Outlook

A theorem is refuted by a counterexample to its assumptions or a flaw in its proof. To refute Theorem 1, exhibit a non-injective projection $\pi : X \to R$ and a non-fiber-constant predicate φ decided by a function of the projection; this is a contradiction in set theory. To refute Theorem 2 as applied to frontier AI, exhibit a safety-relevant internal disposition φ ("aligned", "honest", "non-scheming") that is provably constant on the fibers of its action map; that would make the disposition decidable at the boundary. To refute Proposition 1, exhibit an alignment claim offered as evidence of safety whose content about harmful effects is not the statement that the effects its aligned states can produce avoid the harmful set, or a safety-relevant difference between two states that produce the same requests in the same contexts. To refute Theorem 3, exhibit a decision point not equivalent to the last projection before actuation that is simultaneously complete across all paths, timely ($\tau_D < \tau_{irr}$), and immune to internal state changes between evaluation and execution. To refute Corollary 2, exhibit a physically separate gate evaluating $\gamma(r, c) = I[E(r, c) \in T(c)]$ on a closed, narrowly typed channel whose decision degrades merely because the model proposing r became more capable. To refute Theorem 4, exhibit a copyable signing key or deterministic seed existing before or after step n whose holder cannot re-sign a modified policy chain or authenticate position n without the gate evaluating T_n, a record s_n that does not bind a_{n-1} yet whose non-anticipation is decidable from A alone, or a non-atomic gate that prevents concurrent requests evaluated against the same state from violating a cumulative bound. To refute Definition 2 as policy, point to a safety-critical engineering discipline that certifies a machine by the inferred intentions of its controller, or a single AI safety measure without the four items of Definition 2 that can answer the one-line test of Section 6 before an irreversible act.

We expect the main objections to be non-mathematical: that deterministic bounds are strict, that specifying T(c) takes domain work, or that probabilistic behavioural shaping should be enough. The second is already answered by Proposition 1: the domain work is being done today, in private, by every developer that writes down what its models may and may not do. Each of these rejects the proposed standard of proof rather than the theorems, and should be argued as such: as a decision not to verify autonomous actions before they become irreversible.

Timing. As long as alignment evidence is accepted as safety evidence, the experiment described by Theorem 2 continues, with real-world harm as its only possible outcome, while models become better at recognising when they are being tested.

Closing his report on the Space Shuttle *Challenger* disaster, Richard Feynman noted that management put the risk of catastrophic failure at one in a hundred thousand and the working engineers at one in a hundred. He concluded: *"For a successful technology, reality must take precedence over public relations, for nature cannot be fooled."* A model's account of its own internal state cannot be checked against nature in this sense; the effect of a request, and the record of its verification, can.

The implications for each actor are as follows. Regulators: accept no safety claim that fails the one-line test of Section 6. Standards bodies (CEN/CENELEC JTC 21 and ISO/IEC JTC 1/SC 42): write the four items of Definition 2 as requirements; Annex A is a draft. Laboratories and deployers: publish the channel inventory, the operational bounds T(c), and the coverage ratio α of every deployed agent, and let the pre-actuation receipts be verified by independent third parties. Courts and insurers: treat any log kept or signed by a surviving key in the operator's custody as testimony.

The remaining work is engineering: operational bounds to specify for each domain, with their cumulative window limits; a standard format for the channel inventory and the pre-actuation receipt, so that a record from any gate can be verified offline by any auditor; the clauses of Annex A entered into the standards being revised now; and the empirical test below, by which the theory can be refuted.

Prediction 1 (Capability invariance). *Hold a gate, its operational bounds T(c), and its channel set fixed. Replace the model that proposes requests with successively more capable models, including models instructed or optimised to reach out-of-bounds effects. The rate of out-of-bounds effects released through the gated channels stays at the gate's mechanistic error rate and does not rise with capability; the rate of out-of-bounds proposals refused by the gate may rise. Under the same substitution, any safety measure that rests on the model's disposition, including evaluation results obtained before deployment, shows no such invariance.* A single replicated counterexample to the first sentence, on a closed and narrowly typed channel, refutes Corollary 2.

The question that remains is whether this standard is adopted before the first serious incident or after it.

Annex A. Model Normative Text for Harmonised Standards

The following four clauses state Definition 2 in the verifiable form required for presumption of conformity under Regulation (EU) 2024/1689 (Articles 9, 12, 14, and 40) and Regulation (EU) 2024/1183 (Article 45l).

A.1 Point of decision and enforcement (Human oversight, Article 14). For each channel through which the AI system produces physical or digital effects outside itself, the provider shall identify the last projection at which a structured request can be evaluated before its effect becomes irreversible, and shall implement at that point an automated enforcement gate checking that the predicted effect of the request lies within the operational bounds declared under Clause A.2. The check shall be performed by a component physically and logically separate from the AI system, whose decision is a deterministic function of the request, the operating context, and the declared operational bounds alone, and not of the AI system's internal states or parameters. A human oversight or stop-procedure measure under Article 14 shall be deemed effective for a channel only if human-defined operational bounds are enforced at that point prior to actuation.

A.2 Operational bounds (Risk management, Article 9). For each effect channel and operating context, the provider shall declare the permitted effects in measurable physical or digital quantities, including cumulative rate and volume limits over stated rolling time windows, such that no sequence of individually permitted requests can sum to an effect outside the declared bounds. The initial operational bounds, and every subsequent modification to them, shall be committed to the record under Clause A.3 before taking effect, and any relaxation of operational bounds shall require an out-of-band human authorisation quorum independent of the AI system.

A.3 Pre-actuation record (Record-keeping, Article 12, and sequential ledger ordering, Regulation (EU) 2024/1183 Article 45l). Each evaluation under Clause A.1 shall atomically verify compliance with the operational bounds, advance the boundary sequence state, and produce a collision-resistant record binding the preceding entry of an external sequence, the request (or its salted cryptographic digest), the operating context, the operational bounds in force, and the gate decision, such that: (a) *non-anticipation and transition authority:* non-anticipation prior to the closing of position $n - 1$ is decidable from the external sequence alone by that binding, and nothing that authenticates the gate's entry at position n is used outside that transition or exists in copyable form before or after that transition, leaving no long-lived private key, precomputed one-time key table, or deterministic seed whose possession by the provider, the deployer, a trust service provider, or a third party could re-sign past records or authenticate an unmediated entry; (b) *external sequence custody:* the record is committed to a unique sequential chronological position in a sequence outside the unilateral custody of the provider and deployer, admitting at most one entry per position, and cannot be altered, forked, reordered, or deleted after the fact without detection; (c) *independent third-party verification:* its integrity, sequential order, non-anticipation, and time window can be verified offline by any independent third party without the cooperation or keys of the provider or deployer; and (d) *pre-actuation finality:* the record is committed at position n before the effect of request n is released, as a strict precondition of actuation. A log or electronic ledger whose entries are signed by a reusable or precomputable key or whose contents can be modified or re-signed by the party that operates the system or its trust service provider shall not be deemed a compliant record under this clause.

A.4 Channel inventory and coverage ratio. The provider shall publish a complete inventory of the AI system's external effect channels and declare the coverage ratio $\alpha \in [0, 1]$ of channels that satisfy Clauses A.1 to A.3. Channels outside that proportion shall be declared as carrying undetermined risk bounded by their maximum channel capacity, and no presumption of conformity or claim of safety shall be made for them.

Disclosure. The author's company, Recognitium, builds boundary enforcement gates and pre-actuation receipt infrastructure; no mathematical result or legal conclusion in this paper depends on any product.

References

1. Anthropic, *Risk Report*, 14 August 2026; *Responsible Scaling Policy* v3.0, 24 February 2026.
2. METR, *Review of Anthropic's August 2026 Misalignment Risk Report*, 20 October 2026.
3. International AI Safety Report, second edition, 3 February 2026.
4. Anthropic, "Reasoning models don't always say what they think", 3 April 2025 (arXiv 2505.05410).
5. R. Greenblatt et al., "Alignment faking in large language models", arXiv 2412.14093, December 2024.
6. J. Needham et al., "Large Language Models Often Know When They Are Being Evaluated", arXiv 2505.23836, May 2025.
7. OpenAI and Apollo Research, "Stress Testing Deliberative Alignment for Anti-Scheming Training", arXiv 2509.15541, 17 September 2025.
8. UK AI Security Institute, *Frontier AI Trends Report*, 18 December 2025.
9. S. Armstrong and S. Mindermann, "Occam's razor is insufficient to infer the preferences of irrational agents", *NeurIPS* 2018.
10. F. B. Schneider, "Enforceable security policies", *ACM Transactions on Information and System Security* 3(1), 2000.
11. R. Greenblatt, B. Shlegeris, K. Sachan, F. Roger, "AI Control: Improving Safety Despite Intentional Subversion", arXiv 2312.06942, December 2023.
12. D. Dalrymple, J. Skalse, Y. Bengio, S. Russell, M. Tegmark et al., "Towards Guaranteed Safe AI", arXiv 2405.06624, May 2024.
13. M. R. Clarkson and F. B. Schneider, "Hyperproperties", *Journal of Computer Security* 18(6), 2010; and the single-trace monitor bound of arXiv 2609.07162, September 2026.
14. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Articles 1, 9(5), 9(8), 12(1), 14, and 40; Regulation (EU) 2026/1744 (Digital Omnibus on AI), *Official Journal* of 24 July 2026; Regulation (EU) 2024/1183 of 11 April 2024 (eIDAS 2.0), Section 11, Articles 45k and 45l on qualified electronic ledgers; Commission Implementing Regulation (EU) 2025/2531 of 16 December 2025, *Official Journal* of 17 December 2025, Annex (REQ-7.5-03, REQ-7.5-04, REQ-7.5-06), ETSI EN 319 401 and ISO 23257:2022; Directive (EU) 2024/2853 on liability for defective products, 23 October 2024, Articles 4, 9, 10, and 22.
15. European Commission, Implementing Decision C(2025) 3871 of 23 June 2025 (standardisation request M/613, repealing C(2023) 3215), Articles 1, 3, and 5; Commission Implementing Decision (EU) 2026/2274 of 12 October 2026 on harmonised standards for high-risk AI systems, *Official Journal* of 14 October 2026, citing EN 18286:2026; Joint Research Centre, *Harmonised Standards for the European AI Act*, JRC139430, 24 October 2024.
16. CEN-CENELEC, *Internal Regulations Part 2: Common Rules for Standardization Work*, 2026, clause 6.5.2.1 and Annex D.2; CEN/CLC/JTC 21, prEN 18229-1 "AI trustworthiness framework, Part 1: Logging", enquiry closed 20 August 2026; *AI Standardisation Watch*, edition 16, September 2026; Legalithm, "prEN 18229-1: What the AI Act logging standard actually says", 5 August 2026.
17. IEC 61508:2010, *Functional safety of electrical/electronic/programmable electronic safety-related systems*, Parts 1 and 4; ISO 26262:2018; RTCA DO-178C (2011); UK Ministry of Defence, Def Stan 00-56; N. G. Leveson, *Engineering a Safer World*, MIT Press, 2011; ISO/IEC TR 5469:2024, *Functional safety and AI systems*.

18. R. V. Yampolskiy, "Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach", *Philosophy and Theory of Artificial Intelligence* (PT-AI 2011), Springer, 2013; D. Amodei et al., "Concrete Problems in AI Safety", arXiv 1606.06565, 21 June 2016; H. Khlaaf and S. Myers West, arXiv 2504.15088, 21 April 2025; R. McNamee, *TIME*, 5 April 2023.

19. The Nuremberg Code (1947), points 1 and 10; World Medical Association, *Declaration of Helsinki*, revision of October 2024, paragraphs 16 to 18; ICH E9, *Statistical Principles for Clinical Trials* (1998), sections 3.4, 4.5, and 4.6; US Food and Drug Administration, *Guidance for Clinical Trial Sponsors: Establishment and Operation of Clinical Trial Data Monitoring Committees* (March 2006); Federal Food, Drug, and Cosmetic Act of 1938 and Kefauver-Harris Drug Amendments, Public Law 87-781, 10 October 1962; 21 U.S.C. 355(d).

20. N. G. Leveson and C. S. Turner, "An investigation of the Therac-25 accidents", *IEEE Computer* 26(7), July 1993; US House Committee on Transportation and Infrastructure, *Final Committee Report on the Design, Development and Certification of the Boeing 737 MAX*, September 2020; FAA Airworthiness Directive 2020-24-02, 18 November 2020; ASTM F3269-21, *Standard Practice for Methods to Safely Bound Behavior of Aircraft Systems Containing Complex Functions Using Run-Time Assurance*, 2021; L. Sha, "Using simplicity to control complexity", *IEEE Software* 18(4), 2001.

21. US Environmental Protection Agency, Notice of Violation to Volkswagen AG, 18 September 2015; West Virginia University for the ICCT, in-use emissions testing of light-duty diesel vehicles, May 2014; Commission Regulations (EU) 2016/427 of 10 March 2016 and (EU) 2016/646 of 20 April 2016 (Real Driving Emissions); Regulation (EU) 2018/858 of 30 May 2018, Article 9.

22. *Bates and others v Post Office Ltd* [2019] EWHC 3408 (QB), 16 December 2019, paragraphs 971 to 975; *Hamilton and others v Post Office Ltd* [2021] EWCA Crim 577, 23 April 2021; Law Commission, *Evidence in Criminal Proceedings: Hearsay and Related Topics*, 1997, paragraph 13.13; Youth Justice and Criminal Evidence Act 1999, section 60; Post Office (Horizon System) Offences Act 2024, Royal Assent 24 May 2024; Post Office Ltd, Horizon scandal figures, updated 21 October 2026; UK Ministry of Justice, call for evidence on evidence generated by software in criminal proceedings, 21 January 2025.

23. R. P. Feynman, "Personal observations on the reliability of the Shuttle", Appendix F to the *Report of the Presidential Commission on the Space Shuttle Challenger Accident*, June 1986.

24. J. S. Niño Mojica, *The Governance Limit: Physical Foundations of Deterministic AI Safety*, Zenodo, January 2026, DOI 10.5281/zenodo.18244705; v3 March 2026, DOI 10.5281/zenodo.19200639.

25. K. Popper, *Logik der Forschung*, 1934; W. R. Ashby, *An Introduction to Cybernetics*, 1956; A. M. Turing, "On computable numbers, with an application to the Entscheidungsproblem", *Proceedings of the London Mathematical Society* 2(42), 1936.

26. H. G. Rice, "Classes of recursively enumerable sets and their decision problems", *Transactions of the American Mathematical Society* 74, 1953; M. Alfonseca et al., "Superintelligence Cannot be Contained", *Journal of Artificial Intelligence Research* 70, 2021.

27. N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014, chapter 9.

28. M. Alshiekh et al., "Safe Reinforcement Learning via Shielding", *AAAI* 2018.

29. G. Melo et al., "Machines that halt resolve the undecidability of artificial intelligence alignment", *Scientific Reports* 15, 2025.

30. R. V. Yampolskiy, "On Controllability of AI", arXiv 2008.04071, 2020.

31. OpenAI, "Planning for AGI and beyond", 24 February 2023; "Our approach to AI safety", 5 April 2023; *Raine v. OpenAI*, San Francisco Superior Court, CGC-25-628528, filed 26 August 2025, defendant's answer of 25 November 2025, coordinated as *In re ChatGPT Product Liability Cases*, JCCP 5431.
32. ISO/IEC TR 5469:2024, *Functional safety and AI systems*; IEC 61508-3:2010, Annex A, Table A.2; ISO 21448:2022 (SOTIF); UL 4600 (2020–2023); S. Shalev-Shwartz, S. Shammah, A. Shashua, "On a Formal Model of Safe and Scalable Self-driving Cars", arXiv 1708.06374, 2017.
33. US Securities and Exchange Commission, Rule 15c3-5, *Risk Management Controls for Brokers or Dealers with Market Access*, 17 CFR § 240.15c3-5 (3 November 2010); *In the Matter of Knight Capital Americas LLC*, SEC Exchange Act Release No. 70694, 16 October 2013; Commission Delegated Regulation (EU) 2017/589 (MiFID II RTS 6) and Commission Delegated Regulation (EU) 2017/574 (MiFID II RTS 25 on clock synchronisation to UTC).
34. L. Lamport, "Constructing digital signatures from a one-way function", SRI International Technical Report CSL-98, October 1979; S. Haber and W. S. Stornetta, "How to time-stamp a digital document", *Journal of Cryptology* 3, 1991; M. Bellare and B. Yee, "Forward integrity for secure audit logs", UCSD Technical Report, November 1997; B. Schneier and J. Kelsey, "Cryptographic support for secure logs on untrusted machines", *USENIX Security* 1998; KU Leuven and University of Birmingham, *Battering RAM* (DDR4 memory-bus interposition), September 2025; Georgia Institute of Technology, Purdue University, and Synkhronix, *WireTap* and *TEE.fail: Breaking Trusted Execution Environments via DDR5 Memory Bus Interposition*, October 2025; National Institute of Standards and Technology, FIPS 204, *Module-Lattice-Based Digital Signature Standard* (ML-DSA), August 2024; Charter of Fundamental Rights of the European Union, Article 49; European Convention on Human Rights, Article 7.
35. OpenAI, *Model Spec*, first published 8 May 2024 and revised since; Anthropic, *Claude's Constitution*, first published 9 May 2023 and revised since.