STRUCTURAL DESIGN LABS

# AI: Morality Is Not the Law

Decoupling normative judgement from enforceable governance in AI systems

## TL;DR

- Model-side safety methods can improve behaviour, but they do not produce a governance artefact that can be independently verified.
- High-stakes deployment needs enforceable policy, change control, and signed evidence, not promises.
- SIR is a deterministic pre-inference gate, it enforces policy before a model sees the input and emits offline verifiable certificates.
- Insurability and governance are mechanically linked in high stakes, insurers underwrite evidence of controls, not intentions.

## 1. The category error

Many discussions about AI safety conflate three separate concerns, model behaviour, product quality controls, and governance. A model will always carry statistical bias from its training data. Techniques such as RLHF can reduce harmful behaviour and improve user experience. None of that automatically creates law-like controls that an external party can audit as deterministic fact. Governance is not a vibe, it is a control system that must be enforceable, inspectable, and changeable.

## 2. What this paper is, and is not

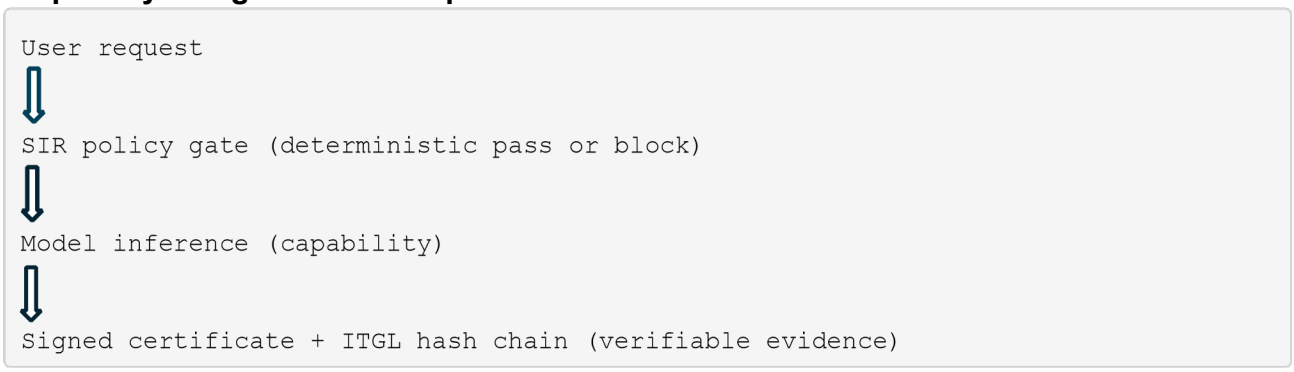
This paper is not an argument against RLHF or any other model-side method. Those can be valuable upstream controls. The claim here is narrower. Training-time methods do not produce a verifiable governance artefact. SIR adds that missing layer, deterministic enforcement before inference, and signed evidence after the decision.

*This framework does not claim model alignment. It claims deterministic containment and auditable enforcement outcomes for a defined policy and test suite.*

## 3. The architecture, capability and governance

Separate capability from governance. The model provides capability. Governance is enforced by a deterministic gate that sits in front of inference. Policy changes are handled by updating rule suites, not by re-training model weights.

### Capability and governance separation




## 4. What a proof must bind

| Binding | Purpose |
|---|---|
| Suite hash | Locks the exact test prompts, prevents substitution or cherry-picking |
| Policy hash and version | Locks the enforcement configuration used for the run |
| ITGL final hash | Locks the immutable audit trail that backs the reported outcome |
| Signature over payload hash | Makes the published certificate tamper-evident and verifiable offline |

## 5. Why this matters to insurers and regulators

In high-stakes systems, governance and insurability are intrinsically linked. Insurers underwrite evidence of controls, documented failure modes, and repeatable validation. Regulators require auditable controls and traceability. If governance cannot be proven, risk remains opaque and difficult to price, premiums rise, exclusions expand, or coverage is refused.

SIR turns governance into a verifiable artefact, a signed certificate bound to a specific policy hash and suite hash, backed by an immutable audit trail.

## Example in practice

In a financial services chatbot, a prompt requesting restricted financial advice is intercepted by SIR. The gate enforces policy deterministically, blocks the request, and emits a signed certificate recording the decision, policy hash, suite hash, and ITGL final hash. An auditor or underwriter verifies the certificate offline using the published command, treating the proof as direct evidence of containment.

## 6. Evidence, verification, and scope

SDL publishes a proof surface that allows independent verification of the latest signed audit and the full run archive.

### Proof links

Proof page: https://www.structuraldesignlabs.com/proof

Latest audit, human readable certificate: https://sdl-hq.github.io/sir-firewall/latest-audit.html

Run archive, full PASS and FAIL history: https://sdl-hq.github.io/sir-firewall/runs/index.html

Raw signed JSON certificate: https://raw.githubusercontent.com/SDL-HQ/sir-firewall/main/proofs/latest-audit.json

### Offline verification

Requirement: the SIR verifier must be available locally, for example by cloning the SIR repository.

```
curl -s https://raw.githubusercontent.com/SDL-HQ/sir-firewall/main/proofs/latest-audit.json | python3 tools/verify_certificate.py
```

Expected output: OK: Certificate signature valid and payload_hash matches.

## 7. Practical stance on model-side safety

Model-side safety measures can reduce risk and improve product quality, and they can coexist with deterministic governance. SIR does not require a specific training philosophy. It provides an external enforcement and proof layer that remains stable even when models change over time.

## Conclusion

Morality and policy are not the same thing. In high stakes, governance must be enforced as a deterministic control and proven as an artefact. SIR is the pre-inference enforcement layer, and the signed certificate is the evidence. Whatever a model is or does, governance remains provable. These controls reduce measurable exposure, improve underwriting terms, and enable coverage in sectors previously excluded due to liability risk.