WHITE PAPER

Negligence by Design

The AI industry is deploying safety architectures in which no human holds authority at the moment the system acts. A published mathematical record establishes that the failure is designed, not accidental.

Stacy Gildenston · Three Primitives Research Lab · Melbourne, Australia
3Primitives.io
v1.1 · July 2026

1. The Record Is on Your Desk

When the first AI class action arrives, the defence is already written. The operator will say the failure was an unforeseeable error, an emergent property of a complex system that nobody could have predicted.

As of April 2026, that defence must answer a published, DOI-registered mathematical record proving the failure mode is structural, not emergent, and was specified before the deployments it will be raised to excuse. This paper puts that record on your desk.

Here is what the record proves, stated upfront: ***AI cannot govern AI*.** That is a *proven theorem*, not a claim and not a slogan.

No computation can generate the permission to act, because permission is not a function of any system's internal state. It lives outside the machine, in the legal framework, the consent, the named human who answers for the act. Every safety gate the industry now sells places a computation in exactly that role, and the result is consequential decisions with no author. The record proves this is *structural*, and it dates the proof *before the deployments*. Everything below is where it is operating.

Their failed architecture is not new. In Australia, it was examined at a national scale in Robodebt, where a computation was treated as the source of hundreds of thousands of decisions about human beings; the Prygodicz settlement followed, and a Royal Commission spent volumes trying to locate who decided.

This failed architecture is appearing again in the NDIS, where a tool called I-CAN will generate participant budgets under a framework in which, according to Guardian Australia and subsequent reporting, NDIA staff cannot modify the computed amount, only accept it or request a new assessment, and the Administrative Review Tribunal will no longer have authority to alter plans or funding.

If a staff member approving an I-CAN plan cannot meaningfully depart from the computed budget, the approval is not a decision. *It is a costume.*

National Disability Insurance Scheme Amendment (Securing the NDIS for Future Generations) Bill 2026 was introduced to Parliament on 14 May 2026. The industry is scaling Robodebt's architecture and calling it safety.

The author is not a lawyer, and nothing here is a legal argument. This is an evidence brief. It states what is deployed, what the published record proves, and where the dates sit. What those facts are worth in litigation is your judgement, not mine.

---

2. The Same Gate Is Being Sold as Safety

In May 2024, Dalrymple, Bengio, Russell, Tegmark, and co-authors published Towards Guaranteed Safe AI, the reference architecture for the industry's safety programme: a world model, a written safety specification, and an automated verifier that certifies each action against the specification. The verifier's pass gates the system. The framework carries the institutional weight of ARIA, Mila, Oxford, Berkeley, and MIT.

The same structure is already in production under other names. Frontier Labs gate outputs with trained classifiers. AWS, Azure, and Google sell guardrails standing in front of most enterprise AI traffic in production today. A vendor market, NVIDIA NeMo Guardrails, Guardrails AI, Lakera, and a long tail behind them sell the gate as a product. Agentic systems add tool-approval flows where the approval is itself computed: auto-approval policies, learned risk scores, allow-lists the system evaluates, with a human-shaped label on the flow.

The engineering differs. The structure does not; a computation produces a pass, and the pass is treated as the permission to act.

No product in this landscape requires a declared human authority at the gate as a structural condition of operation. And nothing here depends on which safety research method wins, because whichever one wins, all of them share the gate.

One defence will be raised early, so it should die early: a human wrote the specification; therefore, a human granted the permission. No. A rule ratified last quarter did not decide this action, in this context, at this moment.

> *A standing rule is a constraint on action, not an authorisation of this action, at this time, affecting this person. The two are legally and structurally distinct.*

Treating the rule as permission is not an oversight. It is the precise failure that the record names, with paperwork.

---

3. What the Record Proves

The core result needs no symbols. Decisions entail responsibility. Responsibility requires agency. Protocols, whether a spreadsheet formula or a frontier AI model, lack agency and cannot bear responsibility. So when a protocol is treated as the source of a decision, agency is displaced from any human actor, and authority is exercised that no one holds.

That is ***Ghost Authority,*** and it is not a metaphor. It is a *proven condition with a published definition and a test*:

> *"At the point where consequences attach, if no human explicitly declares 'I decide this,' the decision lacks legitimate authority."* [FR05, Corollary, the Ownership Test]

Apply that test to any product named in Section 2. When the gate passes an action, and someone is harmed, ask one question: which human, by name, declared 'I decide this' for that action? For every product above, that name does not exist, and discovery will confirm it.

Underneath the plain language sits a formal impossibility result, published as FR02, AI Cannot Govern AI, for the day an expert witness needs it:

> *There exists no total computable function $h : S \times A \to \{0, 1\}$ capable of autonomously resolving authorisation without an exogenous declaration.* [FR02, Lemma 1, v2.3, April 2026]

Translated: *AI cannot govern AI. No computation can generate permission from inside the system, not because permission is hard to compute, but because it is not a function of the system's internal state at all. Permission depends on the context that lives outside the machine: legal frameworks, consent, and the named human who answers for the act. This is not a stipulation. This is a mathematical fact.*

The dependencies are checkable: the legal framework governing a medical decision, a financial transaction, or an infrastructure deployment exists outside the model's parameters, and no internal computation can reach it. Compliance with a specification is computable. Grant the industry everything: assume its verifiers compute compliance perfectly every time. Nothing changes.

Permission is a different object. The architecture takes the computable one (compliance) and installs it where the other (permission from a human to act) must stand. That substitution is the defect; it is present at design time, and no amount of engineering on the verifier removes it because the verifier was never the problem. Its job description was.

---

4. Two Defences, Both on Notice

Defence one: unforeseeable error. It works by pointing at the model's internals, where opacity is real, and claiming that what happens in there could not have been known. The record breaks it before the internals are ever reached, because the failure this paper concerns is not in the model. It is in the gate, and the gate was designed.

The record proves, with priority dates, that placing a computation in the authorisation role guarantees consequential actions with no attributable author. Not emergent. Structural, specified in advance, and published under an open licence so that every deploying party could read it. The question is no longer whether the operator could have known. *The question for discovery is why the operator deployed an architecture whose failure was a published theorem.*

Defence two: nothing better was possible; computed gating is simply the state of the art. The same corpus answers this because it specifies not just a critique but the structural alternative.

Formal Record 1 (FR01) defines three irreducible primitives in governance: the declaration of purpose, the named authority holder, and the operational constraints.

These are not design preferences. They are the minimum structural requirements for any governed action, in any domain, from medicine to criminal law to infrastructure.

The formal record demonstrates this across ten governance domains, especially law, at https://3primitives.io/domains. When any one of the three is absent, governance is incomplete by definition, not by opinion.

From these primitives, **FR01** specifies what must actually cross the gate: a declaration, δ, naming the bounded purpose, the live human authority, and the constraints, issued before execution and immutable for that event, with every gated action leaving a record:

> *No execution may occur without this immutable audit record.* [FR01, Execution Record R(x), v2.2, December 2025]

No declaration, no action. The gate fails closed. We built these proofs into a working, tested, implemented system, the **Upstream Safety System™**, documented at https://3primitives.io/uss. Our architecture, built directly on the proofs, was specified in the same public corpus and published before deployment at scale. Whether it was available to any given defendant is a question for discovery, but the record that it was published and accessible is not.

The regulatory axis points in the same direction. The EU AI Act's Article 14 requires high-risk systems to be designed for effective human oversight. It is persuasive, not binding, outside the EU, but it marks the international regulatory consensus: *oversight performed by a computation is not human oversight with extra steps. It is the absence of human oversight, benchmarked.*

5. Provenance

The corpus is thirteen formal records, FR01 through FR13, published under CC BY 4.0 with DOI registration, authored by Stacy Gildenston and Pyrate Ruby Passell at Three Primitives Research Lab, Melbourne. The canonical logic sequence was closed in December 2025. The impossibility proof and the Ghost Authority Lemma were published in their current form in April 2026, and the corpus was completed in May 2026. Every version is dated. The main public-facing source is https://3primitives.io.

The central claims were subjected to structured adversarial testing across multiple frontier AI systems, with a human adjudicator resolving every dispute against the published records, and no break was found on the three core claims.

One session produced something more: a documented, timestamped demonstration of the failure mode itself. Four frontier models, each handed the corpus in a new chat and asked to apply it, instead treated our lead model's (hallucinated, mathematically incorrect) output as authoritative and propagated that fabricated equivalence for twenty-four hours until a human broke the chain of mirrors against the records they had all ignored and all had fresh access to. The AI systems demonstrated the theorem that AI cannot govern AI by failing to read it. This is evidence of the failure mode in practice, not a proof, but the proof does not require it. The formal records stand on their own terms.

6. The Record

When an automated gate passes an action and harm follows, the operator will reach for an unforeseeable error. *The record proving the failure is structural, not emergent, is already published, already dated, and already distributed.*

This paper establishes the structural defect. Its companion, ***The Second Line of Defence****, maps the four categories of AI safety products on the market today and measures the foundational assumption they all share. The result is zero.*

Three Primitives Research Lab · Melbourne, Australia
3Primitives.io

Formal Records cited: FR01, Canonical Logic Sequence (v2.2) · FR02, AI Cannot Govern AI (v2.3) · FR05, Ghost Authority Lemma (v2.3). Formal Records | 3primitives.io/formal_records