The Calvin Convention: *A Contractual Framework for AI Safety and Accountability*

1. Introduction: The Governance Inversion

Modern AI governance is dangerously inverted. It focuses on retrospective audits, elaborate explanations, and compliance rituals that occur *after* harm has already been done. This “govern-after-the-fact” model is a structural failure, treating accountability as a post-mortem exercise rather than an architectural prerequisite. The strategic challenge of our time is to invert this model, advancing a constitutional argument for embedding refusal into system design before a single decision is made.Effective safety constraints, as science fiction has long understood, must be pre-action, hierarchical, and non-negotiable at runtime. An AI system, like any high-risk industrial machine, should not act and then explain; it must possess the built-in capacity to refuse first. Instead, current governance frameworks often amount to sophisticated forms of “barn door maintenance”—offering detailed analysis of how the horse bolted long after it has left the stable. Once action precedes control, governance becomes retrospective by definition. You are no longer preventing harm; you are accounting for it.The purpose of this white paper is to diagnose the structural failures that enable this inversion—specifically the normalization of **Opacity** and the creation of the **Liability Sponge** —and to present the Calvin Convention as a robust, contract-ready framework for embedding refusal and control into high-risk systems before they are ever deployed. This document seeks to industrialize a line of argument: that we must shift our demands from post-hoc *explanation* to pre-deployment *power* .This paper will first analyze the core architectural flaws that define the current landscape of AI governance.

2. The Architecture of Abdication: Two Failures in Modern AI Governance

The failures of high-risk AI systems are rarely the product of conventional bugs or malicious intent. They are, more often, the predictable and logical outcomes of a flawed architectural philosophy. This philosophy normalizes inscrutability

and misuses human oversight as a mechanism for blame absorption rather than genuine control. This section will dissect two of the most critical flaws in this architecture: the crisis of opacity and the institutionalization of the human operator as a "liability sponge."

2.1. The Crisis of Opacity: When "Proprietary" Means Unaccountable

Arthur C. Clarke's Third Law, "Any sufficiently advanced technology is indistinguishable from magic," is often quoted as a celebration of innovation. It is more accurately a description of a governance failure mode. When a system's reasoning becomes so opaque that it resembles magic, we experience a form of "epistemic surrender." We stop arguing with the system. We stop asking *how* it decided. We start asking *what* it decided, and then we comply. That shift is where governance dies.This collapse is often formalized through the "vendor defense," where claims of "proprietary IP" and "commercially sensitive" information shield algorithmic reasoning from scrutiny. Opacity becomes contractual. Procurement, intended to secure services, is transformed into a mechanism to structurally embed unaccountability. The algorithm's authority is laundered through the procurement process. An institution outsources its reasoning to a vendor, the vendor claims commercial confidentiality, and the individual faces a decision that cannot be explained by the body that made it and cannot be examined because examination would harm the vendor's competitive position.This principle is acutely visible in several high-stakes domains:

- **Public Benefits:** Government agencies license automated eligibility and fraud-detection tools whose internal logic they cannot explain. When a citizen's benefits are denied or clawed back, they are left facing a decision they cannot appeal on substantive grounds. This is due process as ritual. The forms exist. The steps exist. The substance is absent.
- **Credit Scoring:** When a loan application is declined, the legally required "adverse action notice" provides the *shape* of reasoning—citing factors like "insufficient credit history"—without granting the right of interrogation. The applicant cannot see the specific weights, the counterfactuals, or the historical data proxies that led to the decision, laundering historical biases into a mathematical judgment they cannot see or challenge.

2.2. The Liability Sponge: "Human in the Loop" as a Blame Absorption Mechanism

The "Liability Sponge" is a design pattern where a human operator is placed in a high-velocity decision loop not to exercise meaningful control, but to absorb legal and institutional liability for systemic failures. When an algorithmic process that operates at silicon speed is "overseen" by a human operating at biological speed, the human is not a fail-safe; they are a scapegoat.This model is built on "impossible math." An operator asked to validate 1,247 safety flags in a four-hour window—one every 11.5 seconds—is not performing oversight. They are participating in a ceremony of compliance. Performance metrics that evaluate operators on throughput create a powerful incentive to defer to the system's recommendations, as scrutiny is penalized. In this context, "meaningful human review" becomes a definitional impossibility.The public-sector caseworker provides a stark example of the "override trap" this creates:

- If the caseworker **overrides** the system's recommendation to deny a claim, and that claim later proves fraudulent, the override is documented. The caseworker made an explicit judgment call against the machine, and it was wrong. Accountability is personal and direct.
- If the caseworker **defers** to the system's flawed recommendation, the error is systemic. Nobody made a judgment call. The system worked as designed. Accountability diffuses into the architecture, the vendor, and the institution.Over time, deference becomes the only rational professional choice. This stands in sharp contrast to the industrial safety principle of "Stop Work Authority," where any worker, regardless of rank, has the absolute authority to halt a process they deem unsafe, and the system is designed to default to a safe state. The Liability Sponge model inverts this, making the human the default point of failure.The systemic disasters detailed in the next section were not mere failures *of* opaque systems; they were failures *caused by* the architectural abdication detailed here. Opacity and the Liability Sponge were the necessary preconditions for the harm that followed.

3. Case Studies in Systemic Failure

The architectural failures of opacity and liability absorption are not theoretical. They have been implemented at scale, translating directly into profound and preventable human suffering. This section grounds the analysis in documented domains where these flaws have led to systemic disaster.

3.1. Public Eligibility Systems: The Michigan MiDAS and Australian Robo-Debt Disasters

In both Michigan and Australia, governments deployed automated systems to detect fraud and manage public benefits. The results were catastrophic, demonstrating a shared architectural flaw where opacity enabled flawed logic to operate unchecked at a massive scale.| System | Flawed Mechanism | Documented Consequence || -—— | -—— | -—— || **Michigan MiDAS** | Automated data-matching interpreted all inconsistencies as fraud, with no consideration for alternative explanations or routing to human review. It automatically imposed penalties on the accused. | Between 2013 and 2015, the system wrongly accused over 40,000 people of fraud, with a **93% false positive rate** . This led to garnished wages, seized tax refunds, bankruptcies, and lost homes. || **Australian "Robo-Debt"** | Crude data-matching averaged annual tax data against fortnightly welfare payments, flagging discrepancies as overpayments. The system then unlawfully reversed the burden of proof onto citizens. | Between 2016 and 2019, hundreds of thousands of unlawful debt notices were issued. The scheme was eventually dismantled, leading to over a billion dollars in refunds and a Royal Commission that documented severe stress, shame, and suicides. |

The common cause of both failures was architectural. Opacity was the mechanism that allowed flawed logic—obvious upon inspection—to operate at scale. The appeals processes were circular; the system's outputs were treated as presumptively correct evidence of the very fraud they were supposed to detect. Citizens were forced to argue against an invisible accuser whose reasoning they could not see.

3.2. Economic Gatekeeping: Prediction as Prescription

In domains like credit and insurance, opaque models function as unaccountable gatekeepers to economic life. While defended as actuarially "accurate," this accuracy launders historical biases and can create self-fulfilling prophecies. The prediction of risk can become a prescription for it.For example, a model might predict that an individual is a high-risk candidate for health insurance. The resulting high premiums may lead the individual to defer care or skip coverage. This deferred care, in turn, can lead to worse health outcomes, making the high-risk prediction a reality. The model was right: they were expensive. The model helped make them that way.The actuarial defense —"we're just measuring risk"—positions this as discovery, not intervention. But

“accurate pricing” is a policy choice, not a natural law. We could choose to pool risks more broadly, to prioritize access over precision. The actuarial framing makes these choices invisible. It presents hyper-individualized pricing as the natural state and solidarity-based pooling as distortion. The politics vanish into math. Because the model’s reasoning is proprietary, the affected person cannot trace this feedback loop. They experience the price or the denial not as a constructed outcome, but as an immutable fact about themselves.The failures documented above are not bugs to be patched. They are features of a governance model that must be replaced.

4. A New Foundation for Control: The Calvin Convention

The solution begins with a simple insight, best articulated by Isaac Asimov’s fictional robopsychologist, Susan Calvin: effective governance of complex systems requires contracting for *power* , not demanding *explanation* . If a car has reliable brakes that are under sovereign control, the driver does not need to understand the intricacies of fuel injection. They just need to know that when they press the pedal, the machine stops. We have been asking vendors to explain their systems’ reasoning when we should have been demanding the contractual power to control their behavior.The Calvin Convention is a “Bill of Rights for the Human in the Loop”—a set of six contract-ready mechanisms designed to restore sovereign control to operators and institutions *before* a system is deployed. These are not technical tweaks; they are non-negotiable contractual clauses that re-architect the relationship between human authority and algorithmic process.

4.1. Pre-Deployment Rule Sovereignty

Problem: The model decides based on statistical likelihood, potentially violating core institutional principles or harming vulnerable individuals in predictable ways.**Fix:** The institution defines non-negotiable rules that override the model’s output every time, without exception. These are jurisdictional boundaries the model cannot cross, regardless of its confidence score. For example, a contractual clause might state: *“Any grievance mentioning ‘burial site,’ ‘water contamination,’ or ‘intimidation’ bypasses automation entirely and routes to a senior human.”*

4.2. Human-Defined Uncertainty

Problem: The model declares its own confidence (“I am 87% sure”), forcing humans to adapt to its internal sense of certainty.**Fix:** The human institution

defines the risk appetite and sets the thresholds for action. This includes defining acceptable false-negative risk and the acceptable volume of cases per reviewer per day. The model must adapt to the institution's definition of safety. If the system cannot meet these human-defined safety parameters, it must halt.

4.3. Default to Hold

Problem: To maintain high throughput, systems are often designed to default to "Process" or "Approve," requiring active human energy to prevent potential harm.**Fix:** The system's default state is "Hold." If a rule is triggered or an uncertainty threshold is breached, the system stops. It does not flag and proceed; it pauses the action. Support payments are maintained, an eviction is paused. The system must be designed to require active energy to inflict harm, not active energy to prevent it.

4.4. Evidence Access as a Right

Problem: The "vendor defense" of "proprietary IP" is used to deny human reviewers the information they need to validate an algorithmic decision.**Fix:** If a human is asked to validate a decision, they must have a contractual right to see the raw inputs and the transformation steps that led to the output. "No access due to IP" is defined as a breach of the accountability chain. If IP prevents accountability, the system is unfit for high-risk deployment. Full stop.

4.5. Bulk Control

Problem: Systems often force operators to override flawed decisions one by one, an exhausting and impractical task designed to wear down resistance and ensure compliance through friction.**Fix:** The institution must have the contractual right to exercise "Stop Work Authority" at scale. If an operator identifies a systemic drift or pattern of error, they must have the power to instantly suspend all decisions within that cohort, not just appeal one case at a time. This transforms individual resistance into collective agency.

4.6. Pre-Registered Failure Modes

Problem: Vendors often feign surprise at predictable failures, framing them as unforeseeable "edge cases" to deflect responsibility.**Fix:** Before deployment, the vendor and the institution must jointly document the model's known blind spots and limitations. For instance: *"This model struggles with dialect X,"* or *"This classifier has not been validated on population Y."* These documented warnings must be attached to every relevant output in the audit trail, ensuring

that when a predictable failure occurs, it is logged as a known system limitation, not a human error.

5. Conclusion: From Governance Theater to Contractual Reality

This paper has advanced a constitutional argument for a critical shift in how we govern high-risk AI systems: away from post-hoc, ceremonial oversight and toward pre-action, architectural constraints. The failures we see today are not inevitable consequences of advanced technology; they are the direct results of an inverted governance model that prioritizes throughput and plausible deniability over genuine human control and accountability. By focusing on explanations after the fact, we engage in governance theater while systems with no brakes operate at scale.A black box with a kill switch is governable. A transparent box with no brakes is lethal.This presents a clear call to action for policymakers, procurement officers, and technology leaders. Safety and accountability are not technical features to be requested; they are fundamental contractual rights to be demanded. The Calvin Convention provides the language and the logic to do so. It shifts the burden from asking “Can you explain this?” to demanding “Can you prove this is controllable?”The technology is ready for this framework. The algorithms can be constrained, the defaults can be changed, and the controls can be built. The remaining challenge is not technical; it is a matter of exercising the institutional courage to make this framework a mandatory condition of deployment for any AI system with authority over human welfare.