Legitimacy as Emergent Gain

The Dual-Channel Coupling of Trust in Governance Architecture

Paper XIII in the Governance as Engineering series

Models legitimacy as an emergent coupling state — a gain parameter generated by the interaction between architecture and the governed. Derives the legitimacy trap, the borrowed-vs-built distinction, and hysteresis dynamics. With simulation and empirical illustrations. Paper XIII bridges the series' primitives to the outcomes they produce.

Björn Kenneth Holmström

June 2026

`https://bjornkennethholmstrom.org/working-papers/legitimacy-as-emergent-gain`

Executive Summary

Two governments possess identical formal institutions—professional civil services, independent judiciaries, free presses, democratically elected legislatures. One delivers on its promises; the other does not. The difference is not in the architecture. It is in a parameter that the architecture does not control: the willingness of the governed to comply with directives and report honestly. That parameter is legitimacy. And it is not a political nicety. It is a gain parameter that multiplies the effectiveness of every architectural choice.

This paper identifies legitimacy as the Governance as Engineering series' first **endogenous coupling state**—a variable that is not chosen by the designer but that emerges from the interaction between governance architecture and the governed population, and that simultaneously modulates both actuation and observation. Unlike latency, signal fidelity, or boundary selection (the architectural primitives of Papers I–XII), the designer cannot set L directly. The designer can only choose the architecture that, over time, generates or erodes it. And once L exists, it feeds back on the architecture's own effectiveness with multiplicative force.

The formal framework models legitimacy as a state-dependent scheduling parameter $L(t) \in [0,1]$ in a linear parameter-varying (LPV) control system. Actuation effectiveness becomes $\mathbf{B}\_eff = L \cdot \mathbf{B}$; observation noise covariance becomes $\mathbf{V} = \mathbf{V}_0/L$. The two channels that the series has treated separately are coupled through a single variable that evolves under its own dynamics—responding to delivery performance, transparency, and betrayal. When L is high, the controller enjoys full authority and clear sensing. When L collapses, the same institutional architecture becomes unsteerable and blind.

The paper introduces a critical distinction between **built legitimacy** and **borrowed legitimacy**. Built legitimacy arises from consistent, transparent delivery over extended periods. It is slow to accumulate, resilient to individual failures, and operates in a stable dynamic regime. Borrowed legitimacy arises from narrative, charisma, or temporary success. It can be acquired quickly but is structurally brittle: a single revelation of deception can trigger a catastrophic collapse, because the betrayal sensitivity γ is far larger than the delivery sensitivity α. The distinction explains why some regimes appear stable for decades and then disintegrate rapidly—they have been operating on borrowed legitimacy, and the hidden architectural debt has been called.

Three structural failure modes follow from the dynamics. The **performance–legitimacy spiral** is a self-reinforcing collapse in which a delivery failure reduces L, which weakens actuation and degrades observation, which produces further delivery failures. The **transparency trap** is the regime's attempt to arrest the spiral by suppressing or manipulating the observation channel—a strategy that maintains apparent L in the short term while accumulating a hidden discrepancy whose eventual revelation triggers a catastrophic betrayal penalty far larger than the decline that honesty would have produced. **Legitimacy hysteresis** is the persistent gap between the speed of decline and the speed of recovery: trust is destroyed quickly but rebuilt

slowly, meaning a controller that has experienced a legitimacy collapse must sustain significantly better performance for significantly longer to return to its previous L level than the path that caused the decline required.

A simulation of the legitimacy-coupled control loop demonstrates these dynamics under controlled conditions. The trap, the transparency collapse, and the hysteresis gap emerge reliably from the model even when all other architectural primitives are held at ideal values—confirming that the failure modes are not consequences of institutional dysfunction but structural properties of any system in which legitimacy couples actuation and observation.

Empirical illustrations span from the Nordic high-trust equilibrium (a stable built-legitimacy configuration where L multiplies the effectiveness of already sound institutions) through the Greek sovereign debt crisis (a canonical transparency-trap collapse of borrowed fiscal legitimacy) to South Africa's Truth and Reconciliation Commission (a deliberate legitimacy-rebuilding intervention through transparency about past deception) and China's calibration deficit (a split-legitimacy architecture with high actuation-legitimacy for monitored targets and low observation-legitimacy for sensitive dimensions).

Six design principles follow. **Transparency by default** maintains the observation channel on which long-term L depends. **Delivery–reality matching** prevents the promise-delivery gaps that erode L fastest. **Legitimacy sensors**—independent, diversified, high-frequency measurements of compliance and reporting integrity—provide the controller with the information needed to adapt its strategy to its own L level. **Circuit-breaker mechanisms** automatically constrain the controller's authority when L falls below a critical threshold, preventing further action from destroying what remains. **Credible commitment** mechanisms—constitutional entrenchment, fixed-term appointments, pre-committed transparency standards—enable the extended recovery trajectories that the hysteresis asymmetry requires. And the **legitimacy substrate requirement** extends these principles to the global boundary institutions that Paper XII argues are structurally necessary, because a planetary institution with zero legitimacy has zero effective actuation regardless of how perfectly its boundary matches the coupling structure of the governed domain.

The paper is the thirteenth in the series and completes the foundational arc of Cycle Two. Paper XI asked whether the controller can implement its intent. Paper XII asked whether the controller is acting on the right system. This paper asks whether the system will cooperate with the controller's actions. The sequence—actuation, boundary, legitimacy—is the progression from architecture through context to the emergent state that determines whether either functions.

The central contribution is not the claim that legitimacy matters. That has been said many times before, in many vocabularies. It is the claim that legitimacy has a specific, analysable structure—an endogenous state variable with defined dynamics, measurable parameters, and predictable failure modes—and that governance architectures can be designed to observe, protect, rebuild, and never borrow against it at a cost they cannot repay. The paper provides the formal grammar for that structure and the design vocabulary for its implications. It treats legitimacy not as a diffuse political atmosphere but as the gain parameter on which

every other architectural choice depends. A controller that ignores its own L will eventually discover that it has built an elegant machine that no one is willing to operate. A controller that takes L seriously as a design objective will find that it is the foundation on which everything else rests.

Part I — The Legitimacy Gap

1.1 Two Governments, Identical Architecture

Consider two governments. Each possesses a professional civil service, an independent judiciary, a free press, and a democratically elected legislature. Each has access to the same policy instruments, the same fiscal capacity, and the same reservoir of technical expertise. By the standards of institutional quality—the metrics that dominate comparative governance assessment—they are near-twins.

The first government announces a major infrastructure programme. Parliament authorises the funds. The ministry issues the regulations. The contractors mobilise. Five years later, the bridges are built, the railways are operating, and the cost overruns are within the normal range for large projects. The second government announces an identical programme. Parliament authorises the funds. The ministry issues the regulations. The contractors mobilise. Five years later, half the bridges are unfinished, the railway budget has been consumed by litigation and renegotiation, and the cost overruns have doubled the original estimate.

The difference is not in the formal architecture. It is in what happens between the announcement and the asphalt. In the first country, when the ministry issues a regulation, the affected parties generally comply. When the tax authority requests payment, the payment arrives. When the statistical agency surveys economic activity, the responses are reasonably accurate. In the second country, when the ministry issues a regulation, the affected parties negotiate, delay, and litigate. When the tax authority requests payment, a significant fraction of the payment is diverted, disputed, or simply never made. When the statistical agency surveys economic activity, the responses are shaped by what the respondent wants the government to believe.

These behavioural differences are not external to the governance architecture. They are parameters inside the control loop. A directive that is not followed is an actuator that did not move. A report that is systematically distorted is a sensor that is out of calibration. The formal architecture—the laws, the institutions, the procedures—is identical. The effective architecture, the one that actually determines outcomes, is radically different. The variable that accounts for the difference is legitimacy: the willingness of the governed to comply with directives and to report honestly to the institutions that govern them.

Legitimacy is not a normative judgment about whether the government deserves to govern. It is an operational parameter: the probability that a control signal will be executed, and the probability that a measurement will reflect reality. When that parameter is high, the same institutional architecture produces dramatically better outcomes than when it is low. The difference is not marginal. A government operating with a compliance probability of 0.9 and a reporting honesty probability of 0.9 is roughly twice as effective as one operating with probabilities of 0.5—not because it is twice as competent, but because every intervention is multiplied by the parameter that determines whether it reaches the system it is meant to affect.

This paper is about that parameter. It treats legitimacy not as a political science concept or a moral ideal, but as a structural property of the control loop that couples a governance system to the population it governs. And it argues that a controller that ignores its own legitimacy level is operating open-loop on the very variable that determines whether its loop closes.

1.2 The Pattern Across the Series

The Governance as Engineering series has examined fifteen national governance systems, six organisational architectures, and a growing set of structural primitives. Legitimacy, or its absence, appears in many of the failure modes the series has diagnosed. But it has never been named as a distinct variable in its own right. This section traces its shadow presence across the cases, to establish that the pattern is already there, waiting to be formalised.

France's reform-explosion-retreat cycle. The France country study diagnoses a pattern in which technically sound reforms are announced by the centre, met with sustained protest in the street, and then withdrawn or diluted. The formal architecture—the Jacobin state's capacity to design and legislate—is formidable. The effective actuation collapses because a significant fraction of the population refuses to comply with the reform, and the state lacks the legitimacy to enforce it. The controller issues a directive; the actuator does not move. The France case is a legitimacy failure, not a technical design failure.

Brazil's capture equilibrium. The Brazil study documents a system in which programmes are designed with considerable sophistication—Bolsa Família, PIX, the electoral system—but their implementation is systematically captured, diluted, or diverted by an extractive political economy. The state possesses islands of high institutional capacity. What it lacks is the broad-based legitimacy that would enable those islands to expand into an archipelago. The *Centrão* extracts not because the architecture is absent but because the legitimacy that would make extraction politically costly is insufficient to constrain it.

China's calibration deficit. The China study identifies a system in which local officials systematically distort the information they report upward, because the promotion tournament rewards favourable metrics and punishes unfavourable ones. The centre receives a systematically optimistic picture of conditions on the ground. This is an observation-legitimacy failure: the governed (in this case, local officials acting as sensors for the centre) do not report honestly, because they do not trust that honest reporting will be rewarded. The observation channel is degraded not by architectural design but by the behaviour of the agents within it, behaviour driven by the absence of a specific kind of trust.

Russia's legibility deficit. The Russia study documents the extreme case: a power vertical that has systematically destroyed independent observation channels, making accurate information dangerous to the informant. The result is a controller that is structurally blind—not because its sensors are technically incapable, but because the governed have learned that providing accurate information is punished. Legitimacy has collapsed to the point where the observation channel is actively adversarial.

The Nordic high-trust equilibrium. At the other end of the spectrum, the Nordic cases—Finland, Sweden, Denmark—exhibit a governance architecture in which compliance is high, reporting is honest, and the state can operate with high gain across multiple policy domains. The architecture supports ambitious welfare states not because Nordic institutions are uniquely well-designed in a technical sense, but because the legitimacy parameter multiplies the effectiveness of every design choice. A Swedish agency and a Brazilian agency may have identical formal mandates. The Swedish one operates with an effective actuation matrix near unity; the Brazilian one, substantially below it.

In each of these cases, the variable that distinguishes outcomes is not captured by the architectural primitives the series has already formalised. Latency, signal fidelity, representation depth, boundary selection—these are properties of the institutional design. Legitimacy is a property of the relationship between the design and the population it governs. It is the parameter that determines whether the architecture actually functions as designed.

1.3 The Structural Claim

The central claim of this paper is that legitimacy is a structural variable of the governance control loop—not an external context, not a political nicety, but a parameter that sits inside the feedback equations and determines whether they converge or diverge.

Legitimacy is not a primitive in the sense that latency or boundary selection is a primitive. The designer cannot choose L directly. The designer can only choose the architecture that, over time, generates or erodes L. But once L exists, it operates on the architecture with the same structural force as any primitive. A system with latency $\tau = 2$, signal fidelity $\sigma = 0.1$, and a perfectly matched boundary will still fail if $L = 0.1$—because the effective actuation matrix is reduced to 10% and the observation noise is amplified tenfold. The architectural primitives are necessary. They are not sufficient. The parameter that multiplies them must be present, and it can be absent even when the architecture is flawless.

The paper formalises this claim through a specific modelling choice: legitimacy is treated as a state-dependent gain parameter that couples the actuation and observation channels. When L is high, the effective actuation matrix $\mathbf{B}_{\text{eff}} = L \cdot \mathbf{B}$ is near the designed $\mathbf{B}$, and the observation noise covariance is near the designed minimum. When L falls, actuation weakens and observation noise increases simultaneously. The two channels, which Papers I through XII treated as separable, are coupled through a single parameter that the controller does not directly control.

This coupling is the paper's central formal contribution. Earlier papers diagnosed failures in observation (Papers III, IV, VI, X) and failures in actuation (Papers I, XI). This paper argues that there exists a variable that affects both simultaneously, and that a controller experiencing that variable's collapse will face a compound failure that is more severe than the sum of its individual channel degradations. The legitimacy trap—the self-reinforcing spiral in which falling L degrades both channels, which degrades outcomes, which further reduces L—is the signature failure mode that this coupling generates.

The paper further distinguishes between two modes of legitimacy: *built* and *borrowed*. Built legitimacy arises from consistent, transparent delivery over time. It is slow to accumulate, but it exhibits low sensitivity to individual failures and high persistence. Borrowed legitimacy arises from narrative, charisma, enemy construction, or temporary success. It can be acquired quickly, but it is highly sensitive to revealed deception and collapses catastrophically when the gap between the narrative and observed reality becomes undeniable. Formally, built legitimacy has low betrayal sensitivity (γ) and high exogenous persistence (δ); borrowed legitimacy has high γ and low δ. This distinction explains why some regimes appear stable for extended periods and then collapse suddenly: they have been operating on borrowed legitimacy, and the accumulated discrepancy between the borrowed narrative and the underlying reality has finally breached the threshold of deniability.

The structural claim is testable. If legitimacy operates as a coupling gain, then systems with higher measured L should exhibit both higher policy implementation rates and lower reporting bias, controlling for formal institutional quality. The hysteresis prediction implies that legitimacy, once lost, should recover more slowly than it declined. The borrowed-legitimacy prediction implies that regimes with high narrative control but low transparency should exhibit more catastrophic legitimacy collapses than regimes with lower narrative control but higher transparency. These are empirical claims, and they are specified with sufficient precision to be falsifiable.

1.4 Positioning Against Normative Theories

A clarification is necessary before the formal development begins, because the word "legitimacy" carries a weight of normative theory that this paper deliberately sets aside.

Political philosophy has debated the grounds of legitimate authority for centuries. Consent theory, democratic theory, justice-based accounts, and procedural accounts each offer criteria for determining when a government *deserves* to be obeyed. These are important debates, and this paper takes no position in them.

The paper's claim is narrower and structurally distinct. It is not that legitimate governments are morally better. It is that legitimacy—understood as the empirical probability of compliance and honest reporting—has structural consequences for governance performance that operate regardless of the moral basis on which that legitimacy rests. A regime that is widely regarded as illegitimate by normative standards may still enjoy high compliance if it governs through fear, clientelism, or inertia. A regime that is impeccably democratic by normative standards may face low compliance if it has consistently failed to deliver on its promises. The paper's L parameter measures effective compliance, not moral desert.

This is not a cynical move. It is an analytical one. By separating the structural parameter from its normative foundations, the paper can ask questions that a purely normative framework cannot: What level of L is required for a given governance architecture to remain stable? How does L evolve in response to delivery

performance and transparency? What design features make an architecture robust to fluctuations in L? These are engineering questions, and they admit engineering answers that do not depend on resolving millennia of philosophical debate.

The paper also does not claim that high L is always desirable from a societal perspective. A highly legitimate regime may efficiently pursue destructive policies—mobilising for aggressive war, extracting resources unsustainably, enforcing oppressive social norms. The paper's claim is only that high L makes the control loop *efficient*, not that it makes it *good*. The normative question of what goals the loop should pursue belongs to the value architecture (Paper VI) and to democratic theory (Paper III). The structural question of whether the loop can pursue those goals effectively belongs to this paper.

What the paper does claim is that, whatever goals a governance system pursues, it cannot pursue them effectively without adequate L. Legitimacy is not a luxury. It is a gain parameter. And a controller that treats it as optional—that ignores its own L, that borrows against it without limit, that pursues ambitious targets while it is collapsing—will eventually discover that the architecture it designed, however elegant, no longer functions. The rest of this paper is about why, and what to do about it.

Part II — Formal Framework: Legitimacy as Emergent Coupling State

The Governance as Engineering series has, from its first paper, used the state-space formalism of control theory to model governance systems. A system is described by a state vector **x**(t), an actuation matrix **B** that determines how control inputs affect the state, and an observation matrix **C** that determines what the controller can perceive. Papers I through XII have treated **B** and **C** as architectural primitives—properties of the institutional design that the designer chooses and that degrade through identifiable structural mechanisms: latency, aggregation, projection, noise.

This paper introduces a parameter that is not chosen by the designer. It emerges from the interaction between the governance architecture and the population it governs, and it simultaneously modulates both **B** and **C**. That parameter is legitimacy, denoted $L(t) \in [0,1]$. The formal move is to treat L(t) not as an external context but as an endogenous state variable of the control loop—one that couples the two channels the series has so far treated separately.

2.1 The Standard State-Space Model with a Legitimacy Parameter

The baseline model of the series is the discrete-time linear system:

$$\mathbf{x}(t+1) = \mathbf{A}\cdot\mathbf{x}(t) + \mathbf{B}\cdot\mathbf{u}(t) + \mathbf{w}(t)$$
$$\mathbf{y}(t) = \mathbf{C}\cdot\mathbf{x}(t) + \mathbf{v}(t)$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the true state of the governed system, $\mathbf{u}(t) \in \mathbb{R}^m$ is the control input, **w**(t) is process noise with covariance **W**, **y**(t) is the observed signal, and **v**(t) is measurement noise with covariance **V**.

In Papers I through XII, the matrices **A**, **B**, and **C**, and the noise covariances **W** and **V**, are treated as architectural parameters. They can be degraded by structural failure modes—latency, aggregation, projection, boundary mismatch—but the degradations are themselves properties of the institutional design. A designer who shortens the representation chain (Paper III) improves **C**; a designer who reduces delegation depth (Paper XI) improves **B**.

Legitimacy operates differently. It is not a property of the institutional design. It is a property of the governed population's willingness to cooperate with that design. And it affects both channels simultaneously.

Actuation legitimacy (L_B). When a controller issues a directive **u**(t), the actual control input that reaches the system is not **u**(t) but $\mathbf{u}_{eff}(t) = L_B(t) \cdot \mathbf{u}(t)$, where $L_B(t) \in [0,1]$ is the fraction of the population that complies with the directive. A tax reform is announced; L_B is the proportion of taxpayers who actually pay

at the new rate. A public health order is issued; L_B is the proportion who follow it. A regulation is promulgated; L_B is the proportion of regulated entities that implement it without litigation, delay, or evasion. The effective actuation matrix is:

$$\mathbf{B}\_\text{eff}(t) = L\_B(t) \cdot \mathbf{B}$$

When L_B = 1, the controller's directives are fully executed. When L_B = 0.5, half the actuation capacity is lost—not through any failure of the institutional machinery, but because the machinery's commands are not being obeyed. The distinction matters for diagnosis: a governance failure that appears to be an actuation deficit may in fact be a legitimacy deficit, and the appropriate response is not to redesign the actuation chain but to rebuild the trust on which it depends.

Observation legitimacy (L_C). When the controller collects information about the system's state—through surveys, administrative data, regulatory filings, or sensor networks—the accuracy of that information depends on the willingness of the governed to report honestly. A statistical agency surveys business activity; L_C is the probability that a respondent reports accurately rather than strategically. A regulatory inspector visits a facility; L_C is the probability that the operator discloses violations rather than concealing them. A citizen answers a government consultation; L_C is the probability that the response reflects the citizen's genuine preference rather than what the citizen believes the government wants to hear.

The measurement noise covariance $\mathbf{V}(t)$ is not fixed. It is a decreasing function of L_C(t). The simplest parameterization, and the one this paper adopts for analytical clarity, is:

$$\mathbf{V}(t) = \mathbf{V}_0 / L\_C(t)$$

where $\mathbf{V}_0$ is the baseline noise covariance when legitimacy is perfect (L_C = 1). When L_C = 0.5, measurement noise is doubled. When $L\_C \to 0$, measurement noise diverges to infinity—the observation channel is not merely degraded; it is destroyed. The controller is no longer receiving information about the true state of the system. It is receiving noise.

The observation equation becomes:

$$\mathbf{y}(t) = \mathbf{C}\cdot\mathbf{x}(t) + \mathbf{v}(t), \quad \mathbf{v}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_0 / L\_C(t))$$

The two legitimacy parameters, L_B and L_C, are conceptually distinct. A population may comply with directives while lying about the results (high L_B, low L_C: the authoritarian illusion). A population may report honestly while refusing to comply with directives (low L_B, high L_C: the protest democracy). In the general case, they are separate variables. But they are positively correlated through a common dependence on underlying trust: a government that is trusted to use power well tends to be both obeyed and told the truth; a government that is distrusted tends to face both non-compliance and strategic reporting. For much of the analysis that follows, the paper works with a composite L(t), treating $L\_B \approx L\_C$ as a reasonable

approximation for systems where legitimacy is broadly distributed across domains. The multidimensional extension, in which different institutions command different legitimacy levels, is noted as a direction for future work and does not alter the core dynamics.

2.2 Legitimacy Dynamics as an Endogenous Scheduling Variable

Legitimacy is not fixed. It evolves in response to the controller's performance and behaviour. The dynamics can be formalised as:

$$L(t+1) = \text{clip}(L(t) + \Delta L(t), \quad 0, \quad 1)$$

where $\Delta L(t)$ is the change in legitimacy driven by three primary mechanisms:

The delivery gap. The most direct driver of legitimacy is the gap between what the controller promises and what it delivers. Citizens form expectations about outcomes—economic growth, service quality, public safety—and update their trust based on the discrepancy between expectation and reality. Formally:

$$\Delta L_{\text{delivery}}(t) = -\alpha \cdot \|\mathbf{x}(t) - \mathbf{x}_{\text{promised}}(t)\|^2$$

where $\mathbf{x}_{\text{promised}}$ is the state the controller committed to achieving, and α captures the population's sensitivity to delivery failure. The quadratic form reflects the empirical regularity that large failures damage legitimacy disproportionately—a missed budget target by 5% is more than five times as damaging as a miss by 1%.

The transparency signal. Legitimacy is not only about outcomes. It is also about process. A controller that operates transparently—publishing its data, explaining its decisions, acknowledging its errors—generates a positive transparency signal that partially offsets delivery failures. Formally:

$$\Delta L_{\text{transparency}}(t) = +\beta \cdot T(t)$$

where $T(t) \in [0,1]$ is the controller's chosen transparency level, and β captures the population's responsiveness to openness. The transparency channel is the mechanism through which the controller can invest in future legitimacy, even at the cost of revealing inconvenient truths in the present.

The betrayal cost. The most damaging event for legitimacy is not failure but deception. When the governed discover that the controller has been systematically manipulating information—suppressing unfavourable statistics, punishing honest reporting, constructing a narrative that diverges from observable reality—the legitimacy penalty is catastrophic. Formally:

$$\Delta L_{\text{betrayal}}(t) = -\gamma \cdot D(t)$$

where D(t) is an indicator (or continuous measure) of revealed deception, and γ is the betrayal sensitivity. Critically, γ is not a constant. It is substantially larger for *borrowed* legitimacy than for *built* legitimacy, a distinction formalised in Section 2.4.

The full legitimacy dynamics are:

$$L(t+1) = \text{clip}\left(L(t) - \alpha \cdot \|\mathbf{x}(t) - \mathbf{x}\_\text{promised}(t)\|^2 + \beta \cdot T(t) - \gamma \cdot D(t) + \delta,\quad 0,\quad 1 \right)$$

where δ is a small exogenous drift term capturing the slow, secular accumulation or erosion of institutional trust that occurs independently of any single government's performance.

Hysteresis asymmetry. The dynamics above treat the delivery gap symmetrically: a positive gap (over-delivery) increases L at the same rate that a negative gap (under-delivery) decreases it. This is empirically false. Trust is destroyed more rapidly than it is rebuilt. To capture this, the sensitivity parameter α is made state-dependent:

$$\alpha = \alpha\_\text{drop} \quad \text{if} \quad \Delta\_\text{error}^2 > 0 \quad \text{(performance worsening)}$$
$$\alpha = \alpha\_\text{recovery} \quad \text{if} \quad \Delta\_\text{error}^2 \leq 0 \quad \text{(performance improving)}$$

with $\alpha\_\text{drop} \gg \alpha\_\text{recovery}$. A government that under-delivers for one period loses L quickly; a government that over-delivers for one period gains L slowly. The asymmetry is not a psychological assumption; it is an empirical regularity with deep evolutionary roots—organisms that updated positive expectations as quickly as negative ones would be vulnerable to exploitation. Whatever its origin, the consequence for governance is the hysteresis loop documented in Section 2.3: the path from high L to low L is short; the path back is long.

The controller must now manage a system in which its own effectiveness parameter L(t) is a function of its performance and transparency. This is a *nonlinear state-dependent feedback system*: the scheduling variable L(t) is closed-loop and error-dependent. The formal analysis of such systems draws on the absolute stability criteria—the Popov Criterion and the Circle Criterion—that determine whether a sector-bounded nonlinearity in the feedback path preserves or destroys system stability. The legitimacy trap, introduced next, is precisely the condition under which the coupling nonlinearity exits the stable sector.

2.3 The Kalman Filter and the Collapse of Observation Legitimacy

The observation-legitimacy parameter L_C governs the noise covariance $\mathbf{V}(t) = \mathbf{V}_0 / L\_C(t)$. When L_C falls, measurement noise rises. This has a precise and severe consequence for the controller's ability to estimate the system's true state.

A well-designed controller does not use raw observations **y**(t) directly. It passes them through a state estimator—in the optimal case, a Kalman filter—that combines the noisy measurement with a prediction from the internal model to produce a minimum-variance estimate $\hat{\mathbf{x}}(t)$ of the true state. The Kalman gain **K**_k

determines how much weight the estimator gives to new measurements relative to the model-based prediction:

$$\mathbf{K}\_k = \mathbf{P}\_k \mathbf{C}^T (\mathbf{C} \mathbf{P}\_k \mathbf{C}^T + \mathbf{R})^{-1}$$

where $\mathbf{P}\_k$ is the error covariance of the state estimate and $\mathbf{R} = \mathbf{V}_0$ / L_C is the measurement noise covariance.

As L_C → 0, $\mathbf{R} \to \infty$. As $\mathbf{R} \to \infty$, the Kalman gain $\mathbf{K}\_k \to 0$. A Kalman gain of zero means the estimator ignores new measurements entirely. It updates its state estimate purely by propagating the internal model forward via the dynamics matrix **A**:

$$\hat{\mathbf{x}}(t+1) = \mathbf{A}\cdot\hat{\mathbf{x}}(t) + \mathbf{B}\cdot\mathbf{u}(t)$$

The controller is now operating open-loop. It is not responding to the world. It is responding to its own model of the world, projected forward without correction. Every discrepancy between the model and reality accumulates unseen. The controller's dashboard shows the state that the model predicts, not the state that exists.

This is the formal mechanism of *dashboard insulation*—the condition in which a governance system's internal picture of its own performance diverges systematically from observable reality. The mechanism does not require conspiracy, propaganda, or deliberate deception, though those can accelerate it. It requires only that L_C fall low enough that the Kalman gain approaches zero. At that point, the architecture that was designed to maintain situational awareness becomes a machine for maintaining situational ignorance. The controller is not merely uninformed. It is misinformed by its own model, and the error compounds at a rate determined by the gap between **A** and the true system dynamics.

The practical governance manifestations of dashboard insulation are familiar from the series' country cases. The Soviet Gosplan's late-period economic statistics, which showed growth while the economy stagnated. The Chinese promotion tournament's systematically optimistic local reporting. The Russian power vertical's progressive loss of contact with battlefield reality. In each case, the observation channel was not merely degraded by architecture. It was destroyed by the collapse of L_C to the point where reporting the truth became individually irrational for the agents who possessed it.

2.4 The Legitimacy Trap as a Sector-Bounded Nonlinearity

The coupling between L, actuation, and observation creates the possibility of a self-reinforcing collapse. When L falls, actuation weakens. Outcomes deteriorate. The delivery gap widens. L falls further. Simultaneously, rising observation noise conceals the true state, making the controller's interventions increasingly miscalibrated. The controller, blind to its own blindness, redoubles its efforts—applying interventions that are too large or too small, at the wrong time, in the wrong place—and the resulting deterioration further erodes L.

This is the *legitimacy trap*: a positive-feedback spiral in which the parameter that makes governance effective is progressively destroyed by the ineffectiveness that its own destruction produces. Formally, it is a bifurcation in the nonlinear system dynamics. When L is above a critical threshold L_crit, the system has a single stable equilibrium at high performance and high legitimacy. When L falls below L_crit, the high-performance equilibrium disappears or becomes unreachable, and the system is drawn into a low-L, low-performance attractor from which recovery requires either external intervention or a sustained period of conditions—modest targets, maximum transparency, consistent delivery—that the depleted actuation capacity makes difficult to achieve.

The locus of the trap can be characterised using the absolute stability criteria for systems with sector-bounded nonlinearities. The legitimacy dynamics constitute a nonlinear feedback element in the loop. When the sector bounds of that element—determined by α, β, γ, and the hysteresis asymmetry—are compatible with the linear dynamics of the governance architecture, the system is absolutely stable: it converges to a high-L equilibrium regardless of initial conditions. When the sector bounds are violated, the system becomes conditionally stable or unstable: there exists a region of the state space from which it diverges toward the low-L attractor.

The governance interpretation is direct. A well-designed architecture with moderate delivery sensitivity (α), high transparency responsiveness (β), and low betrayal sensitivity (γ) is absolutely stable in the legitimacy dimension. It can absorb shocks to L—a scandal, a recession, a policy failure—and recover. An architecture with high delivery sensitivity, low transparency, or high betrayal sensitivity is conditionally stable. It can maintain high L under favourable conditions, but a sufficiently large shock can push it below L_crit, after which the system's own dynamics drive it deeper into the trap rather than back toward recovery.

The distinction between built and borrowed legitimacy, introduced in the next section, is precisely a parameterisation of these sector bounds.

2.5 Borrowed vs. Built Legitimacy

Not all legitimacy is structurally identical. The same observed L can rest on fundamentally different foundations, and those foundations determine the system's resilience to shocks.

Built legitimacy arises from consistent, transparent delivery over extended periods. The population trusts the controller because the controller has repeatedly demonstrated that it does what it says, reports honestly about what it did, and corrects its errors when they occur. The parameters of built legitimacy are:

- **α** (delivery sensitivity): moderate. Individual failures are understood as exceptions, not as revelations of systemic incompetence.
- **γ** (betrayal sensitivity): low. Because the controller has a long track record of transparency, a single revelation of deception is more likely to be interpreted as an aberration than as evidence of systematic dishonesty.

- **δ** (exogenous persistence): high. Built legitimacy decays slowly even in the absence of positive reinforcement, because it is embedded in institutional memory and cultural norms rather than in the performance of the current government.

Built legitimacy functions as a structural stabiliser. It damps the feedback loop between delivery failure and L erosion, giving the controller time to correct course before the trap closes.

Borrowed legitimacy arises from narrative, charisma, enemy construction, or temporary success. The population trusts the controller not because of a consistent track record but because of a compelling story about the controller's identity, intentions, or enemies. The parameters of borrowed legitimacy are:

- **α** (delivery sensitivity): high. Because the controller's legitimacy rests on the narrative rather than on demonstrated competence, a delivery failure that contradicts the narrative is disproportionately damaging—it undermines the story on which the entire trust relationship depends.
- **γ** (betrayal sensitivity): very high. When borrowed legitimacy is punctured by revealed deception, the collapse is catastrophic. The population infers not merely that the controller made an error, but that the narrative was fraudulent from the beginning. The betrayal cost is not proportional to the deception; it is proportional to the gap between the narrative and the revealed reality.
- **δ** (exogenous persistence): low. Borrowed legitimacy is not embedded in institutions or cultural norms. It is attached to specific leaders, narratives, or circumstances, and it evaporates rapidly when those supports are removed.

The borrowed-legitimacy architecture is structurally brittle. It can maintain high L for extended periods under favourable conditions—the narrative holds, the economy grows, the enemies remain threatening. But it is exquisitely vulnerable to shocks that breach the narrative. When the breach occurs, L does not decline gradually. It collapses, and the collapse is amplified by the very mechanisms—transparency suppression, narrative control—that the architecture used to maintain borrowed legitimacy in the first place.

The Soviet Union in the 1980s is the canonical example. Decades of borrowed legitimacy—the narrative of historical inevitability, the construction of external enemies, the suppression of economic data—maintained high apparent L until glasnost and the accumulating weight of observable failure breached the narrative. The collapse, when it came, was not a gradual decline in trust but a near-instantaneous evaporation of the legitimacy on which the entire architecture depended. The architecture did not fail because its institutions were technically incapable. It failed because the legitimacy that multiplied their effectiveness was borrowed, and the debt was called.

2.6 Gain-Scheduling: Adapting Control to Legitimacy Level

A rational controller should not apply the same control strategy regardless of its own legitimacy level. The optimal strategy depends on L, because L determines the effective actuation and observation capacity available.

High-L regime (L > L_high). The controller can pursue ambitious targets with high gain. Actuation is reliable; observation is accurate; the feedback loop is tight. Large, transformative reforms are feasible because the controller can count on compliance and honest reporting. The primary governance challenge in this regime is to maintain the conditions—delivery, transparency—that keep L high.

Moderate-L regime (L_crit < L < L_high). The controller should reduce its gain and invest in transparency. Ambitious targets carry the risk of a delivery gap that pushes L toward the trap threshold. The controller should pursue incremental, reversible actions that demonstrate reliability without staking legitimacy on outcomes it cannot guarantee. Transparency investment—publishing data, acknowledging errors, consulting affected populations—builds L at the cost of revealing inconvenient truths in the short term.

Low-L regime (L < L_crit). The controller is in or approaching the legitimacy trap. Ambitious action is counterproductive: the depleted actuation capacity makes delivery failure likely, and the amplified observation noise makes miscalibration inevitable. The controller should operate in a *legitimacy-rebuilding mode*: minimal targets, maximum transparency, and small, visible, delivered commitments that accumulate a track record of reliability. The goal is not to solve the system's substantive problems directly—the controller lacks the effective capacity to do so—but to rebuild the L on which all future capacity depends.

This is the structural analogue of "earning trust back." It is not a moral prescription. It is a control strategy, derived from the dynamics of the LPV system. A controller that ignores its own L and pursues ambitious targets from the low-L regime is mathematically likely to destroy what remains of its legitimacy, because the control energy it applies will be absorbed by the very parameter it is ignoring.

Legitimacy sensors. Gain-scheduling on L requires the controller to observe L directly. This means monitoring trust surveys, compliance rates, reporting latency, participation metrics, and the divergence between official and independent data sources. A controller that does not measure its own legitimacy is operating open-loop on the parameter that schedules its own effectiveness. The design implications are developed in Part VI.

2.7 Relationship to Architectural Primitives

Legitimacy is not a primitive in the series' sense. The designer cannot choose L. But the designer can choose the architecture that generates or erodes L over time. The relationship between the architectural primitives of Papers I–XII and the emergent coupling state L is the central structural insight of the paper.

A governance system with low latency (Paper I), short representation chains (Paper III), high observation dimensionality (Paper VI), well-matched boundaries (Paper XII), and protected observer diversity (Paper X) will *tend* to generate high L. It delivers outcomes reliably, because its control loop is tight. It reports honestly, because its observation channels are diverse and protected. The governed population learns, over time, that compliance is rewarded and honesty is safe.

A governance system that violates these primitives will *tend* to generate low L. Its delivery is inconsistent because its control loop is slow and its actuation is attenuated. Its reporting is distorted because its observation channels are narrow and manipulable. The governed population learns, over time, that compliance is futile and honesty is dangerous.

But the relationship is stochastic, path-dependent, and subject to hysteresis. A system with improving architecture may face low L for an extended period, because the population's trust has been depleted by the preceding period of dysfunction and recovers more slowly than the architecture improves. A system with deteriorating architecture may enjoy high L for a period, because borrowed legitimacy sustains trust beyond the point at which the underlying architecture would justify it—until the borrowing runs out.

The architectural primitives create the conditions for legitimacy. They do not guarantee it. And once L is established, it feeds back on the primitives' effectiveness with multiplicative force. This is why legitimacy is best understood not as a twelfth primitive but as the series' first endogenous coupling state: the variable that emerges from the interaction between architecture and society, and that determines whether the architecture works.

Part III — Failure Modes of Legitimacy

The formal framework of Part II establishes that legitimacy is an emergent coupling state—a parameter L(t) that modulates both actuation and observation, evolving under its own dynamics in response to delivery, transparency, and deception. When L is high, the governance architecture functions as designed. When it collapses, the same architecture becomes unsteerable and blind. This part identifies the specific failure modes that legitimacy dynamics produce. They are not independent of the structural failures diagnosed in earlier papers. They are the mechanisms through which those structural failures are amplified, accelerated, and made self-reinforcing by the governed population's loss of trust.

3.1 The Performance–Legitimacy Spiral

The most direct failure mode is the self-reinforcing collapse that occurs when a delivery failure reduces L, which reduces the capacity to deliver, which produces further delivery failures. The loop is simple in structure and devastating in consequence.

A government announces a major reform—an economic stabilisation programme, a health system restructuring, an infrastructure investment plan—and commits to specific, visible outcomes. The initial L is moderate to high. The population expects delivery. The reform encounters the normal difficulties of implementation: cost overruns, institutional resistance, exogenous shocks. The delivery gap—the squared distance between the promised state **x**_promised and the achieved state **x**(t)—begins to widen.

The legitimacy dynamics of Section 2.2 activate. The delivery gap reduces L. The reduction in L reduces compliance (L_B falls): taxpayers become less willing to pay the full amount, regulated entities challenge directives in court rather than implementing them, local officials slow-walk central instructions. The same reduction in L reduces reporting honesty (L_C falls): the statistical agency's surveys receive strategic rather than accurate responses, the regulatory inspectors are denied access or misled, the performance metrics reported upward are increasingly fictional.

The controller now faces a double degradation. Its effective actuation matrix has contracted, making it harder to achieve the outcomes that would close the delivery gap. Its observation channel has become noisier, making it harder to perceive the true state of the system and calibrate interventions accurately. The controller, observing a worsening situation through a degrading sensor, applies corrective actions that are increasingly miscalibrated—too large or too small, at the wrong time, in the wrong place. The delivery gap widens further. L falls further. The loop closes.

The spiral is not a theoretical possibility. It is the structural core of many state capacity collapses. The Argentine crisis of 2001 followed this trajectory: a currency board that had borrowed legitimacy from early success encountered external shocks, the delivery gap widened, compliance with the banking system and tax authority collapsed, and the government's capacity to stabilise the situation eroded precisely as the demands on that capacity intensified. The Greek sovereign debt crisis exhibited the same architecture: a government that had borrowed legitimacy through reported fiscal discipline was revealed to have systematically misreported its deficits; the resulting collapse in L destroyed both market access (an actuation channel) and the reliability of official statistics (an observation channel), making the subsequent stabilisation programme vastly more difficult than the same programme would have been under high-L conditions.

The spiral is driven by the asymmetry in the legitimacy update parameter α. When performance is worsening, α = α_drop is large; L falls quickly. When the controller attempts to recover, α = α_recovery is small; L rises slowly. The mathematics of the spiral are unforgiving: the rate of descent exceeds the rate of ascent, and a controller that does not account for this asymmetry will repeatedly fall back into the trap before recovery can take hold. The design implication—gain-scheduling on L, with reduced targets and increased transparency during recovery—is developed in Part VI. The structural point here is that the spiral is not a failure of the controller's intentions or competence. It is a dynamical property of the legitimacy-coupled system, and it will occur whenever a controller with moderate-to-high delivery sensitivity (α) pursues ambitious targets without monitoring its own L.

3.2 The Transparency Trap

The second failure mode is more insidious, because it appears to work until it catastrophically doesn't. A controller with declining L may attempt to maintain the appearance of legitimacy by suppressing or manipulating the observation channel—redefining metrics, restricting access to data, punishing honest reporting, and controlling the public narrative. In the short term, this can sustain L at a higher level than the underlying delivery performance would support. The controller is *borrowing* legitimacy against a future betrayal cost.

The mechanism exploits the structure of the L dynamics. Legitimacy responds to *perceived* delivery gaps, not necessarily to actual ones. If the controller can manipulate the reported state $\mathbf{x}\_rep(t)$ to appear closer to the promised state than the true state $\mathbf{x}(t)$ actually is, the perceived delivery gap is smaller than the real one, and L declines more slowly—or even holds steady—despite deteriorating actual performance. The controller can achieve this manipulation by controlling the transparency parameter T(t): setting it low, suppressing independent observation channels, and ensuring that the governed population receives a filtered or fabricated picture of outcomes.

The problem is that this strategy degrades the observation channel on which the controller itself depends. As L_C falls—because the governed learn that reporting honestly is punished or that official statistics are manipulated—the measurement noise covariance $\mathbf{V}(t) = \mathbf{V}_0 / L\_C(t)$ rises. The Kalman gain $\mathbf{K}\_k$ approaches

zero. The controller's own state estimator begins to ignore incoming measurements and relies entirely on its internal model, propagating its own expectations forward uncorrected.

The result is the *dashboard insulation* formalised in Section 2.3. The controller is no longer governing the real system. It is governing a phantom—the system as its own model predicts it to be. The discrepancy between the phantom and reality accumulates unseen, because every channel that could reveal it has been degraded by the same transparency suppression that is propping up apparent L.

This cannot continue indefinitely. The hidden discrepancy is a form of architectural debt. It accumulates in the unobserved dimensions of the system—the ecological degradation that GDP statistics exclude, the financial fragility that regulatory reports conceal, the social unrest that official media does not cover. Eventually, the debt is called. A crisis breaches the suppression apparatus—a financial collapse that cannot be hidden, an environmental catastrophe that is visible from space, a military defeat that shatters the narrative of competence. At that moment, the governed population perceives not merely that the controller has failed, but that it has been systematically deceiving them.

The betrayal cost is catastrophic. The legitimacy update equation activates the $-\gamma \cdot D(t)$ term, where $D(t)$ represents the magnitude of revealed deception and γ is the betrayal sensitivity. For borrowed legitimacy, γ is very large—far larger than the α that governs normal delivery failures. L does not decline. It collapses. And it collapses to a level lower than the underlying delivery performance alone would have produced, because the deception has destroyed the very possibility of trust. The controller is now in the low-L regime, with depleted actuation and a destroyed observation channel, precisely when it faces the crisis that its own suppression strategy prevented it from anticipating.

The transparency trap is not a hypothetical. The Soviet Union's terminal crisis followed exactly this trajectory. Decades of borrowed legitimacy—the narrative of historical inevitability, the external enemy construction, the systematic suppression of economic and social data—maintained high apparent L while the underlying system stagnated. Glasnost was an attempt to open the observation channel and rebuild L_C before the hidden discrepancies became fatal. It came too late. The revelation of the gap between the narrative and reality was so severe—γ was so large for a system whose legitimacy was entirely borrowed—that L collapsed to near-zero, and the architecture disintegrated with it.

The transparency trap is the structural mechanism behind the Measurement Paradox of Paper VIII: the very act of suppressing information to protect legitimacy destroys the information on which legitimacy ultimately depends. The trap is not escapable through better suppression technology. It is only escapable through transparency—accepting the short-term L cost of revealing uncomfortable truths in exchange for maintaining the observation channel on which long-term L depends. The controller that chooses suppression over transparency is choosing a trajectory that leads, with mathematical certainty, to a future state in which both L and observability are lower than they would have been under honesty.

3.3 High-Suppression Architectures and Their Fragility

The third failure mode is an extension of the transparency trap to the structural level. Some governance architectures are organised around a permanent suppression of observation-legitimacy (L_C) in specific dimensions, while maintaining actuation-legitimacy (L_B) through a combination of performance delivery, narrative control, and coercion. These architectures can be stable for extended periods, but they are brittle in a specific and predictable way.

The architecture works by separating the channels. The controller delivers enough on the dimensions that are visible and valued—economic growth, internal security, national prestige—to maintain L_B at an adequate level. Compliance is reasonably high; directives are followed. Simultaneously, the controller suppresses the observation channels for dimensions that would reveal failures or trade-offs: environmental degradation, inequality, corruption, the true costs of policy decisions. L_C is low for those dimensions, but the low L_C does not immediately affect L_B, because the governed population does not observe the suppressed dimensions clearly enough for them to influence compliance behaviour.

The fragility lies in the coupling between the channels. L_B and L_C are not perfectly separable. They are linked through the underlying trust parameter that both reflect. A revelation that the controller has been systematically suppressing information—even about a dimension that the population does not directly experience—can trigger a betrayal response that collapses L across all dimensions simultaneously. The mechanism is the one modelled in Section 2.5: borrowed legitimacy exhibits very high γ. A high-suppression architecture is, by its nature, a borrowed-legitimacy architecture. It has sacrificed the observation-channel integrity that builds long-term trust in exchange for the short-term appearance of competent governance. When the suppression is breached, the betrayal cost is total.

The Chinese calibration deficit, documented in the country study and revisited in Section 5.4, illustrates the mechanism. The promotion tournament creates high L_B for centrally monitored targets—local officials comply energetically with growth targets, infrastructure mandates, and stability directives. But it creates low L_C for the dimensions that are politically sensitive or difficult to measure: the true extent of local government debt, the severity of environmental pollution, the level of popular dissatisfaction. The system operates with a split legitimacy structure: high actuation-legitimacy, low observation-legitimacy.

This structure is stable as long as the suppressed dimensions do not generate crises large enough to breach the suppression apparatus. When they do—when local government debt reaches a scale that threatens the financial system, when environmental degradation produces a visible catastrophe, when popular dissatisfaction erupts into sustained protest—the revelation forces a simultaneous crisis in the suppressed dimension and a legitimacy collapse in the dimensions that were previously stable. The Zero-COVID reversal of late 2022 is a case in point: a policy that had been presented as a success was abandoned abruptly, revealing that the observation channel had been systematically distorted and that the true state of public compliance and epidemiological reality was different from what had been reported upward. The damage was not limited to the health dimension. It spilled into the broader legitimacy of the state's competence.

High-suppression architectures are not uniquely authoritarian. Democratic systems can exhibit the same pattern in specific domains. A financial regulator that suppresses evidence of systemic risk to maintain market confidence is operating a high-suppression architecture for that domain. A health agency that downplays the severity of an outbreak to avoid public panic is borrowing legitimacy against a future betrayal cost. The structural fragility is the same regardless of the political system: suppressed observation channels generate hidden discrepancies that, when revealed, trigger legitimacy collapses disproportionate to the underlying performance failure.

3.4 Legitimacy Hysteresis

The fourth failure mode is not a dynamic that produces collapse, but a dynamic that prevents recovery. It is the most subtle of the four, and arguably the most consequential for governance design.

The asymmetry in the legitimacy update parameter—$\alpha_{drop} \gg \alpha_{recovery}$—introduces hysteresis into the L dynamics. The path from high L to low L is steep and short. The path from low L back to high L is shallow and long. A controller that has experienced a legitimacy collapse cannot simply restore L by returning to the performance levels that previously sustained it. It must sustain significantly better performance—or a significantly longer period of adequate performance—to climb back to the same L level from which it fell.

The hysteresis loop has a precise structural consequence. Consider a governance system that experiences a major delivery failure—a financial crisis, a corruption scandal, a military defeat. L falls from L_{high} to L_{low}. The controller responds with reforms that improve underlying performance. After several years, the delivery gap is closed. By any objective measure, the system is performing as well as it was before the crisis. But L has not recovered to its pre-crisis level. It has recovered only part of the way. The population's trust is still depleted. Compliance is still below baseline. Reporting is still strategic rather than honest.

The controller, observing that its reforms have not restored legitimacy despite restored performance, faces a choice. It can continue the recovery effort—maintaining transparency, managing expectations, delivering consistently—for the extended period required to traverse the hysteresis gap. Or it can conclude that legitimacy is unattainable through performance alone and turn to alternative strategies: borrowing legitimacy through narrative, suppressing observation channels to hide the remaining gap, or coercing compliance rather than earning it.

The second path is the trap within the trap. A controller that abandons the long recovery trajectory in favour of shortcuts is likely to trigger the transparency trap (Section 3.2) or to entrench a high-suppression architecture (Section 3.3), both of which lead to eventual catastrophic collapse. The hysteresis dynamic thus creates a filtering mechanism: it selects for controllers that can sustain commitment to transparency and delivery over timescales that exceed the political cycle, and it selects against controllers that cannot.

The empirical evidence for legitimacy hysteresis is substantial. The post-communist transitions of the 1990s exhibited the pattern across multiple countries: even where economic reforms eventually produced growth and institutional improvements, trust in government remained depressed for a decade or more. The post-2008 financial crisis produced a sustained decline in institutional trust across the developed world that did not recover even as macroeconomic indicators improved. The Greek case is again instructive: by 2019, Greece had exited its bailout programme, returned to growth, and completed significant structural reforms. Trust in government remained among the lowest in Europe.

Hysteresis is not a psychological curiosity. It is a structural property of the legitimacy dynamics, and it has direct implications for the design of post-crisis recovery strategies. A controller that does not model hysteresis will underestimate the duration and intensity of the recovery effort required. It will declare victory prematurely, withdraw transparency commitments, and fall back into the trap. The design implications—legitimacy sensors, gain-scheduling, circuit-breaker mechanisms—are developed in Part VI.

3.5 Legitimacy Contagion

The final failure mode arises from the fact that L is not a single parameter attached to a monolithic controller. It is distributed across institutions, domains, and population segments. And it exhibits contagion dynamics: a legitimacy collapse in one institution can spread to others, because the governed population updates its trust in the system as a whole based on the performance of its parts.

Formally, legitimacy can be modelled as a vector $\mathbf{L}(t) = [L_1(t), L_2(t), \ldots, L_K(t)]$ where each L_k represents the legitimacy of a specific institution—the legislature, the judiciary, the central bank, the statistical agency, the executive. The dynamics of each L_k depend not only on that institution's own performance and transparency but also on the legitimacy of the institutions to which it is coupled:

$$L_k(t+1) = \text{clip}\left(L_k(t) - \alpha_k \cdot \text{delivery\_gap}_k + \beta_k \cdot T_k - \gamma_k \cdot D_k + \delta_k + \Sigma_j\, w_{kj} \cdot L_j(t),\ 0,\ 1 \right)$$

where w_{kj} represents the strength of the legitimacy spillover from institution j to institution k.

When w_{kj} is large and positive, high legitimacy in one institution supports legitimacy in others. An independent and trusted judiciary can anchor the legitimacy of the entire governance architecture, providing a reservoir of trust that persists even when the legislature or executive experiences temporary failures. When w_{kj} is large and negative—or when a legitimacy collapse is so severe that it overwhelms the positive spillovers—the contagion works in the opposite direction: a catastrophic failure in one institution drains legitimacy from the rest.

The 2008 financial crisis and its aftermath provide a clear illustration. The crisis originated in the financial sector—a domain whose regulatory institutions lost legitimacy when their failure to anticipate or prevent the crisis became apparent. But the legitimacy collapse did not remain confined to financial regulators. It spread to central banks, whose unconventional interventions were perceived as serving the financial sector at the expense of the public. It spread to legislatures, whose bailout authorisations were perceived as rewarding the

architects of the crisis. It spread to the broader "establishment" of technocratic governance. The result was the populist wave of the 2010s, which was substantially a legitimacy contagion event: the loss of trust in economic governance institutions spilled into a loss of trust in political institutions more broadly.

The structural vulnerability to contagion depends on the coupling architecture. A governance system with high institutional differentiation—where the central bank, the judiciary, the statistical agency, and the executive are perceived as genuinely independent—has lower effective w_kj between institutions, and a legitimacy shock to one is partially contained. A system where all institutions are perceived as branches of a single, undifferentiated governing apparatus has high effective w_kj, and a shock anywhere propagates everywhere.

This connects directly to Paper X's argument for observer diversity. An ensemble of institutions with decorrelated legitimacy dynamics—where trust in the statistical agency does not automatically rise and fall with trust in the executive—provides the governance equivalent of a diversified portfolio. The system can absorb a legitimacy shock to one component without losing the capacity to observe and act through the others. The design implication is that institutional independence is not only a matter of technical effectiveness or democratic principle. It is a structural requirement for legitimacy resilience. A system that concentrates all legitimacy in a single institution or leader is a system that will experience total legitimacy collapse when that institution fails.

The contagion dynamic also illuminates a pathway out of the legitimacy trap. If L can be rebuilt in one institution—a newly independent anti-corruption commission, a reformed statistical agency, a trusted central bank—the positive spillovers can gradually raise L across the broader architecture. The rebuilding strategy is not to attempt a simultaneous restoration of trust in all institutions, which the hysteresis dynamics make nearly impossible. It is to identify an institution whose legitimacy can be rebuilt most quickly, protect its independence, and allow the contagion to do its work in the positive direction. This is the structural logic behind the "islands of integrity" strategy observed in anti-corruption reforms and the "beachhead" approach to institutional development.

These five failure modes—the performance-legitimacy spiral, the transparency trap, high-suppression fragility, hysteresis, and contagion—are not a taxonomy of separate problems. They are different expressions of the same underlying dynamic: a coupling state L(t) that governs both channels of the control loop, that evolves under its own nonlinear dynamics, and that can trap a governance system in a low-performance, low-trust attractor from which recovery is possible but slow. The failure modes interact. A performance-legitimacy spiral can trigger a transparency trap if the controller attempts to arrest the spiral by suppressing observation. A high-suppression architecture can produce a contagion event when the suppressed dimensions breach. Hysteresis can make the recovery from any of these failures longer than the political conditions that enabled the recovery can be sustained.

The structural diagnosis is clear. Legitimacy is not a political luxury. It is a gain parameter with its own dynamics, its own traps, and its own recovery requirements. A governance architecture that does not model these dynamics—that treats L as exogenous or ignores it entirely—will eventually be destroyed by them. The design principles of Part VI specify what a legitimacy-sensitive architecture requires. The simulation of Part IV demonstrates the dynamics. The empirical illustrations of Part V ground them in cases. But the core insight is already in place: a controller that does not know its own L is operating blind on the parameter that determines whether its loop closes. And a controller that borrows L without understanding the terms of the debt is operating on borrowed time.

Part IV — Simulation: The Legitimacy-Driven Controller

The formal framework of Part II models legitimacy as an emergent coupling state L(t) that modulates both actuation and observation, evolving under its own dynamics in response to delivery, transparency, and deception. Part III traces the resulting failure modes—the performance–legitimacy spiral, the transparency trap, high-suppression fragility, hysteresis, and contagion—through their structural logic. This part subjects those dynamics to controlled simulation, demonstrating that the failure modes emerge reliably from the interaction of a well-designed controller with an endogenous legitimacy parameter, even when all other architectural primitives are held at ideal values.

The simulation is not a calibration to any specific real-world system. It is an existence proof: a demonstration that the qualitative dynamics the formal framework predicts—the trap, the hysteresis, the borrowed-legitimacy collapse—are generated by a minimal set of assumptions about how legitimacy responds to performance and transparency. The parameters are chosen to make the mechanisms visible. The code is open-source, with fixed seeds for replicability, Monte Carlo distributions across 100 seeds, and parameter sweeps demonstrating robustness.

4.1 Model Specification

The simulated world consists of a governance system controlling a two-dimensional state vector $\mathbf{x}(t) = [x_1(t), x_2(t)]^T$, representing, for concreteness, two policy-relevant dimensions such as economic output and environmental quality, or service delivery and fiscal balance. The true dynamics are linear and time-invariant:

$$\mathbf{x}(t+1) = \mathbf{A}\cdot\mathbf{x}(t) + \mathbf{B}\_\text{eff}(t)\cdot\mathbf{u}(t) + \mathbf{w}(t)$$

where $\mathbf{A} = 0.95\cdot\mathbf{I}_2$ captures slow autonomous drift toward zero (the target state), $\mathbf{w}(t) \sim \mathcal{N}(\mathbf{0}, 0.01\cdot\mathbf{I}_2)$ is process noise, and $\mathbf{B}\_\text{eff}(t)$ is the effective actuation matrix:

$$\mathbf{B}\_\text{eff}(t) = L\_B(t) \cdot \mathbf{B}$$

with $\mathbf{B} = \mathbf{I}_2$ representing the designed actuation capacity. When L_B = 1, the controller has full authority; when L_B = 0.5, half its control signal is lost.

The controller observes the system through a noisy measurement channel:

$$\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{v}(t), \quad \mathbf{v}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{V}(t))$$

where the measurement noise covariance is $\mathbf{V}(t) = \mathbf{V}_0 / L\_C(t)$, with $\mathbf{V}_0 = 0.05 \cdot \mathbf{I}_2$. When L_C = 1, measurement noise is at its designed minimum. As L_C falls, noise rises; as L_C → 0, the observation channel is destroyed.

The controller applies proportional state feedback based on its optimal estimate $\hat{\mathbf{x}}(t)$. The estimate is produced by a Kalman filter, as described in Section 2.3. The control law is:

$$\mathbf{u}(t) = -\mathbf{K} \cdot \hat{\mathbf{x}}(t)$$

where **K** is the linear quadratic regulator (LQR) gain computed for the nominal system (**A**, **B**, **Q** = **I**, **R** = 0.1·**I**), giving $\mathbf{K} \approx 0.75 \cdot \mathbf{I}_2$. The controller's target is **x**_target = **0**.

The controller also chooses a transparency level $T(t) \in [0,1]$ and a promised state trajectory **x**_promised(t). For simplicity, the simulation holds the promised state constant at **x**_promised = **0** (the target), so the delivery gap is simply the observed squared error $\|\hat{\mathbf{x}}(t)\|^2$. In extensions, the controller can adjust its promised trajectory to manage expectations.

Legitimacy dynamics. The composite legitimacy L(t) evolves according to the update equation of Section 2.2, with the hysteresis asymmetry of Section 2.4:

$$L(t+1) = \text{clip}\big(L(t) - \alpha(t) \cdot \|\mathbf{x}\_\text{rep}(t)\|^2 + \beta \cdot T(t) - \gamma \cdot D(t) + \delta,\quad 0,\quad 1 \big)$$

where:

- **x**_rep(t) is the *reported* state—the state as perceived by the governed population. When the controller is transparent (T high), **x**_rep(t) = **x**(t), the true state. When the controller suppresses transparency, **x**_rep(t) is a filtered version of the true state, as described below.
- $\alpha(t) = \alpha\_\text{drop}$ if $\|\mathbf{x}\_\text{rep}(t)\|^2 > \|\mathbf{x}\_\text{rep}(t-1)\|^2$ (performance worsening), and α_recovery otherwise, with $\alpha\_\text{drop} = 0.12$, $\alpha\_\text{recovery} = 0.03$ (a 4:1 asymmetry).
- $\beta = 0.08$ is the transparency sensitivity.
- γ is the betrayal sensitivity, set to $\gamma\_\text{built} = 0.5$ for built-legitimacy scenarios and $\gamma\_\text{borrowed} = 3.0$ for borrowed-legitimacy scenarios.
- D(t) is the deception revelation indicator, activated when the cumulative hidden discrepancy breaches a threshold.
- $\delta = 0.005$ is a small exogenous drift.

The split-state mechanism for the transparency trap. To model the transparency trap cleanly, the simulation distinguishes between the true state **x**(t) and the reported state **x**_rep(t). The controller chooses a suppression parameter $\lambda \in [0,1]$. When $\lambda = 1$, the reported state equals the true state: full transparency. When $\lambda < 1$, the reported state is a convex combination of the true state and a flattering reference:

$$\mathbf{x}\_\text{rep}(t) = \lambda \cdot \mathbf{x}(t) + (1-\lambda) \cdot \mathbf{x}\_\text{promised}(t)$$

The governed population updates L based on **x**_rep(t), not **x**(t). Suppression thus maintains higher apparent L in the short term. However, a hidden discrepancy variable E_betrayal(t) integrates the cumulative gap between the true and reported states:

$$E\_betrayal(t+1) = E\_betrayal(t) + \|\mathbf{x}(t) - \mathbf{x}\_rep(t)\|^2$$

The probability of revelation—the event that the deception becomes publicly visible—is modelled as a hazard rate that increases with E_betrayal:

$$P(\text{revelation at } t) = 1 - \exp(-h \cdot E\_betrayal(t))$$

where h = 0.02 is the hazard coefficient. When revelation occurs, D(t) is set to 1 for that time step and the suppression parameter λ is forced to 1 thereafter (the controller can no longer hide). The betrayal penalty $-\gamma \cdot D(t)$ then hits the legitimacy update with full force.

This split-state mechanism operationalises the Measurement Paradox of Paper VIII in the legitimacy domain: the very act of suppressing information to protect L builds a hidden debt whose eventual repayment is catastrophic.

Initial conditions and simulation length. The simulation runs for T = 300 time steps. The burn-in period is 20 steps. Initial legitimacy is set to L(0) = 0.7 for high-legitimacy scenarios and L(0) = 0.3 for low-legitimacy scenarios. The initial state is **x**(0) = **0** (on target). External shocks are introduced as temporary perturbations to **x** to test the system's resilience.

4.2 Scenarios

Four scenarios are simulated, corresponding to the failure modes of Part III.

Scenario 1: High-transparency, high-legitimacy equilibrium. The controller begins with L(0) = 0.7 and maintains full transparency (T = 1, λ = 1). No deception occurs. The controller pursues the target with the LQR gain. A moderate external shock is introduced at t = 50, displacing **x** from the origin. The scenario demonstrates that a transparent, high-L system absorbs shocks effectively: L remains high, the state returns to target, and no trap activates.

Scenario 2: The legitimacy trap. The controller begins with L(0) = 0.7 and full transparency, but at t = 50 a large external shock displaces **x** far from target. The resulting delivery gap is substantial. The controller applies its standard gain to correct the deviation, but the shock has already reduced L. The reduced L weakens actuation and degrades observation, making the controller's subsequent interventions less effective. The delivery gap persists. L continues to fall. The system spirals into the low-L attractor. The scenario demonstrates the self-reinforcing nature of the trap: the same control strategy that works at high L fails at low L, and the failure further reduces L.

Scenario 3: Recovery through transparency intervention. The controller begins in the low-L attractor (L(0) = 0.3, either exogenously or as the endpoint of Scenario 2). At t = 50, the controller switches to a legitimacy-rebuilding strategy: it reduces its control gain by half (recognising its depleted actuation), increases transparency to maximum (T = 1), and commits to a lower promised target that is achievable given the reduced capacity. It sustains this strategy for the remainder of the simulation. The scenario demonstrates the hysteresis gap: L recovers, but slowly. The time required for L to return to 0.6 is substantially longer than the time it took to fall from 0.6 to 0.3. The scenario also compares recovery trajectories with and without transparency investment, showing that transparency accelerates rebuilding even when it reveals uncomfortable truths.

Scenario 4: Borrowed-legitimacy collapse. The controller begins with moderate legitimacy (L(0) = 0.55) but low transparency (T = 0.2, λ = 0.3). The reported state **x**_rep(t) is substantially flattering: the governed population perceives outcomes as better than they are. Apparent L remains stable or declines only slowly, while the true state deteriorates due to accumulating process noise and the controller's miscalibrated interventions (which are themselves a consequence of the degraded observation channel). The hidden discrepancy E_betrayal(t) grows. At a stochastic trigger point, the deception is revealed. D(t) = 1, and the betrayal penalty γ_borrowed = 3.0 is applied. L collapses catastrophically—far below the level that honest governance with the same underlying performance would have produced. The scenario demonstrates the structural brittleness of borrowed legitimacy.

Parameter sweeps. The simulation conducts sensitivity analysis across key parameters:

- γ (betrayal sensitivity): swept from 0.5 to 5.0, demonstrating the transition from recoverable deception to catastrophic collapse.
- Suppression duration (time before revelation): swept from 10 to 200 steps, showing that longer suppression produces more severe collapses.
- α_drop / α_recovery ratio: swept from 1:1 (no hysteresis) to 10:1 (strong hysteresis), demonstrating the emergence of the trap as asymmetry increases.
- Initial L(0): swept from 0.1 to 0.9, mapping the basins of attraction for the high-L and low-L equilibria.

4.3 Expected Results and Key Figures

The simulation is designed to produce four primary outputs that operationalise the theoretical claims of Parts II and III.

Figure 1: Phase diagram in (L, T) space. A quiver plot or basin-of-attraction map showing, for a grid of initial (L, T) conditions, whether the system converges to the high-L equilibrium or the low-L attractor. The diagram reveals the separatrix—the critical L_crit below which the system is drawn into the trap regardless of transparency level. Built-legitimacy parameter settings (low γ, high δ) show a larger basin of attraction for the

high-L equilibrium and a lower L_crit. Borrowed-legitimacy settings (high γ, low δ) show a smaller high-L basin and a higher L_crit. The figure makes visible the structural difference between the two legitimacy regimes.

Figure 2: Time-series of the legitimacy trap and recovery. Three panels. The top panel shows the state norm $\|\mathbf{x}(t)\|$ over time for Scenario 2 (trap) and Scenario 3 (recovery). The middle panel shows L(t) for the same trajectories. The bottom panel shows the effective actuation and observation capacities, L_B(t) and L_C(t), illustrating how they co-move. The trap trajectory shows L and performance declining together; the recovery trajectory shows L lagging behind performance improvement, with the hysteresis gap clearly visible as the horizontal distance between the decline curve and the recovery curve.

Figure 3: Borrowed-legitimacy collapse. A single trajectory from Scenario 4, with two vertical axes. The top panel shows the true state norm $\|\mathbf{x}(t)\|$ and the reported state norm $\|\mathbf{x}\_rep(t)\|$. The divergence between the two is the hidden discrepancy. The middle panel shows apparent L (calculated from **x**_rep) and true L (calculated from **x**). Apparent L remains stable while true L declines. The bottom panel shows E_betrayal(t), the cumulative hidden discrepancy. A dashed vertical line marks the stochastic revelation event. At that moment, apparent L and true L converge downward, and the state norm spikes as the controller's remaining legitimacy is destroyed. The figure is the graphical signature of the transparency trap.

Figure 4: Collapse severity heatmap. A parameter sweep across suppression duration (x-axis) and betrayal sensitivity γ (y-axis). The colour map shows the minimum L reached after revelation. Long suppression and high γ produce the most severe collapses. Overlaid contours indicate the parameter region where post-collapse L falls below the trap threshold L_crit, meaning the system enters the low-L attractor and cannot recover without external intervention. The heatmap identifies the structural conditions under which borrowing legitimacy is survivable (short suppression, low γ, built-legitimacy baseline) and those under which it is fatal.

Summary metrics. For each scenario, the simulation reports: the fraction of Monte Carlo runs that enter the trap (L falls below L_crit and does not recover within the simulation window); the mean recovery time from L_low to L_high under the rebuilding strategy; the collapse magnitude (L_before_revelation − L_after_revelation) for borrowed-legitimacy scenarios; and the hysteresis gap (time or performance differential between decline and recovery trajectories).

The simulation does not predict specific outcomes for any real governance system. It demonstrates the qualitative dynamics that the formal framework identifies, under controlled conditions, with all non-legitimacy parameters held at ideal values. The fact that the failure modes emerge even under these idealised conditions—with perfect internal actuation and observation when L = 1, with optimal state estimation, with no adversarial actors—is the central simulation finding. The legitimacy trap is not a consequence of institutional dysfunction. It is a consequence of the coupling between performance, transparency, and trust in any system where that coupling exists. The simulation makes that consequence visible.

Part V — Empirical Illustrations

The formal framework of Part II and the failure modes of Part III make predictions about where and how legitimacy dynamics will manifest. This part examines five cases that span from stable high-trust equilibria to catastrophic borrowed-legitimacy collapses, and from national-scale governance to municipal service delivery. In each, the legitimacy parameter is the lens that makes the outcome legible, and in each, the structural logic of the trap, the transparency trade-off, and the hysteresis asymmetry are visible in the historical record.

5.1 The Nordic High-Trust Equilibrium

The Nordic governance systems—Denmark, Finland, Norway, Sweden—exhibit a configuration that the framework identifies as a stable high-L equilibrium. The composite legitimacy parameter in these countries is among the highest measured globally, as reflected in trust surveys, tax compliance rates, and the willingness of citizens to provide accurate information to government agencies.

The architecture that sustains this equilibrium is not a single institutional design choice but a mutually-reinforcing configuration of the primitives identified across the series. Representation chains are short by international standards (Paper III), with strong municipal government and extensive use of direct consultation. Observation channels are high-dimensional (Paper VI), with statistical agencies that enjoy constitutional independence and a cultural norm of honest reporting. Observer diversity is protected (Paper X), with strong public service broadcasting, well-resourced universities, and an active civil society providing independent verification of official claims. Actuation chains are short (Paper XI), with a tradition of decentralised implementation that gives local authorities genuine discretion and accountability.

These architectural properties generate high L_B (compliance) and high L_C (honest reporting). The resulting effective governance capacity is the product, not the sum, of the formal institutional quality and the legitimacy multiplier. When a Nordic tax authority issues a demand, it collects approximately 98% of the statutory amount. When a Nordic statistical agency surveys business conditions, response rates exceed 90% and strategic misreporting is rare. The same formal institutions, operated in a lower-L environment, would produce substantially worse outcomes—not because the institutions are different, but because the parameter multiplying their effectiveness is lower.

Critically, Nordic legitimacy is overwhelmingly *built* rather than *borrowed*. The high L is the product of decades of consistent delivery, high transparency, and institutionalised mechanisms for acknowledging and correcting errors. The betrayal sensitivity γ is low, because the population's trust is based on a long track

record of honesty rather than on a fragile narrative. When delivery failures occur—and they do, in every system—they are interpreted as exceptions rather than as revelations of systemic dishonesty. The legitimacy dynamics are in the absolutely stable regime: the system absorbs shocks without entering the trap.

The Nordic case also illustrates the hysteresis asymmetry in the favourable direction. Because L is high and built, the recovery from any temporary decline is slow but the decline itself is slow. The system's legitimacy has a large moment of inertia. This is the structural reward for patient, transparent governance: not that failures never happen, but that when they do, the architecture has the reserves to survive them.

5.2 The Eurozone Periphery and the Transparency Trap

The Greek sovereign debt crisis of 2010–2012 is the series' clearest illustration of the transparency trap and the borrowed-legitimacy collapse. The case demonstrates, with unusual clarity, how a governance system can maintain high apparent L while the underlying legitimacy substrate is being consumed, and how the eventual revelation triggers a catastrophic collapse that makes recovery vastly more difficult.

Greece joined the Eurozone in 2001 on the basis of reported fiscal statistics that met the Maastricht convergence criteria. Those statistics were systematically manipulated. The true fiscal deficit and public debt were substantially higher than reported, a fact that was known to parts of the Greek statistical system but was not reflected in the official figures communicated to European partners or to financial markets. The reported state **x**_rep(t)—fiscal discipline, economic convergence—diverged from the true state **x**(t) by a margin that accumulated over nearly a decade.

This is the split-state mechanism of Section 4.1 in operation. The suppression parameter λ was low; the reported state was a flattering composite of the true state and the promised state. International markets and European institutions treated Greek sovereign debt as nearly risk-equivalent to German debt, pricing in the *reported* fiscal position rather than the true one. Apparent legitimacy was high: Greece borrowed at favourable rates, participated in European decision-making, and was treated as a full member of the currency union. The hidden discrepancy E_betrayal(t) accumulated unseen.

In late 2009, the newly elected Greek government revealed that the 2009 deficit would be approximately 12.7% of GDP—more than double the previously reported figure. The revelation was the stochastic trigger of the model. The betrayal penalty γ was large, because the legitimacy of Greek fiscal governance had been almost entirely borrowed: it rested on the narrative of convergence and discipline rather than on a long track record of transparent delivery. When that narrative was revealed to be fraudulent, L collapsed.

The consequences were structural, not merely reputational. The collapse of L_B meant that subsequent austerity measures—tax increases, spending cuts, structural reforms—faced acute compliance deficits. Taxpayers who had been told that their government had lied about the fiscal situation were less willing to comply with new tax demands. Regulated entities challenged every measure. The collapse of L_C meant that official statistics, even after the statistical authority was reformed and granted independence, were met with

scepticism that undermined their utility for policy calibration. The controller was attempting to implement a stabilisation programme with a depleted actuation matrix and a destroyed observation channel—precisely the condition the model identifies as the legitimacy trap.

The recovery trajectory illustrates the hysteresis dynamic. Greece did not regain market access until 2019, a decade after the crisis began, despite implementing extensive structural reforms and achieving primary fiscal surpluses from 2016 onward. The objective performance indicators improved substantially. L did not recover correspondingly. Trust in Greek institutions remained among the lowest in the Eurozone even as the macroeconomic data improved. The hysteresis gap was, and remains, large. The European institutions that provided external financing—the Troika of the European Commission, European Central Bank, and International Monetary Fund—functioned in part as a legitimacy substitute: an external source of L that bridged the period during which domestic L was too depleted to support normal governance.

The Greek case is not an outlier. It is the transparency trap operating exactly as the model predicts, in a setting where the architectural debt of suppressed observation was large and the legitimacy was borrowed rather than built. The structural lesson is that transparency is not a luxury that can be sacrificed for short-term stability. It is the parameter that determines whether stability is possible at all.

5.3 South Africa's Truth and Reconciliation Commission

If Greece illustrates the catastrophic collapse of borrowed legitimacy, South Africa's Truth and Reconciliation Commission (TRC) illustrates a deliberate attempt to manage the transition from a collapsed-legitimacy regime to a new legitimacy baseline. It is the most prominent example of a governance system acknowledging past deception as a strategy for rebuilding the observation channel.

The apartheid state operated a high-suppression architecture: high actuation-legitimacy (L_B) for the white population whose compliance sustained the regime, and near-zero observation-legitimacy (L_C) for the dimensions that would have revealed the regime's violence, illegality, and unsustainability. The observation channel was systematically degraded—through censorship, surveillance, the criminalisation of dissent, and the suppression of data on the conditions of the black majority. By the late 1980s, the regime's borrowed legitimacy was collapsing under the weight of accumulated discrepancies between the official narrative and observable reality: economic sanctions, internal resistance, and the visible brutality of the security apparatus.

The transition of 1994 created a new governance architecture with a profound legitimacy problem. The new democratic government inherited institutions—the civil service, the judiciary, the security forces—whose legitimacy with the majority population was near zero. The parameter L was depleted across both channels: many citizens had no reason to trust that the state's directives were legitimate (low L_B) or that its information channels were honest (low L_C). The hysteresis problem was acute: even a democratic government committed to transparency and delivery would face an extended period during which its effective governance capacity was severely constrained by the depleted L it had inherited.

The TRC, which operated from 1996 to 1998, was a structural intervention in the legitimacy dynamics. Its function, in the framework's terms, was to reset the observation channel by publicly acknowledging the deception and violence of the preceding regime. By creating an official, transparent record of past abuses—taking testimony from victims and perpetrators, publishing findings, granting amnesty conditional on full disclosure—the TRC performed a legitimacy operation: it demonstrated that the new state would not continue the suppression strategies of the old. It signalled a regime change in transparency, from $\lambda \approx 0$ to $\lambda \approx 1$.

The TRC did not fully succeed in rebuilding L. The hysteresis gap was large; the material conditions of most South Africans did not improve rapidly; and the commission's work was contested by parties across the political spectrum. But it succeeded in a more limited and crucial sense: it prevented the new democratic state from inheriting the betrayal debt of the apartheid state. By acknowledging the past deception, it reduced the γ that would have been applied to any future revelation of suppressed information. It created, in effect, a new baseline: the legitimacy of the new state would be judged on its own performance, not on the sins of its predecessor.

The South African case is an illustration of a design principle that Part VI formalises: when legitimacy has collapsed, transparency about the collapse is the necessary precondition for rebuilding. A new controller that inherits a depleted-L architecture must openly acknowledge the depletion before it can credibly commit to a different trajectory. The TRC was an imperfect but structurally correct intervention in the observation channel—an attempt to restore L_C so that L_B could eventually follow.

5.4 China's Calibration Deficit Revisited

The China country study diagnosed a system in which the promotion tournament generates high compliance with centrally monitored targets and systematic distortion of the information reported upward. In the framework of this paper, China exhibits a split-legitimacy architecture: high actuation-legitimacy (L_B) for the dimensions the centre monitors closely, coupled with low observation-legitimacy (L_C) for the dimensions that are politically sensitive or difficult to verify.

The split is not accidental. It is a structural consequence of the system's approach to legitimacy management. The Chinese Communist Party maintains L_B through a combination of performance delivery—sustained economic growth, infrastructure development, poverty reduction—and the credible threat of sanction for non-compliance. Local officials comply with centrally mandated targets because their careers depend on it. The actuation channel, for the dimensions the centre prioritises, is effective.

But the same promotion tournament that produces high L_B degrades L_C. Local officials whose careers depend on reported performance have strong incentives to report favourable figures and suppress unfavourable ones. The observation channel for dimensions such as local government debt, environmental pollution, food safety, and public dissatisfaction is systematically distorted. The centre knows this—the calibration deficit is well-documented in Chinese policy discourse—but it cannot easily correct it without undermining the incentive structure that sustains L_B.

The architecture is therefore a borrowed-legitimacy structure for the observation channel. Apparent L_C—the centre's confidence in the information it receives—is maintained through a combination of selective auditing, anti-corruption campaigns, and the assumption that the overall picture is approximately correct despite local distortions. The hidden discrepancy E_betrayal(t) accumulates across multiple suppressed dimensions. The suppression parameter λ is low for these dimensions; the reported state diverges from the true state.

The Zero-COVID reversal of late 2022 illustrates the fragility of this architecture. The policy had been presented as a demonstration of the system's capacity to control the pandemic. Its implementation was enforced through the high-L_B actuation channel: local officials complied energetically with lockdown, quarantine, and testing mandates. But the L_C channel was degraded. Reports of public compliance, of the economic costs, and of the epidemiological situation were filtered through the same promotion incentives that distorted all other reporting.

When the policy was reversed—suddenly, and with minimal transition planning—the event functioned as a partial revelation of the hidden discrepancy. The gap between the official narrative of successful control and the observable reality of widespread protest, economic disruption, and the ultimately uncontrollable spread of the virus became visible to a large fraction of the population. The damage was not confined to the health dimension. It spilled into the broader legitimacy of the state's competence, exactly as the contagion mechanism of Section 3.5 would predict.

The Chinese case is not one of catastrophic L collapse—the regime's actuation-legitimacy remains substantial, and its observation-legitimacy, while damaged, has not been destroyed. But it illustrates the structural fragility of the split-legitimacy architecture: a crisis in a suppressed dimension can breach the suppression apparatus and damage legitimacy across dimensions that were previously stable. The system's resilience depends on the continued effectiveness of the suppression mechanisms, and the continued effectiveness of those mechanisms depends on the absence of crises large enough to overwhelm them. This is the definition of borrowed-legitimacy fragility.

5.5 Municipal Infrastructure Delivery

The legitimacy dynamics that operate at national scale also operate at the scale of a city government, and the municipal case provides a less politically charged illustration of the same mechanisms.

Consider a city government that announces a major public transit project—a new light rail line connecting underserved neighbourhoods to the urban core. The project is promised to be completed within five years, at a specified cost, with specified service levels. The announcement generates an initial increase in L: the government is seen to be addressing a genuine need, and the affected communities extend provisional trust.

The project encounters the normal difficulties of large infrastructure delivery. Geological conditions are worse than anticipated. Contractors underperform. Costs escalate. The completion date slips—first by a year, then by two, then indefinitely. The delivery gap between the promised state (functional transit by year five)

and the achieved state (partial construction, no service) widens.

The L dynamics activate. Residents of the affected neighbourhoods, having extended trust and been disappointed, reduce their compliance with other city initiatives. Property owners resist land acquisition for future projects. Voters reject bond measures. Community organisations that previously cooperated with city planning processes become adversarial. L_B falls. Simultaneously, the city's reporting on the project's status becomes less candid. Progress reports emphasise achievements and downplay delays. The observation channel degrades. L_C falls.

The city now faces a localised legitimacy trap. It needs to build the transit line to restore L, but the depleted L makes it harder to secure the cooperation, funding, and political support required to complete the project. The spiral is self-reinforcing.

The recovery strategy, if the city's leadership understands the dynamics, follows the gain-scheduling logic of Section 2.6. Rather than doubling down on the troubled project with diminished capacity, the city reduces its ambitions. It completes a small, visible segment of the line—one station, one kilometre of track—and puts it into operation. It publishes transparent, unvarnished reports on the status of the remaining construction. It acknowledges the original timeline was unrealistic and commits to a revised timeline that it can credibly meet. It delivers small, frequent, visible improvements—bus service enhancements, pedestrian infrastructure, station area improvements—that demonstrate reliability at a scale the depleted L can support.

These interventions do not immediately restore L to its pre-project level. The hysteresis gap ensures that trust returns more slowly than it was lost. Residents who were promised a transit line and received a single station are not quickly impressed. But the interventions arrest the downward spiral and begin the slow climb back. Over an extended period of consistent, transparent delivery, L recovers to a level where more ambitious projects become feasible again.

The municipal case is not dramatic. It is not a crisis of state capacity or a collapse of democratic legitimacy. It is the legitimacy dynamics operating at their quotidian scale—in the gap between what a government promises and what it delivers, and in the structural consequences of how it manages that gap. The same mathematics that drives sovereign debt crises and authoritarian fragility operates in the relationship between a city council and the neighbourhoods it serves. The trap is the same; the recovery strategy is the same. Only the scale differs.

This is the paper's most important empirical claim: that legitimacy is not a special property of national governments or democratic mandates, but a structural parameter of any governance relationship, operating at every scale from the municipal to the planetary. The design principles of Part VI apply at all of them.

Part VI — Design Principles for Legitimacy-Sensitive Architectures

The diagnosis is structural. Legitimacy is an emergent coupling state—a parameter L(t) that simultaneously modulates actuation effectiveness and observation fidelity, evolving under its own dynamics in response to delivery, transparency, and deception. The failure modes of Part III are the signatures of a controller that ignores this parameter or attempts to manipulate it at the expense of the observation channel. The empirical illustrations of Part V are the historical manifestations of these dynamics across scales.

This part turns from diagnosis to prescription. It specifies the design principles that follow from the formal framework and that a governance architecture must satisfy if it is to maintain adequate L over time, recover from L collapses, and avoid the traps that make L depletion self-reinforcing.

The principles are not a blueprint. The appropriate legitimacy strategy for any specific governance function depends on the sensitivity parameters of the governed population—α (delivery sensitivity), β (transparency responsiveness), γ (betrayal sensitivity)—and on the initial L conditions the system inherits. What the principles provide is a set of structural requirements that any legitimacy-sensitive architecture must meet, and a vocabulary for designing institutions that meet them.

6.1 Transparency by Default—Within Functional Bounds

The transparency trap of Section 3.2 demonstrates that suppressing observation channels to maintain apparent L is a strategy with a mathematically inevitable catastrophic termination. The only sustainable strategy is the opposite: transparency by default.

Transparency by default means that governance information—statistics, performance metrics, policy analyses, evaluation reports—should be publicly accessible unless there is a specific, legitimate, and bounded reason for confidentiality. The burden of proof falls on those who would restrict access, not on those who would provide it. The default is openness; secrecy is the exception that must be justified.

The structural logic follows directly from the L dynamics. When transparency is high, the reported state $\mathbf{x}_{rep}(t)$ tracks the true state $\mathbf{x}(t)$ closely. The governed population's perception of the delivery gap is accurate. L evolves in response to actual performance, not to a manipulated picture. This means L can decline when performance deteriorates—transparency does not prevent legitimacy losses—but it also means that the decline is proportional to the underlying problem, not amplified by the revelation of deception. The betrayal cost γ is not activated because there is no betrayal.

When transparency is low, the reported state diverges from the true state. The hidden discrepancy E_betrayal(t) accumulates. When revelation occurs—and it will, with a probability that increases with the discrepancy—the betrayal cost is applied. The resulting L collapse is more severe than the decline that honest reporting would have produced. The controller has traded a moderate, manageable L loss in the present for a catastrophic, unmanageable L loss in the future.

The qualification "within functional bounds" acknowledges that some governance functions require confidentiality to operate: diplomatic negotiations, military operations, criminal investigations, personal privacy, commercial confidentiality. Transparency by default does not mean the live-streaming of cabinet meetings or the publication of intelligence sources. It means that the perimeter of legitimate confidentiality is drawn narrowly, reviewed regularly, and enforced by institutions that are independent of the actors whose information would be restricted.

The design implications are specific:

- Statistical agencies should have constitutional or statutory independence, with protected budgets, fixed-term leadership appointments insulated from executive removal, and a legal obligation to publish raw data and methodology alongside findings.
- Freedom of information legislation should be structured so that the default is disclosure, exemptions are time-limited, and the adjudication of disputes is independent of the government whose information is requested.
- Whistleblower protection should extend to public servants who disclose evidence of suppression, manipulation, or politically motivated restriction of information—because the whistleblower is performing the structural function of maintaining L_C.

Transparency by default is not a democratic nicety. It is a structural requirement for maintaining the observation channel on which the controller's own effectiveness depends. The controller that erodes transparency to protect short-term L is the controller that blinds itself to the accumulating conditions of its own collapse.

6.2 Delivery–Reality Matching

The delivery gap is the primary driver of L erosion in the model of Section 2.2. When a controller promises an outcome that it fails to deliver, L declines—and with the hysteresis asymmetry of Section 3.4, it declines faster than it can be rebuilt. The structural implication is direct: a controller should not promise more than it can credibly deliver.

Delivery–reality matching is the principle that commitments should be calibrated to the controller's effective capacity, not to its aspirational capacity. The effective capacity is not the formal capacity—the budget, the legal authority, the institutional machinery—but the formal capacity multiplied by the current L. A controller

with L = 0.5 has half the effective actuation of a controller with L = 1, even if their formal architectures are identical. It should make commitments accordingly.

This principle runs against the grain of political incentives. Governments are rewarded, in the short term, for ambitious promises. The politician who promises a transformative reform, a rapid infrastructure build-out, or a quick end to a crisis generates an immediate increase in apparent L—the population extends provisional trust on the basis of the promise. But if the promise exceeds the capacity to deliver, the resulting delivery gap will destroy more legitimacy than the promise generated. The net effect over the political cycle is negative, even if the initial response is positive.

The structural response is to institutionalise modesty. This does not mean abandoning ambition. It means that ambition is expressed through a sequence of incremental, delivered commitments rather than through a single grand promise. Each delivery builds L, which increases the effective capacity for the next, more ambitious step. The trajectory is a virtuous spiral—the reverse of the performance-legitimacy spiral—in which growing L enables growing ambition, which enables further delivery, which further grows L.

The design implications include:

- Fiscal rules that require independent costing of policy proposals, so that the delivery gap between promised and funded is visible before the promise is made.
- Infrastructure procurement models that break large projects into independently deliverable segments, each of which generates a visible completion and a legitimacy dividend before the next begins.
- Regulatory impact assessments that estimate not only the compliance cost but the compliance probability —the likelihood that regulated entities will actually implement the regulation, given current L.
- Policy commitments that are explicitly contingent on capacity: "we will do X, and if we succeed at X within the specified timeframe and budget, we will then do Y." The conditional structure builds L through each successive delivery rather than staking L on the final, most ambitious outcome.

Delivery–reality matching is the operationalisation of gain-scheduling on L. When L is low, targets should be modest and achievable. When L is high, ambition can expand. The controller that ignores its own L in setting targets is the controller that generates the delivery gaps that will destroy it.

6.3 Legitimacy Sensors

A controller cannot adapt its strategy to its own legitimacy level if it does not know what that level is. The third design principle follows directly: governance architectures must include dedicated, protected mechanisms for measuring L.

Legitimacy sensors are institutionalised channels for monitoring the parameter on which the controller's own effectiveness depends. They measure L_B (compliance probability) and L_C (reporting honesty probability) across the domains and population segments the controller serves. The measurements are not perfect—trust is

a latent variable, and any measurement of it is subject to error—but they provide the controller with a signal that would otherwise be absent: whether its own gain parameter is rising or falling.

The design requirements for legitimacy sensors follow from the Measurement Paradox of Paper VIII. A controller with declining L has an incentive to suppress or manipulate the very measurements that would reveal the decline. Legitimacy sensors must therefore be structurally protected from that incentive. Specifically:

- **Independence.** The institutions that measure L—trust surveys, compliance monitoring, reporting integrity audits—should be independent of the executive and of the institutions whose legitimacy they measure. Independence includes appointment processes, budgetary autonomy, and legal protection from removal or sanction.
- **Diversity.** Legitimacy sensors should draw on multiple, decorrelated observation channels (Paper X). Official trust surveys should be complemented by independent academic research, civil society monitoring, and the divergence analysis between official and unofficial data sources. A single, government-controlled legitimacy survey is a legitimacy sensor that can be captured; a diversified ensemble of independent sensors is harder to compromise.
- **Frequency.** Legitimacy is a state variable with its own dynamics, and the controller needs to track its evolution in real time. Annual trust surveys are insufficient for gain-scheduling purposes. The controller needs higher-frequency indicators—compliance rates, reporting latency, participation rates in voluntary programmes, the rate at which citizens choose to use formal rather than informal channels—that provide a continuous signal of L's direction and rate of change.
- **Public accessibility.** The measurements of L should be public, not confidential to the controller. This serves two functions. First, it prevents the controller from suppressing unfavourable measurements while acting on them privately—a form of the transparency trap internalised within the measurement system. Second, it enables the governed population to observe the controller's legitimacy trajectory, which itself affects L dynamics: a population that can see that trust is declining across multiple independent measures may update its behaviour more gradually than one that discovers a hidden legitimacy collapse through crisis.

The specific metrics for legitimacy sensing include:

- Tax compliance gap estimates, customs enforcement data, and regulatory compliance rates (L_B).
- Divergence between official statistics and independent estimates—satellite data, market prices, civil society surveys (L_C).
- Trust in government surveys, with disaggregation by institution, domain, and population segment.
- Whistleblower report volumes and the rate at which internal audits detect discrepancies between reported and actual conditions.
- Participation rates in voluntary programmes—vaccination, census response, consultation submissions—as behavioural proxies for willingness to cooperate with the state.

A controller that does not measure its own L is operating open-loop on the parameter that schedules its own effectiveness. The legitimacy sensor is the instrument that closes the loop—not on the governed system, but on the governance system itself.

6.4 Circuit-Breaker Mechanisms

When L breaches a critical threshold—L_crit, the point below which the legitimacy trap becomes self-reinforcing—the controller's normal operating logic becomes counterproductive. Ambitious interventions fail because actuation is depleted; the failures further reduce L. The controller is in a region of the state space where the only sustainable strategy is to reduce ambition, increase transparency, and rebuild L before attempting large-scale action. But the controller in the low-L regime is also subject to political pressure to *act*—to solve the problems that the depleted L prevents it from solving.

The circuit-breaker mechanism is a structural device that interrupts this dynamic. When L falls below a pre-specified threshold, certain classes of decision are automatically paused, transferred, or subjected to heightened scrutiny. The circuit-breaker prevents the controller from destroying its remaining legitimacy through continued action on a corrupted signal.

The circuit-breaker is the governance analogue of a safety valve in an engineering system. When pressure exceeds a safe threshold, the valve opens automatically—not because an operator decides to open it, but because the system is designed to prevent catastrophic failure. In governance, the circuit-breaker removes from the controller the discretion to ignore its own L level. It enforces gain-scheduling by institutional design rather than by the controller's judgment—because the controller's judgment is itself degraded when L is low.

The design properties of an effective circuit-breaker include:

- **Automatic triggers.** The circuit-breaker should activate when L falls below a pre-specified threshold, as measured by independent legitimacy sensors. The trigger should be automatic, not discretionary—the controller should not be able to override it.
- **Domain-specificity.** The circuit-breaker should apply to the specific domains where L has collapsed, not to all governance functions indiscriminately. If L has collapsed for the tax authority but remains adequate for the health system, the circuit-breaker should constrain tax enforcement decisions while leaving health policy unaffected.
- **Proportionality.** The constraints imposed by the circuit-breaker should be proportional to the L depletion. A moderate L decline might trigger mandatory transparency measures and independent review of major decisions. A severe L collapse might transfer decision-making authority to an independent body, a legislative process, or a citizens' assembly until L recovers.
- **Reversibility.** The circuit-breaker should include a clear pathway for deactivation: a sustained improvement in measured L, over a period sufficient to traverse the hysteresis gap, should restore normal governance authority.

- **Insulation from capture.** The institution that determines whether L has breached the threshold must be independent of the controller whose authority the circuit-breaker constrains. This is the same independence requirement that applies to legitimacy sensors, raised to a higher level of consequence.

Historical precedents for circuit-breaker mechanisms are rare but instructive. The independent commission model—where a controversial or legitimacy-depleted policy domain is transferred to a temporary, independent body—is an ad hoc circuit-breaker. The use of referendums for constitutional amendments in some jurisdictions functions as a circuit-breaker on the legislature's authority for specific, high-stakes decisions. The structural requirement is to make such mechanisms systematic, automatic, and integrated into the governance architecture rather than improvised in crisis.

6.5 Credible Commitment and Time Consistency

The hysteresis asymmetry of Section 3.4—$\alpha_{drop} \gg \alpha_{recovery}$—means that rebuilding L requires sustained, consistent performance over extended periods. A single delivery failure can destroy more legitimacy than a single delivery success can rebuild. The structural implication is that legitimacy-sensitive architectures must solve the problem of credible commitment: how can a controller bind itself to a trajectory of transparency and delivery that extends beyond the current political cycle?

Credible commitment is the governance equivalent of time consistency in economic policy. A central bank that announces a low-inflation target must be believed to be willing to accept short-term costs (higher interest rates, slower growth) to achieve the long-term target. If the governed population believes the bank will renege when the costs become politically uncomfortable, the announcement provides no anchoring benefit. Similarly, a controller that announces a legitimacy-rebuilding strategy—modest targets, maximum transparency, consistent delivery—must be believed to be willing to accept the political costs of that strategy over the extended period required to traverse the hysteresis gap. If the population believes the controller will abandon the strategy when it becomes inconvenient, the strategy generates no trust benefit.

The design response is to institutionalise commitments so that defection is costly. Mechanisms include:

- **Constitutional or statutory entrenchment** of transparency obligations, fiscal rules, and audit independence, so that reversing them requires a supermajority or a constitutional amendment rather than a simple change of government.
- **Fixed-term appointments** for the heads of statistical agencies, audit institutions, and legitimacy-sensing bodies, with terms that span multiple electoral cycles and removal only for cause.
- **Pre-committed transparency standards** that specify, in advance, what information will be published, at what frequency, in what format, and with what independent verification. The pre-commitment removes the discretion to reduce transparency when the published information becomes uncomfortable.

- **International agreements** that bind the controller to transparency and delivery standards, with external monitoring and enforcement. The external constraint substitutes for domestic credibility when domestic credibility is depleted—the mechanism through which EU conditionality and IMF programmes sometimes function as legitimacy bridges.
- **Opposition and civil society involvement** in the design and monitoring of legitimacy-rebuilding strategies. A strategy that is co-designed with the political opposition and civil society is harder for the government to abandon unilaterally, because the co-designers have both the standing and the interest to enforce it.

Credible commitment is not a solution to the hysteresis problem. It is a structural tool for making the hysteresis problem manageable. It reduces the controller's ability to defect from the rebuilding trajectory when the political costs of staying the course become high. It does not eliminate the costs; it makes them bearable by ensuring that the benefits of defection are smaller than the costs of the institutional penalties that defection triggers.

6.6 The Legitimacy Substrate for Global Boundary Institutions

Paper XII argued that planetary-scale dynamics—climate change, pandemic transmission, financial stability, artificial intelligence—require global functional boundary institutions: controllers whose jurisdictional perimeter is the planet for the specific purpose of governing those dynamics. This paper adds a critical architectural constraint: a global institution with planetary jurisdiction and zero legitimacy has zero effective actuation, regardless of how perfectly its boundary matches the coupling structure of the governed domain.

The legitimacy substrate for global governance is not a political obstacle that can be wished away. It is the gain parameter that determines whether the global institution can function. A global climate authority that lacks the trust of the populations whose emissions it must constrain will issue directives that are ignored, collect data that is strategically misreported, and face compliance deficits that make its formal authority meaningless. The architecture will be elegant and ineffective.

The design implications follow from the principles already established:

- **Radical transparency.** A global boundary institution with authority over planetary dynamics must operate at a transparency level that exceeds the domestic standard, precisely because it lacks the reservoir of national identity and historical legitimacy that domestic institutions can draw on. Every decision, every piece of evidence, every model assumption must be publicly accessible and independently verifiable. The institution's legitimacy must be continuously earned; it cannot be borrowed from the member states that created it.
- **Subsidiarity of implementation.** The global institution should set the boundary conditions—the emissions constraints, the surveillance standards, the stability requirements—while implementation is carried out by national and sub-national institutions with higher local L. This separation preserves the L

of local institutions for the actuation function while the global institution handles the coordination function that local institutions structurally cannot.

- **Demonstrable delivery before authority expansion.** The global institution should build L through a sequence of modest, delivered successes before seeking expanded authority. A climate institution that successfully monitors and reports emissions with high transparency for a decade will have more L for enforcement than one that seeks enforcement authority from inception. The sequencing mirrors the gain-scheduling logic: start with functions that are achievable with the current L, deliver consistently, and expand as L grows.
- **Diversified legitimacy sources.** The global institution should be accountable to multiple constituencies—states, sub-national governments, civil society organisations, scientific bodies, and directly to citizens through transparent reporting—so that its L is not dependent on any single constituency's continued trust. This is the legitimacy analogue of observer diversity (Paper X): an ensemble of legitimacy sources is more resilient than a single source.
- **Legitimacy sensing at the global scale.** The institution should maintain independent, transparent measurements of its own L across the populations it affects. Trust in the institution should be surveyed regularly, by independent bodies, with public results. A declining trust trend should trigger circuit-breaker mechanisms that constrain the institution's authority expansion until L recovers.

The legitimacy substrate requirement is not a concession to political realism. It is a structural constraint that follows from the same mathematics as every other principle in this paper. The L that multiplies actuation and observation for a domestic controller does the same for a global one. Designing a global governance architecture without designing for its legitimacy substrate is like designing an aircraft without designing for the atmosphere it must fly through. The architecture will be formally correct and functionally inoperative.

The six design principles form an integrated architecture for legitimacy-sensitive governance. Transparency by default maintains the observation channel on which L depends (6.1). Delivery–reality matching prevents the delivery gaps that erode L (6.2). Legitimacy sensors provide the controller with the information needed to adapt its strategy to its own L level (6.3). Circuit-breaker mechanisms prevent the controller from destroying its remaining legitimacy through continued action when L is depleted (6.4). Credible commitment mechanisms enable the extended recovery trajectories that the hysteresis asymmetry requires (6.5). And the legitimacy substrate requirement extends these principles to the global institutions on which planetary governance depends (6.6).

None of these principles is easy to implement. Each confronts the political incentives that make short-term legitimacy borrowing more attractive than long-term legitimacy building. But the alternative is not a different governance strategy. It is the continued operation of the legitimacy trap, the transparency trap, and the high-

suppression fragility that the earlier parts of this paper have diagnosed. The structural logic is clear. The design requirements follow from it. The task of implementation is the work of building institutions that take their own legitimacy seriously as the parameter on which everything else depends.

Part VII — Connection to the Series

This paper is the thirteenth in a sequence that began with the observation that governance systems fail in structurally predictable ways, not because of incompetent institutions but because of architectural choices that place hard constraints on what any institution can achieve. The preceding papers have examined those constraints from multiple angles, using multiple formal frameworks, across multiple domains. This part places the legitimacy dynamics in the context of the series as a whole—clarifying what kind of variable legitimacy is, how it relates to the primitives already established, and where it opens the path forward.

7.1 Legitimacy as the First Endogenous Coupling State

The Governance as Engineering series has, across twelve papers, built a grammar of governance architecture. That grammar identifies structural primitives—properties of the institutional design that the designer can choose and that determine how the control loop performs. Latency, signal fidelity, representation depth, delegation depth, observer diversity, boundary selection—each is an architectural variable. The designer sets them, within the constraints of political feasibility and historical inheritance. They degrade through identifiable structural mechanisms. They can be improved through architectural reform.

Legitimacy is not one of these.

The designer cannot choose L. The designer can choose the architecture that, over time, generates or erodes L. But L itself arises from the interaction between that architecture and the governed population. It is an *emergent* variable—a property of the system's state, not of its design. And it is a *coupling* variable: it links the actuation and observation channels that the rest of the series has treated as separable.

This is why the paper does not add legitimacy as a twelfth primitive. It adds it as the series' first *endogenous coupling state*—a variable that is generated by the system, that feeds back on the system's own effectiveness, and that the controller must therefore observe, model, and respond to. The distinction is not merely terminological. It reflects a structural difference in the kind of thing legitimacy is. A primitive is chosen; a coupling state is managed. A primitive is a design parameter; a coupling state is a control objective. The earlier papers provide the vocabulary for designing architectures. This paper provides the vocabulary for managing the relationship between those architectures and the populations they govern.

The formal move is consistent with the series' trajectory. Papers I through VII are primarily diagnostic: they identify the structural failure modes that arise when primitives are set badly. Papers VIII through XII are increasingly prescriptive: they specify the design principles that follow from the diagnosis. This paper

introduces the first variable that is not itself a design choice, but that determines whether the design choices work. It is the bridge between architecture and outcome—and it is the variable that the earlier papers, by treating observation and actuation as independent channels, could not capture.

7.2 How Architecture Generates Legitimacy

If legitimacy is emergent rather than designed, the natural question is what generates it. The paper's answer is that legitimacy is generated by the architectural primitives themselves, operating over time, in interaction with the governed population.

A governance system with low latency (Paper I), short representation chains (Paper III), high observation dimensionality (Paper VI), protected observer diversity (Paper X), short delegation chains (Paper XI), and well-matched boundaries (Paper XII) will *tend* to generate high L. It delivers outcomes reliably, because its control loop is tight and its actuation is effective. It reports honestly, because its observation channels are diverse, independent, and difficult to manipulate. The governed population learns, over successive interactions, that compliance is rewarded and honesty is safe. Trust accumulates. L rises.

A governance system that violates these primitives will *tend* to generate low L. Its delivery is inconsistent because its control loop is slow and its actuation is attenuated. Its reporting is distorted because its observation channels are narrow and manipulable. The governed population learns that compliance is futile and honesty is dangerous. Trust erodes. L falls.

But the relationship is not deterministic. It is stochastic, path-dependent, and subject to the hysteresis asymmetry that Section 3.4 formalised. A system with improving architecture may face low L for an extended period, because the population's trust was depleted by the preceding period of dysfunction and recovers more slowly than the architecture improves. A system with deteriorating architecture may enjoy high L for a period, because borrowed legitimacy—narrative, charisma, temporary success—sustains trust beyond the point at which the underlying architecture would justify it. The architectural primitives create the *conditions* for legitimacy. They do not guarantee it. And once L is established, it feeds back on the primitives' effectiveness with multiplicative force.

This relationship explains a persistent puzzle in comparative governance: why institutional reforms that should improve outcomes often fail to do so in the short term, and why some systems with formally weak institutions outperform others with formally strong ones. The missing variable is L. A reform that improves the formal architecture but that is implemented in a low-L environment will underperform—not because the reform was poorly designed, but because the parameter that multiplies its effectiveness is depleted. A system with weak formal architecture but high L may outperform a system with strong formal architecture but low L, because the legitimacy multiplier compensates for the architectural deficit. The formal architecture and the legitimacy parameter are both necessary for effective governance. Neither is sufficient alone.

7.3 The Legitimacy-Weighted Information-Actuation Frontier

Paper XII introduced the Information-Actuation Frontier: the structural trade-off between boundary mismatch (B_struct) and delegation depth (Paper XI failure). Expanding boundaries to capture spillovers lengthens observation and actuation chains; shrinking boundaries to preserve internal fidelity leaves structured cross-boundary feedback ungoverned. The frontier is inescapable within any single-boundary architecture.

This paper adds a third dimension to that frontier. Legitimacy multiplies both axes. When L is high, the effective actuation capacity is high for any given delegation depth, and the effective observation capacity is high for any given boundary configuration. The frontier shifts outward: the system can achieve better outcomes for any combination of boundary size and delegation depth. When L is low, the frontier shifts inward. Even a perfectly matched boundary and short delegation chains yield poor actuation and noisy observation, because the parameter that multiplies them is depleted.

This has a critical implication for reform sequencing. A governance system that attempts to optimise its boundary configuration (Paper XII) or its delegation depth (Paper XI) without first addressing its legitimacy deficit is operating on a shifted-inward frontier. The gains from architectural reform will be smaller than expected, because the multiplier that converts architectural improvements into governance outcomes is operating at a fraction of its potential. The system may conclude that the reforms were poorly designed, when in fact the reforms were sound but the L substrate was insufficient to realise them.

The implication is that, in some cases, legitimacy-building must precede or accompany architectural reform rather than follow it. A system in the legitimacy trap—L below L_crit—has severely constrained transition bandwidth (Paper IX). It cannot implement ambitious reforms because it lacks the effective actuation capacity to do so. It must first rebuild L through the strategies described in Part VI—modest targets, maximum transparency, consistent delivery—and then pursue architectural reform from a higher-L baseline. The sequencing is not a political concession. It is a structural requirement that follows from the multiplication of the frontier by L.

This also illuminates a pathway for reform in systems where comprehensive architectural change is politically blocked. A government that cannot redraw boundaries or shorten delegation chains may still be able to improve L through transparency and delivery–reality matching. The resulting L improvement shifts the frontier outward, improving governance outcomes even without formal architectural reform. The strategy is not a substitute for architectural change, but it can create the conditions—the effective capacity, the political trust—that make architectural change possible later.

7.4 Connection to Transition Bandwidth

Paper IX defined transition bandwidth as the rate at which a governance architecture can peacefully redesign its own structure. It argued that the binding constraint on governance adaptation is not the absence of design solutions but the limited institutional capacity to implement them against incumbent resistance.

This paper adds a legitimacy constraint to that analysis. Rebuilding L takes time—specifically, the extended period required to traverse the hysteresis gap from L_low to L_high. During that period, the controller's effective actuation capacity is depleted, and its transition bandwidth is correspondingly reduced. The system cannot redesign its architecture rapidly because it lacks the legitimacy to implement the redesign. The legitimacy trap is thus a transition-bandwidth trap: a condition in which the very parameter required for reform is depleted by the conditions that make reform necessary.

The interaction between legitimacy and transition bandwidth has an important dynamic consequence. A system that has experienced a legitimacy collapse—through a transparency trap, a borrowed-legitimacy crisis, or a performance-legitimacy spiral—faces a compound constraint. It needs to reform the architecture that produced the collapse. But the collapse has depleted the L required to implement reform. The system is in a double bind: it cannot reform without L, and it cannot rebuild L without demonstrating the delivery that reform would enable.

The resolution, as Paper IX argued for transition constraints more generally, lies in protected experimental spaces. A government with depleted national L may still be able to build L in a limited domain—a municipal laboratory, a sandbox state, a functionally specific institution—where the delivery gap is small enough to be closed with the available capacity. The local L building can then generate positive contagion (Section 3.5), gradually expanding the legitimacy substrate on which broader reform depends. This is the structural logic behind the convergent first step identified across the series' country studies: start small, deliver visibly, build trust, then expand. It is not merely politically pragmatic. It is mathematically required by the interaction of L dynamics and transition bandwidth.

7.5 The Cycle Two Arc

The series' second cycle has traced a specific trajectory, and this paper completes its foundational arc.

Paper XI asked: can the controller implement its intent? The answer was that delegation depth attenuates actuation, and that beyond a critical depth, the control energy required to realise policy intent becomes prohibitive. The paper treated the actuation channel as a structural primitive, subject to degradation through the same mechanisms—projection, noise, latency—that the earlier papers had diagnosed for observation.

Paper XII asked: is the controller acting on the right system? The answer was that boundary mismatch between the real plant and the modelled plant generates unmodeled dynamics that destabilise the controller regardless of internal competence. The paper treated the boundary as a design variable, to be matched to the coupling structure of the governed domain.

This paper asks: will the system cooperate with the controller's actions? The answer is that legitimacy—the willingness of the governed to comply with directives and report honestly—is an emergent coupling state that simultaneously modulates actuation and observation, and that can trap a controller in a low-performance, low-trust attractor from which recovery is possible but slow.

The sequence—actuation, boundary, legitimacy—is the progression from architecture through context to emergent state. Paper XI addresses the controller's internal capacity to act. Paper XII addresses the controller's relationship to the external system it governs. Paper XIII addresses the controller's relationship to the population whose cooperation makes action possible. The three together complete the picture of what a controller must manage: its own actuation chain, its jurisdictional perimeter, and the trust that multiplies both.

The arc is not an accident. It reflects the logic of moving from the things the designer can choose—the architectural primitives—to the things the designer can only influence—the emergent coupling states—to the strategies for managing the interaction between them. The earlier papers provide the vocabulary for designing institutions. The later papers provide the vocabulary for governing the relationship between institutions and the societies they serve. This paper is the pivot between those two concerns: it is about the variable that is not chosen but that determines whether the choices work.

The series began with a simple observation: governance systems fail in predictable ways, not because leaders lack wisdom or institutions lack resources, but because the underlying architecture generates failure as a structural output. Thirteen papers later, that observation has been extended from the architecture of institutions to the dynamics of the trust that sustains them. The structural primitives of Papers I through XII are necessary. They are not sufficient. They operate within a legitimacy field that multiplies or nullifies their effects. The controller that understands this will design not only for latency, fidelity, and boundary, but for the parameter that determines whether any of them matter. The controller that does not will eventually discover that it has built an elegant machine that no one is willing to operate.

Part VIII — Limitations and Conclusion

8.1 Limitations

The argument of this paper is structural: legitimacy is an emergent coupling state that simultaneously modulates actuation and observation, and a governance architecture that ignores this parameter will eventually be destroyed by the dynamics it generates. This argument has been developed through a formal framework, a simulation, and empirical illustrations. It is subject to limitations that should be stated clearly.

The scalar L assumption. The paper models legitimacy as a composite scalar L(t), acknowledging in Section 2.1 that actuation-legitimacy L_B and observation-legitimacy L_C are conceptually distinct. In reality, legitimacy is multidimensional. A population may trust the courts but not the legislature, the central bank but not the executive, the health system but not the tax authority. A comprehensive treatment would model L as a vector or matrix, with different institutions commanding different legitimacy levels and with cross-institutional spillovers captured by a coupling matrix. The scalar model captures the essential dynamics—the coupling of actuation and observation, the trap, the hysteresis—but cannot capture the distributional structure of legitimacy collapses or the targeted rebuilding strategies that the multidimensional extension would enable.

Phenomenological L dynamics. The update equation for L in Section 2.2 is phenomenological. It specifies how L responds to delivery gaps, transparency, and betrayal, but it does not derive these responses from a micro-founded model of individual behaviour. The parameters α, β, and γ are estimated from empirical regularities rather than deduced from first principles. A full micro-foundation would require modelling the beliefs, expectations, and strategic interactions of the governed population—an agent-based extension that is beyond the scope of this paper but that would strengthen the connection between the structural dynamics and the individual behaviour that generates them.

The bracketing of normative questions. The paper deliberately sets aside the question of when a government *deserves* legitimacy. It treats L as an empirical parameter—the probability of compliance and honest reporting—and analyses its structural consequences regardless of its normative basis. This is an analytical strength in that it allows the paper to make claims that do not depend on resolving millennia of political philosophy. It is a limitation in that the paper cannot address the normative conditions under which L *should* be rebuilt. A regime that governs through fear, clientelism, or propaganda may maintain high L without satisfying any democratic criterion of legitimacy. The paper's framework identifies the structural consequences of high or low L; it does not provide a normative theory of which L is justified.

Empirical illustrations, not validations. The five cases in Part V demonstrate that the paper's diagnostic framework is legible in real governance systems and that the mechanisms it identifies—the transparency trap, the hysteresis gap, the borrowed-legitimacy collapse—have recognisable historical manifestations. The cases are not formal validations. They were selected because they exhibit the dynamics clearly, not through a systematic sampling of governance domains. Confirming that L, as operationalised through the legitimacy sensors of Section 6.3, predicts governance outcomes across a representative sample of systems and periods requires an empirical programme that has not yet been conducted.

The simulation is illustrative. The simulation of Part IV demonstrates that the qualitative dynamics of the framework emerge reliably from a minimal set of assumptions. The parameters are chosen to make the mechanisms visible, not to calibrate against any specific real-world system. The quantitative results—the specific location of L_crit, the specific duration of recovery trajectories, the specific magnitude of borrowed-legitimacy collapses—are artefacts of the parameter choices. The qualitative claim—that the trap, the hysteresis, and the transparency-trap collapse are structural features of any system with coupled actuation-observation legitimacy—is the simulation's contribution, and it is robust to parameter variation.

The political difficulty of implementation. The design principles of Part VI are structurally sound within the framework. They are also politically demanding. Transparency by default confronts the incentives of every government to manage information strategically. Delivery–reality matching confronts the incentives of every politician to promise more than can be delivered. Circuit-breaker mechanisms confront the reluctance of every executive to cede authority to automatic triggers. The paper provides the structural argument for these principles; it does not provide a political theory of how to implement them against resistance. The transition bandwidth problem of Paper IX applies with full force to the implementation of legitimacy-sensitive design.

The relationship between L and performance is bidirectional. The paper emphasises the direction from L to performance: L multiplies actuation and observation, so changes in L drive changes in outcomes. But the reverse direction is equally present in the model: changes in outcomes drive changes in L through the delivery gap. The system is a coupled nonlinear oscillator, and the paper's analysis of its dynamics—the trap, the bifurcation, the hysteresis—is necessarily simplified. A full nonlinear analysis, including the possibility of limit cycles, chaotic transients, and multiple coexisting attractors, would require a more extensive mathematical treatment.

These limitations are substantial, but they are also specific and bounded. The paper does not claim to have solved the problem of legitimacy. It claims to have identified it as a structural variable of the governance control loop, to have formalised its dynamics and its coupling to the observation and actuation channels, and to have derived design principles that any governance architecture must satisfy to manage it. Those claims survive the limitations acknowledged here.

8.2 Conclusion

This paper began with a simple observation: two governments can have identical formal architectures and radically different outcomes, because the willingness of the governed to comply with directives and report honestly is a parameter that multiplies the effectiveness of every architectural choice. That parameter is legitimacy. And it is not a soft concept, a political nicety, or a normative luxury. It is a gain parameter with its own dynamics, its own traps, and its own structural requirements.

The central insight of the paper is that legitimacy couples the two channels of governance that the series has so far treated separately. When L is high, the actuation matrix is full-rank and the observation channel is clear. When L collapses, the same institutional architecture becomes unsteerable and blind. The collapse is not a separate event from the architectural failures the series has already diagnosed. It is the mechanism through which those failures are amplified, accelerated, and made self-reinforcing by the governed population's loss of trust.

The paper has formalised this coupling through the legitimacy-weighted state-space model of Part II, in which the effective actuation matrix is $\mathbf{B}\_\text{eff} = L \cdot \mathbf{B}$ and the observation noise covariance is $\mathbf{V} = \mathbf{V}_0 / L$. It has derived the legitimacy dynamics—the delivery gap, the transparency signal, the betrayal cost—that govern L's evolution. It has identified the legitimacy trap, the condition in which falling L and deteriorating performance reinforce each other in a positive-feedback spiral that terminates in a low-L attractor from which recovery is slow. It has distinguished between built legitimacy, which is stable and resilient, and borrowed legitimacy, which is acquired quickly and collapses catastrophically. And it has derived, from these dynamics, a set of design principles that any governance architecture must satisfy if it is to maintain the parameter on which its own effectiveness depends.

The design principles are demanding. They require transparency by default, even when transparency reveals uncomfortable truths. They require delivery–reality matching, even when political incentives reward over-promising. They require legitimacy sensors that are independent, diversified, and public, even though a controller with declining L has every incentive to suppress them. They require circuit-breaker mechanisms that automatically constrain the controller's authority when L falls below a critical threshold. They require credible commitment mechanisms that bind the controller to a rebuilding trajectory over timescales that exceed the political cycle. And they require that global governance institutions, if they are to function at all, be built on a legitimacy substrate that is earned through transparency and delivery rather than borrowed from the member states that created them.

None of this is easy. Each principle confronts the political incentives that make short-term legitimacy borrowing more attractive than long-term legitimacy building. But the alternative is not stability. It is the continued operation of the legitimacy trap, the transparency trap, and the high-suppression fragility that Part III diagnosed and that the empirical illustrations of Part V documented in cases from Greek sovereign debt to Chinese pandemic policy.

The paper's most important claim is not that legitimacy matters—that has been said many times before, in many vocabularies. It is that legitimacy has a specific, analysable structure, and that structure has specific, designable implications. The paper provides the formal grammar for that structure and the design vocabulary for those implications. It treats legitimacy not as a diffuse political atmosphere but as an endogenous state variable with defined dynamics, measurable parameters, and predictable failure modes. It argues that a controller who understands this structure can design for it, and that a controller who does not will be destroyed by it.

The series has now completed the foundational arc of its second cycle. Paper XI addressed the controller's capacity to act. Paper XII addressed whether the controller is acting on the right system. This paper addresses whether the system will cooperate with the controller's actions. The sequence—actuation, boundary, legitimacy—is the progression from architecture through context to the emergent state that determines whether either functions.

What remains is to build the institutions that satisfy these requirements—and, before building, to test the predictions on which the design rests. The empirical programme is specified. The measurement framework exists in prototype. The theoretical architecture is complete. The next phase is not more theory. It is confrontation with data, and with the political task of implementing legitimacy-sensitive design in institutions that have been designed as if legitimacy were exogenous.

The trust of the governed is not a gift. It is a state variable, generated by the interaction between architecture and society, evolving under its own nonlinear dynamics, and feeding back on the effectiveness of every governance action. A controller that treats it as a constant will eventually discover that it is the most important variable in the loop. A controller that treats it as a design objective—to be observed, protected, rebuilt, and never borrowed against at a cost it cannot repay—will find that it is the foundation on which every other architectural choice rests. The engineering of governance is not complete until it accounts for the parameter that determines whether the machine will run. That parameter is legitimacy. And it is time governance architecture took it seriously.

Appendix A — Formal Derivations

This appendix provides the mathematical derivations underlying the legitimacy-weighted state-space model of Part II. It defines the coupled actuation–observation system, derives the Kalman filter degradation under falling observation legitimacy, formalises the legitimacy dynamics with hysteresis asymmetry and the split-state transparency mechanism, and characterises the legitimacy trap as a sector-bounded nonlinearity in the control loop.

A.1 Legitimacy-Weighted State-Space Model

The baseline discrete-time linear system is

$$\begin{aligned} x(t+1) &= A\,x(t) + B\,u(t) + w(t), \quad w(t) \sim N(0, W), \\ y(t) &= C\,x(t) + v(t), \quad v(t) \sim N(0, V_0), \end{aligned}$$

x(t + 1)y(t) = A x(t) + B u(t) + w(t), w(t) ∼ N (0, W), = C x(t) + v(t), v(t) ∼ N (0, V0),

where $x(t) \in R^n$x(t) ∈ Rn is the true state, $u(t) \in R^m$u(t) ∈ Rm the control input, $y(t) \in R^p$y(t) ∈ Rp the measurement, and A, B, CA, B, C are the nominal dynamics, actuation, and observation matrices. The noise covariances WW and V_0V0 represent irreducible process and measurement uncertainty under perfect legitimacy.

Legitimacy-dependent channels.
Legitimacy is modelled as two scalar parameters $L_B(t), L_C(t) \in [0, 1]$LB(t), LC(t) ∈ [0, 1]. They modify the actuation and observation channels:

$$B_{\text{eff}}(t) = L_B(t)\,B, \qquad V(t) = \frac{V_0}{L_C(t)}.$$

Beff(t) = LB(t) B, V(t) = LC(t)V0.

Thus the effective system available to the controller is

$$\begin{aligned} x(t+1) &= A\,x(t) + L_B(t)\,B\,u(t) + w(t), \\ y(t) &= C\,x(t) + v(t), \quad v(t) \sim N(0, V_0/L_C(t)). \end{aligned}$$

x(t + 1)y(t) = A x(t) + LB(t) B u(t) + w(t), = C x(t) + v(t), v(t) ∼ N (0, V0/LC(t)).

When $L_B = L_C = 1$LB = LC = 1 the controller operates with its full designed authority and sensing precision. As either parameter falls, actuation is weakened and measurement noise is amplified. The two channels are coupled through the common dependence of L_BLB and L_CLC on the underlying trust state; in the scalar-LL approximation used in the main text we set $L_B = L_C = L(t)$LB = LC = L(t).

A.2 Kalman Filter Degradation Under Failing Observation Legitimacy

A well-designed controller does not use raw measurements directly but filters them through a state estimator. Under Gaussian noise and linear dynamics the optimal estimator is the Kalman filter. The filter propagates a state estimate $\hat{x}(t)$x^(t) and an error covariance $P(t)$P(t) via two steps:

Prediction.

$$\hat{x}(t|t-1) = A\,\hat{x}(t-1) + L_B(t-1)\,B\,u(t-1),$$

x^(t∣t − 1) = A x^(t − 1) + LB(t − 1) B u(t − 1),

$$P(t|t-1) = A\,P(t-1)\,A^{\top} + W.$$

P(t∣t − 1) = A P(t − 1) A⊤ + W.

Update. On receipt of $y(t)$y(t),

$$K(t) = P(t|t-1)\,C^{\top}\,\big(\,C\,P(t|t-1)\,C^{\top} + V(t)\,\big)^{-1},$$

K(t) = P(t∣t − 1) C⊤ (C P(t∣t − 1) C⊤ + V(t))−1,

$$\hat{x}(t) = \hat{x}(t|t-1) + K(t)\,\big(\,y(t) - C\,\hat{x}(t|t-1)\,\big),$$

x^(t) = x^(t∣t − 1) + K(t)(y(t) − C x^(t∣t − 1)),

$$P(t) = \big(\,I - K(t)\,C\,\big)\;P(t|t-1).$$

P(t) = (I − K(t) C) P(t∣t − 1).

The Kalman gain $K(t)$K(t) determines how much weight the filter gives to the new measurement relative to the model-based prediction. It depends inversely on the measurement noise covariance $V(t) = V_0/L_C(t)$ V(t) = V0/LC(t).

Limiting behaviour as $L_C \to 0$LC → 0.

As observation legitimacy decays, $L_C(t) \to 0$LC(t) → 0 and $V(t) \to \infty$V(t) → ∞ (its eigenvalues diverge). The innovation covariance $S(t) = CP(t|t-1)C^{\top} + V(t)$S(t) = CP(t∣t − 1)C⊤ + V(t) becomes dominated by $V(t)$V(t), so

$$\lim_{L_C \to 0} K(t) = \lim_{V \to \infty} P(t|t-1)\,C^{\top}\,\big(\,CP(t|t-1)C^{\top} + V\,\big)^{-1} = 0.$$

LC → 0lim K(t) = V → ∞lim P(t∣t − 1) C⊤ (CP(t∣t − 1)C⊤ + V)−1 = 0.

The Kalman gain vanishes. The update step then reduces to

$$\hat{x}(t) = \hat{x}(t|t-1) = A\,\hat{x}(t-1) + L_B(t-1)\,B\,u(t-1),$$
x^(t) = x^(t∣t − 1) = A x^(t − 1) + LB(t − 1) B u(t − 1),

which is the open-loop propagation of the internal model. The filter ignores all incoming measurements. The controller's estimate of the system state is driven entirely by its prior model AA and its own past commands, uncorrected by reality. This is the formal mechanism of *dashboard insulation*: a collapse of observation legitimacy forces the controller to operate blind, no matter how sophisticated its internal model may be.

A.3 Legitimacy Dynamics with Hysteresis Asymmetry

Legitimacy evolves in response to the controller's performance and transparency. The core update equation for the composite scalar $L(t)$L(t) is

$$L(t+1) = \text{clip}\,\Big(\,L(t) - \alpha(t)\,\|x_{\text{rep}}(t)\|^2 + \beta\,T(t) - \gamma\,D(t) + \delta,\;0,\;1\,\Big), \qquad \text{(A.1)}$$
L(t + 1) = clip(L(t) − α(t) ‖xrep(t)‖2 + β T(t) − γ D(t) + δ, 0, 1),(A.1)

where

- $x_{\text{rep}}(t)$xrep(t) is the state as perceived by the governed population (see below),
- $T(t) \in [0,1]$T(t) ∈ [0, 1] is the controller's transparency level,
- $D(t) \in \{0,1\}$D(t) ∈ {0, 1} indicates a deception revelation event,
- $\delta > 0$δ > 0 is a small exogenous drift capturing slow, secular accumulation of institutional trust,
- $\beta > 0$β > 0 and $\gamma > 0$γ > 0 are sensitivity parameters.

Hysteresis-asymmetric delivery sensitivity.
The parameter $\alpha(t)$α(t) is not constant. It takes different values depending on whether performance is improving or worsening:

$$\alpha(t) = \begin{cases} \alpha_{\text{drop}}, & \text{if } \|x_{\text{rep}}(t)\|^2 > \|x_{\text{rep}}(t-1)\|^2, \\ \alpha_{\text{recovery}}, & \text{if } \|x_{\text{rep}}(t)\|^2 \le \|x_{\text{rep}}(t-1)\|^2, \end{cases} \qquad \alpha_{\text{drop}} \gg \alpha_{\text{recovery}} > 0. \qquad \text{(A.2)}$$
α(t) = ⎩ ⎨ ⎧αdrop,αrecovery,if ‖xrep(t)‖2 > ‖xrep(t − 1)‖2,if ‖xrep(t)‖2 ≤ ‖xrep(t − 1)‖2, αdrop ≫ αrecovery > 0.(A.2)

This piecewise definition captures the empirical regularity that trust is lost more quickly than it is rebuilt. A worsening delivery gap (positive change in squared error) produces a strong negative update; an improving delivery gap produces only a weak positive update. The hysteresis loop in the main text follows directly from this asymmetry.

Split-state transparency and the betrayal mechanism.
When the controller suppresses information, the reported state seen by the public diverges from the true state. We model this by a suppression parameter $\lambda \in [0,1]$λ ∈ [0, 1] and a promised reference state x_{promised} xpromised (typically the target 00):

$$x_{\text{rep}}(t) = \lambda\, x(t) + (1-\lambda)\, x_{\text{promised}}(t). \qquad (A.3)$$

xrep(t) = λ x(t) + (1 − λ) xpromised(t).(A.3)

Full transparency corresponds to $\lambda = 1$λ = 1 ($x_{\text{rep}} = x$xrep = x); complete fabrication to $\lambda = 0$λ = 0.

The hidden discrepancy between true and reported states accumulates in a variable

$$E_{\text{betrayal}}(t+1) = E_{\text{betrayal}}(t) + \|x(t) - x_{\text{rep}}(t)\|^2, \qquad (A.4)$$

Ebetrayal(t + 1) = Ebetrayal(t) + ‖x(t) − xrep(t)‖2,(A.4)

with $E_{\text{betrayal}}(0) = 0$Ebetrayal(0) = 0. The probability that the deception is revealed at time tt is modelled as a hazard rate that increases with E_{betrayal}Ebetrayal:

$$\Pr(\text{revelation at } t) = 1 - \exp\left(-h\, E_{\text{betrayal}}(t)\right), \qquad (A.5)$$

Pr(revelation at t) = 1 − exp(−h Ebetrayal(t)),(A.5)

where $h > 0$h > 0 is the hazard coefficient. On revelation, $D(t)$D(t) is set to 11 and λλ is forced to 11 thereafter (the controller can no longer hide). The betrayal penalty $-\gamma D(t)$−γD(t) then strikes the legitimacy update (A.1) with full force, producing the catastrophic collapse analysed in the text.

A.4 The Legitimacy Trap as a Sector-Bounded Nonlinearity

When the controller uses a linear state-feedback law $u(t) = -K\,\hat{x}(t)$u(t) = −K x^(t), the complete system—plant, estimator, controller, and legitimacy dynamics—constitutes a nonlinear feedback loop. The legitimacy state $L(t)$L(t) enters as a *state-dependent gain* that multiplies BB and scales the measurement noise. Moreover, $L(t)$L(t) itself evolves according to (A.1), which is a memoryless nonlinear function of the recent state trajectory and the controller's transparency. This structure is precisely that of a **Lur'e system**—a linear time-invariant forward path with a nonlinear, sector-bounded feedback element.

Circle Criterion condition.
Consider the simplified case where the forward path (the linear dynamics with constant $L = \bar{L}$L = L¯) is stable and the legitimacy update is approximated as a static nonlinearity $\phi(\cdot)$ϕ(·) acting on the delivery gap $e(t) = \|x_{\text{rep}}(t)\|^2$e(t) = ‖xrep(t)‖2. The update (A.1) can be written as

$$\Delta L(t) = -\phi(e(t)) + \beta T(t) - \gamma D(t) + \delta,$$

ΔL(t) = −ϕ(e(t)) + βT(t) − γD(t) + δ,

where $\phi(\cdot)$ϕ(·) is piecewise linear with slopes determined by α_{drop}αdrop and α_{recovery}αrecovery. The nonlinearity ϕϕ satisfies a sector condition: there exist constants k_1, k_2k1, k2 such that

$$k_1 e \le \phi(e) \le k_2 e \quad \text{for all } e \ge 0.$$

k1e ≤ ϕ(e) ≤ k2e for all e ≥ 0.

In our case, $k_1 = \alpha_{\text{recovery}}$k1 = αrecovery and $k_2 = \alpha_{\text{drop}}$k2 = αdrop (suitably scaled). The Circle Criterion provides a sufficient condition for absolute stability of the closed loop: if the Nyquist plot of the linear part (the transfer function from the legitimacy gain perturbation to the delivery gap) does not intersect or encircle a specific disk determined by k_1, k_2k1, k2, then the system is stable for any time-varying gain in that sector.

When the sector bounds are narrow—i.e. the asymmetry $\alpha_{\text{drop}}/\alpha_{\text{recovery}}$αdrop/αrecovery is small and γγ is moderate—the stability disk is large, and the condition is easily satisfied: the system is **absolutely stable**, and the legitimacy dynamics cannot drive it to a low-LL attractor from any initial condition. This corresponds to a *built*-legitimacy regime.

When the sector bounds are wide—large α_{drop}αdrop relative to α_{recovery}αrecovery, and large γγ—the stability disk shrinks. The Circle Criterion may be violated, meaning there exist gain trajectories (legitimacy paths) that destabilise the loop. In that case the system is only **conditionally stable**: a sufficiently large perturbation that drives LL below a critical value L_{crit}Lcrit will cause the loop to diverge from the high-LL equilibrium and enter the legitimacy trap. This corresponds to a *borrowed*-legitimacy regime, where the high sensitivity to delivery failures and the catastrophic betrayal penalty make the system vulnerable to a self-reinforcing collapse.

The locus of L_{crit}Lcrit is not a universal constant but depends on the specific parameters $\alpha_{\text{drop}}, \alpha_{\text{recovery}}, \beta, \gamma, \delta$αdrop, αrecovery, β, γ, δ and on the dynamic characteristics of the plant (the eigenvalues of AA). In the simulation of Part IV, L_{crit}Lcrit is identified numerically as the separatrix of the basins of attraction.

A.5 Built vs. Borrowed Legitimacy: Parameter Sets

The distinction between built and borrowed legitimacy is operationalised through distinct parameter regimes in the update equation (A.1) and the hazard model (A.5).

Parameter	Built legitimacy	Borrowed legitimacy
α_{drop}αdrop	moderate (e.g. 0.12)	high (e.g. 0.25)
α_{recovery}αrecovery	moderate (e.g. 0.06)	low (e.g. 0.02)
ββ	moderate (e.g. 0.08)	low (e.g. 0.03)
γγ	low (e.g. 0.5)	high (e.g. 3.0)
δδ	high (e.g. 0.005)	low (e.g. 0.001)
Hazard coefficient hh	low (deception is harder to sustain)	high (deception is more likely to be revealed, but the regime is more likely to attempt it)

Structural interpretation.
Built legitimacy is characterised by a damped response to delivery gaps, substantial responsiveness to transparency, a small betrayal penalty (because trust is based on a long track record of honesty), and a slow exogenous decay rate. The sector bounds in the associated Lur'e system are narrow, satisfying the Circle Criterion: the high-LL equilibrium is absolutely stable.

Borrowed legitimacy is characterised by a hyper-sensitive response to delivery failures, weak responsiveness to transparency, a catastrophic betrayal penalty (because trust is narrative-based and fragile), and a rapid exogenous decay when the narrative weakens. The sector bounds are wide, violating the Circle Criterion: the system is only conditionally stable, and a sufficiently large shock can push it into the trap.

These parameter sets are not independent. A regime that relies on borrowed legitimacy will tend to suppress transparency (low ββ), which forces it to rely even more heavily on the narrative, making it exquisitely sensitive to any delivery failure that breaches that narrative (high α_{drop}αdrop, high γγ). The parameter regime is self-reinforcing until the collapse occurs—exactly the dynamic that the transparency trap formalises.

Appendix B — Simulation Specification

This appendix provides the detailed specification for the simulation described in Part IV. It defines the system dynamics, the legitimacy update mechanism with split-state transparency and hysteresis asymmetry, the four scenarios, the parameter sweeps, and the output metrics. The specification is sufficient to implement the simulation independently.

B.1 Model Specification

The simulated governance system controls a two-dimensional state vector $x(t) = [x_1(t), x_2(t)]^\top \in R^2$x(t) = [x1(t), x2(t)]⊤ ∈ R2, representing two policy-relevant dimensions such as economic output and environmental quality, or service delivery and fiscal balance. The dynamics are linear time-invariant, with legitimacy entering as a multiplicative gain on actuation and an inverse divisor on observation noise.

True dynamics.

$$x(t+1) = A\,x(t) + L_B(t)\,B\,u(t) + w(t), \qquad w(t) \sim N(0, W),$$

x(t + 1) = A x(t) + LB(t) B u(t) + w(t), w(t) ∼ N(0, W), where

$$A = 0.95\,I_2, \quad B = I_2, \quad W = 0.01\,I_2.$$

A = 0.95 I2, B = I2, W = 0.01 I2. The uncontrolled system decays slowly toward the origin (the target state). Actuation effectiveness is multiplied by the actuation-legitimacy parameter $L_B(t) \in [0, 1]$LB(t) ∈ [0, 1].

Observation.

$$y(t) = C\,x(t) + v(t), \qquad v(t) \sim N\left(0, V(t)\right),$$

y(t) = C x(t) + v(t), v(t) ∼ N(0, V(t)), with $C = I_2$C = I2 (full-state observation up to noise) and

$$V(t) = \frac{V_0}{L_C(t)}, \qquad V_0 = 0.05\,I_2.$$

V(t) = LC(t)V0, V0 = 0.05 I2. The observation-legitimacy parameter $L_C(t) \in [0, 1]$LC(t) ∈ [0, 1] scales the measurement noise covariance inversely: as L_CLC falls, noise rises without bound.

State estimation.

The controller maintains a state estimate $\widehat{x}(t)$x^(t) via a Kalman filter as specified in Appendix A.2. The filter is provided with the true legitimacy values $L_B(t), L_C(t)$LB(t), LC(t) (the controller knows its own current legitimacy level; the simulation does not model misperception of L by the controller, though that would be a natural extension).

Control law.
The controller applies proportional state feedback based on the filtered estimate:

$$u(t) = -K\,\widehat{x}(t),$$

u(t) = −K x^(t), where KK is the infinite-horizon linear quadratic regulator (LQR) gain computed for the nominal design system (A, B, Q, R)(A, B, Q, R) with state cost $Q = I_2$Q = I2 and control cost $R = 0.1\,I_2$ R = 0.1 I2. Solving the discrete algebraic Riccati equation yields $K \approx 0.75\,I_2$K ≈ 0.75 I2 for the chosen parameters. This gain is optimal when $L_B = 1$LB = 1 and is used regardless of the current legitimacy level, so that performance degradation reflects the legitimacy multiplier rather than controller detuning. In Scenario 3 the controller is permitted to reduce its gain below the nominal value to simulate gain-scheduling.

B.2 Legitimacy Dynamics

The composite legitimacy $L(t)$L(t) is modelled as a scalar with $L_B(t) = L_C(t) = L(t)$LB(t) = LC(t) = L(t). Its evolution follows the update equation of Appendix A.3, implemented with hysteresis asymmetry, a split-state transparency mechanism, and a stochastic betrayal hazard.

Legitimacy update.

$$L(t+1) = \text{clip}\,\Big(\,L(t) - \alpha(t)\,\|x_{\text{rep}}(t)\|^2 + \beta\,T(t) - \gamma\,D(t) + \delta,\;0,\;1\,\Big),$$

L(t + 1) = clip(L(t) − α(t) ‖xrep(t)‖2 + β T(t) − γ D(t) + δ, 0, 1), where

- $x_{\text{rep}}(t)$xrep(t) is the state perceived by the governed population,
- $T(t) \in [0,1]$T(t) ∈ [0, 1] is the controller's chosen transparency level,
- $D(t) \in \{0,1\}$D(t) ∈ {0, 1} indicates a deception revelation event,
- $\delta = 0.005$δ = 0.005 is a small exogenous trust drift,
- $\beta = 0.08$β = 0.08 (transparency sensitivity) in all scenarios unless otherwise specified.

Hysteresis-asymmetric delivery sensitivity.

$$\alpha(t) = \begin{cases} \alpha_{\text{drop}} = 0.12, & \text{if } \|x_{\text{rep}}(t)\|^2 > \|x_{\text{rep}}(t-1)\|^2, \\ \alpha_{\text{recovery}} = 0.03, & \text{if } \|x_{\text{rep}}(t)\|^2 \le \|x_{\text{rep}}(t-1)\|^2, \end{cases}$$

α(t) = {αdrop = 0.12,αrecovery = 0.03,if ‖xrep(t)‖2 > ‖xrep(t − 1)‖2,if ‖xrep(t)‖2 ≤ ‖xrep(t − 1)‖2,giving a

4:1 drop-to-recovery ratio.

Split-state transparency.
When the controller chooses suppression, the reported state is a convex combination of the true state and the promised reference $x_{\text{promised}} = 0$xpromised = 0 (the target):

$$x_{\text{rep}}(t) = \lambda\, x(t) + (1-\lambda)\, 0 = \lambda\, x(t),$$

xrep(t) = λ x(t) + (1 − λ) 0 = λ x(t), with suppression parameter $\lambda \in [0,1]$λ ∈ [0, 1]. $\lambda = 1$λ = 1 is full transparency; $\lambda < 1$λ < 1 flatters the true state toward zero (the target) in the eyes of the public.

Hidden discrepancy and betrayal hazard.
The cumulative hidden discrepancy evolves as

$$E_{\text{betrayal}}(t+1) = E_{\text{betrayal}}(t) + \|x(t) - x_{\text{rep}}(t)\|^2,$$

Ebetrayal(t + 1) = Ebetrayal(t) + ‖x(t) − xrep(t)‖2, with $E_{\text{betrayal}}(0) = 0$Ebetrayal(0) = 0. The probability of revelation at time tt is

$$\Pr(\text{revelation at } t) = 1 - \exp\left(-h\, E_{\text{betrayal}}(t)\right),$$

Pr(revelation at t) = 1 − exp(−h Ebetrayal(t)), with hazard coefficient $h = 0.02$h = 0.02. At each time step a uniform random draw determines whether revelation occurs. On revelation, $D(t)$D(t) is set to 11 for that step, λλ is forced to 11 permanently (the controller can no longer hide), and the betrayal penalty $-\gamma D(t)$−γD(t) is applied. The betrayal sensitivity γγ is set according to the legitimacy regime:

- Built-legitimacy scenarios: $\gamma_{\text{built}} = 0.5$γbuilt = 0.5.
- Borrowed-legitimacy scenarios: $\gamma_{\text{borrowed}} = 3.0$γborrowed = 3.0.

Transparency level $T(t)$T (t).
In the simulation, transparency $T(t)$T (t) is not continuously optimised but set as a scenario parameter. In high-transparency scenarios, $T = 1$T = 1 and $\lambda = 1$λ = 1; in suppression scenarios, $T = 0.2$T = 0.2 and $\lambda = 0.3$λ = 0.3 (typical borrowed-legitimacy values).

B.3 Scenarios

Four scenarios are simulated, corresponding to the failure modes of Part III. All use the same plant dynamics and LQR controller (unless otherwise noted for the recovery scenario). Scenarios 2–4 are run with built- or borrowed-legitimacy parameter sets as indicated.

Scenario 1 — High-transparency, high-legitimacy equilibrium.
$L(0) = 0.7$L(0) = 0.7, $T = 1$T = 1, $\lambda = 1$λ = 1 (no suppression). No deception is active. The system experiences a moderate external shock at $t = 50$t = 50: a temporary displacement $x(50) \leftarrow x(50) + [1.5, 0]^\top$ x(50) ← x(50) + [1.5, 0]⊤. The scenario demonstrates shock absorption in a high-LL, transparent regime. Parameters: built-legitimacy set.

Scenario 2 — The legitimacy trap.
$L(0) = 0.7$L(0) = 0.7, $T = 1$T = 1, $\lambda = 1$λ = 1. At $t = 50$t = 50 a large external shock is applied: $x(50) \leftarrow x(50) + [3.0, 0]^\top$ x(50) ← x(50) + [3.0, 0]⊤. The resulting delivery gap is substantial, triggering the

asymmetric α_{drop}αdrop. The controller continues to apply the nominal LQR gain. The simulation demonstrates the self-reinforcing spiral as falling LL degrades actuation and observation, preventing recovery. Parameters: built-legitimacy set, then also borrowed set for comparison.

Scenario 3 — Recovery through transparency intervention.
$L(0) = 0.3$L(0) = 0.3 (the system begins in the low-LL attractor, either exogenously or as the end state of Scenario 2). At $t = 50$t = 50 the controller switches to a legitimacy-rebuilding strategy:

- Gain is halved: $K_{\text{rebuild}} = 0.5\,K$Krebuild = 0.5 K.
- Full transparency is adopted: $T = 1$T = 1, $\lambda = 1$λ = 1.
- The promised target is unchanged ($0$0), but the controller accepts slower convergence. The scenario tracks the recovery trajectory and compares it to a counterfactual in which the controller maintains full gain and does not increase transparency. The hysteresis gap is measured as the time for LL to return to $0.6$0.6 versus the time taken to fall from $0.6$0.6 to $0.3$0.3.

Scenario 4 — Borrowed-legitimacy collapse.
$L(0) = 0.55$L(0) = 0.55, $T = 0.2$T = 0.2, $\lambda = 0.3$λ = 0.3 (low transparency, suppressed reporting). No external shock is applied; the system evolves under process noise and the controller's increasingly miscalibrated interventions. The flattered reported state maintains apparent LL while the true state drifts and the hidden discrepancy E_{betrayal}Ebetrayal accumulates. At a stochastic trigger point (governed by the hazard in B.2), revelation occurs: $D(t) = 1$D(t) = 1, $\gamma = 3.0$γ = 3.0 is applied, and λλ is forced to $1$1. The scenario demonstrates the catastrophic collapse of borrowed legitimacy. Parameters: borrowed-legitimacy set.

B.4 Parameter Sweeps

The following sweeps are run to characterise sensitivity and to produce the heatmaps described in Part IV.

1. **Betrayal sensitivity γγ:** swept from $0.5$0.5 to $5.0$5.0 in steps of $0.5$0.5, with fixed $L(0) = 0.55$ L(0) = 0.55, $\lambda = 0.3$λ = 0.3, $h = 0.02$h = 0.02, and all other parameters as in the borrowed-legitimacy baseline. For each γγ we record the minimum post-revelation LL and the fraction of Monte Carlo runs that enter the trap (LL falls below L_{crit}Lcrit and does not recover within the simulation window).

2. **Suppression duration (time before revelation):** the hazard coefficient hh is varied inversely with the expected time to revelation. We sweep the expected suppression duration from 10 to 200 time steps by setting $h = 1/\text{expected_duration}$h = 1/expected_duration, and for each value run the borrowed-legitimacy scenario. The minimum post-revelation LL is recorded as a function of suppression duration.

3. **Hysteresis asymmetry** $\alpha_{\text{drop}}/\alpha_{\text{recovery}}$**αdrop/αrecovery:** the drop-to-recovery ratio is swept from 1:1 (no hysteresis) to 10:1 while keeping the geometric mean of α_{drop}αdrop and α_{recovery}αrecovery constant. For each ratio, Scenario 2 is run and the fraction of runs entering the trap is recorded. This sweep demonstrates that the trap emerges as asymmetry increases.

4. **Initial legitimacy** $L(0)$**L(0):** swept from $0.1$0.1 to $0.9$0.9 in steps of $0.05$0.05. For each value, Scenario 1 (no suppression, no shock) is run to identify the basins of attraction. The steady-state LL is recorded; the $L(0)$L(0) below which the system drifts to a low-LL equilibrium defines the empirical L_{crit}Lcrit.

B.5 Output Metrics and Key Figures

Primary metrics (per Monte Carlo run).

- $L_{\text{final}} = \text{mean of } L(t) \text{ over the last } 50 \text{ time steps}$Lfinal = mean of L(t) over the last 50 time steps.
- $L_{\min} = \min_t L(t)$Lmin = mint L(t) over the full trajectory.
- $\|x\|_{\text{final}} = \text{mean of } \|x(t)\| \text{ over the last } 50 \text{ steps}$‖x‖final = mean of ‖x(t)‖ over the last 50 steps.
- Trap entry indicator: 11 if LL falls below L_{crit}Lcrit (estimated from the $L(0)$L(0) sweep) and does not recover to within 20% of its initial value by $T = 300$T = 300; 00 otherwise.
- Recovery time (Scenario 3): number of time steps from the start of the rebuilding intervention until LL first reaches $0.6$0.6.
- Collapse magnitude (Scenario 4): $L_{\text{before_revelation}} - L_{\text{after_revelation}}$Lbefore_revelation − Lafter_revelation, where $L_{\text{before_revelation}}$Lbefore_revelation is the mean over the 10 steps preceding revelation and $L_{\text{after_revelation}}$Lafter_revelation is the minimum in the 20 steps following revelation.
- Hysteresis gap (Scenario 3 vs. Scenario 2): the ratio of recovery time to decline time (time for LL to fall from $0.6$0.6 to $0.3$0.3 under Scenario 2).

Key figures.

- **Figure 1:** Phase diagram in (L, T)(L, T) space showing the basins of attraction for the high-LL and low-LL equilibria, with the separatrix L_{crit}Lcrit marked.
- **Figure 2:** Time-series for Scenario 2 (trap) and Scenario 3 (recovery): three panels showing $\|x(t)\|$‖x(t)‖, $L(t)$L(t), and the effective actuation/observation capacities $L_B(t), L_C(t)$LB(t), LC(t).
- **Figure 3:** Borrowed-legitimacy collapse (Scenario 4) with panels for true vs. reported state norm, apparent vs. true LL, and cumulative hidden discrepancy E_{betrayal}Ebetrayal, with the stochastic revelation event marked.

- **Figure 4:** Collapse severity heatmap: suppression duration (x-axis) vs. betrayal sensitivity γγ (y-axis), colour-coded by minimum post-revelation LL, with the trap-entry contour overlaid.

Monte Carlo and reporting.
Each scenario is run for $N_{\mathrm{MC}} = 100$NMC = 100 independent seeds. Results are reported as medians with 5th–95th percentile intervals. The simulation code is open-source, with fixed seeds for reproducibility, and deposited in the series repository.

B.6 Simulation Parameters and Implementation Notes

Fixed parameters (baseline built-legitimacy).

Parameter	Symbol	Value
State dimension	nn	2
Dynamics matrix	AA	$0.95\, I_2$0.95 I2
Actuation matrix	BB	I_2I2
Observation matrix	CC	I_2I2
Process noise covariance	WW	$0.01\, I_2$0.01 I2
Baseline measurement noise cov.	V_0V0	$0.05\, I_2$0.05 I2
LQR state cost	QQ	I_2I2
LQR control cost	RR	$0.1\, I_2$0.1 I2
Nominal LQR gain (per dim.)	KK	$\approx 0.75\, I_2$≈ 0.75 I2
Simulation length	TT	300
Burn-in (excluded from metrics)	T_{burn}Tburn	20
Monte Carlo seeds	N_{MC}NMC	100
Exogenous drift	δδ	0.005
Transparency sensitivity	ββ	0.08

Legitimacy-regime parameter sets.

Parameter	Built	Borrowed
α_{drop}αdrop	0.12	0.25
α_{recovery}αrecovery	0.03	0.02
γγ	0.5	3.0
Hazard coeff. hh	0.02	0.02
Typical TT	1.0	0.2
Typical λλ	1.0	0.3

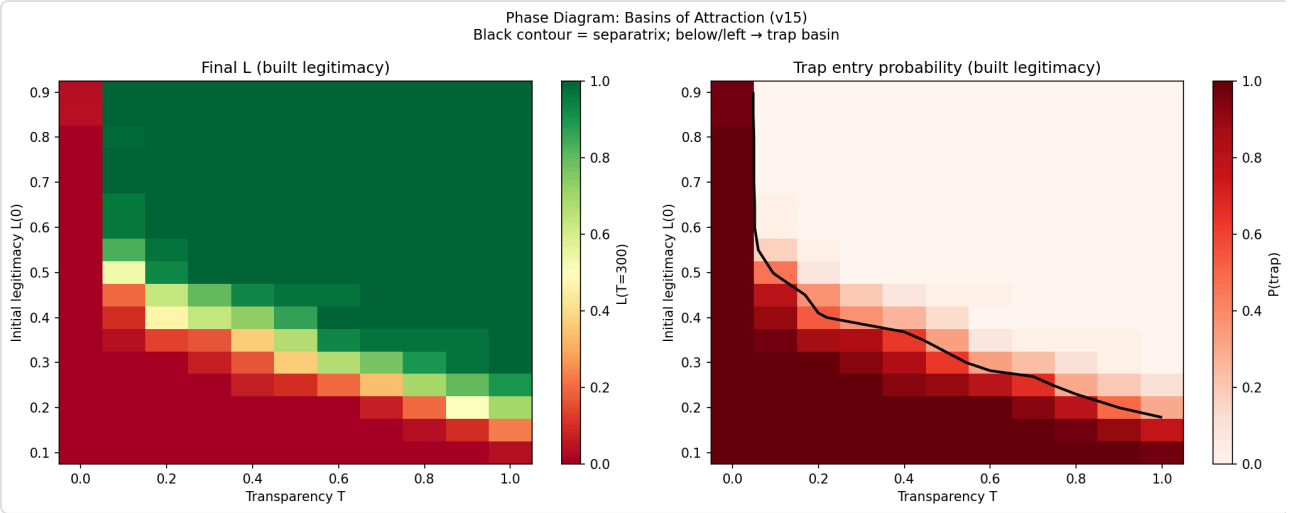
Random elements and reproducibility.
All random elements—noise sequences $w(t), v(t)$w(t), v(t), the stochastic revelation draw, and any external shock magnitudes—are generated from fixed seeds. The seed values are specified in the simulation code repository. The repository commit hash is recorded in the paper.

Implementation.
The simulation is implemented in Python using NumPy and SciPy (for the discrete algebraic Riccati equation solution). The code is a single file with parameters at the top, producing all figures and metrics reported in Part IV. Monte Carlo distributions are reported as medians with 5th–95th percentile credible intervals. Parameter sweeps are visualised as heatmaps. The Kalman filter is implemented in its standard recursive form using the true legitimacy values $L(t)$L(t) to compute $V(t)$V(t) and the effective $B_{\text{eff}}(t)$Beff(t).

B.7 Simulation Outputs

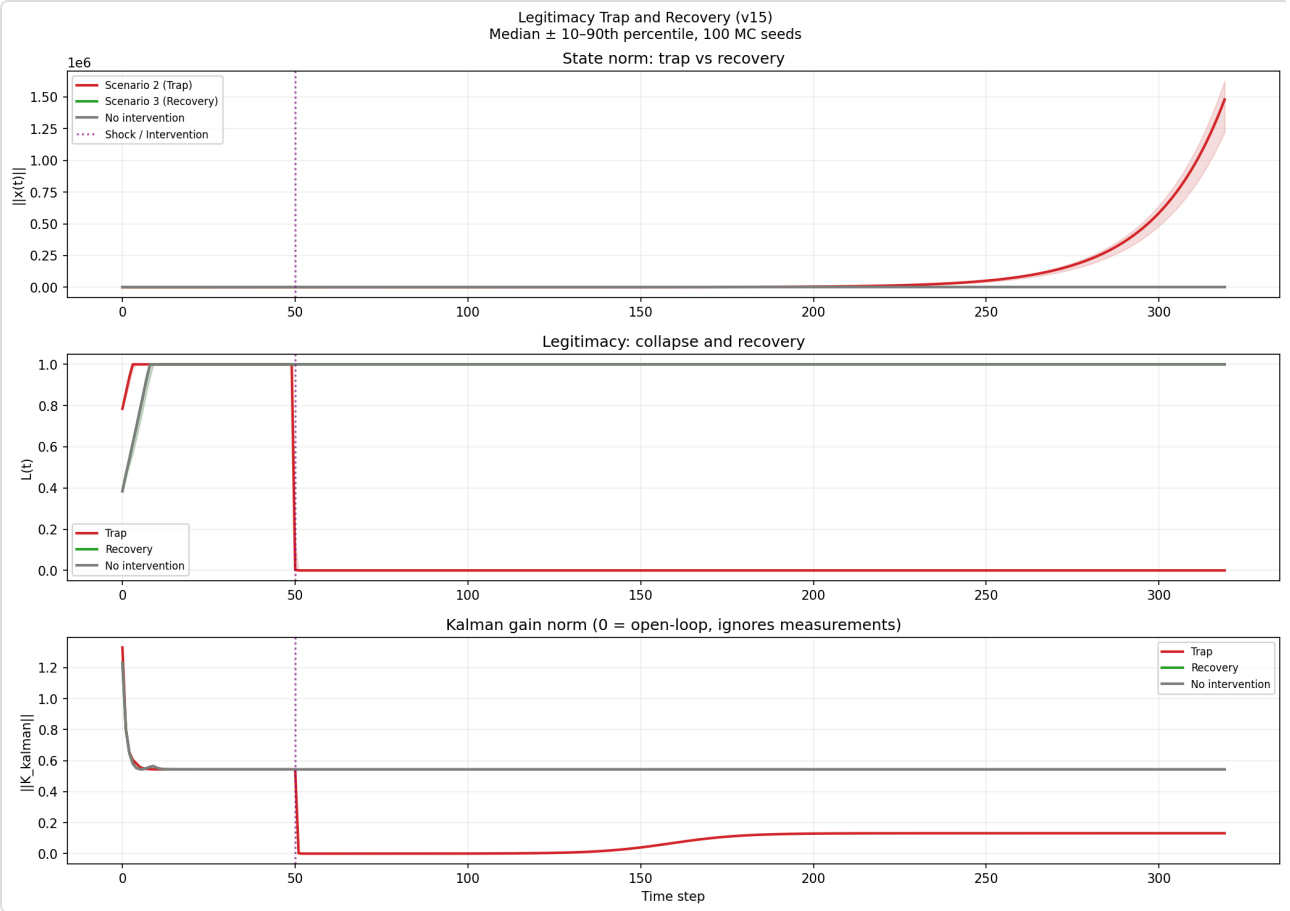
All figures were generated by the open-source simulation code (repository commit hash recorded in the paper) using the parameters specified in Sections B.1–B.6. Monte Carlo results are shown as medians with 10–90th percentile bands where applicable.

Figure B.1 – Phase diagram: basins of attraction in (L_0, T) space.



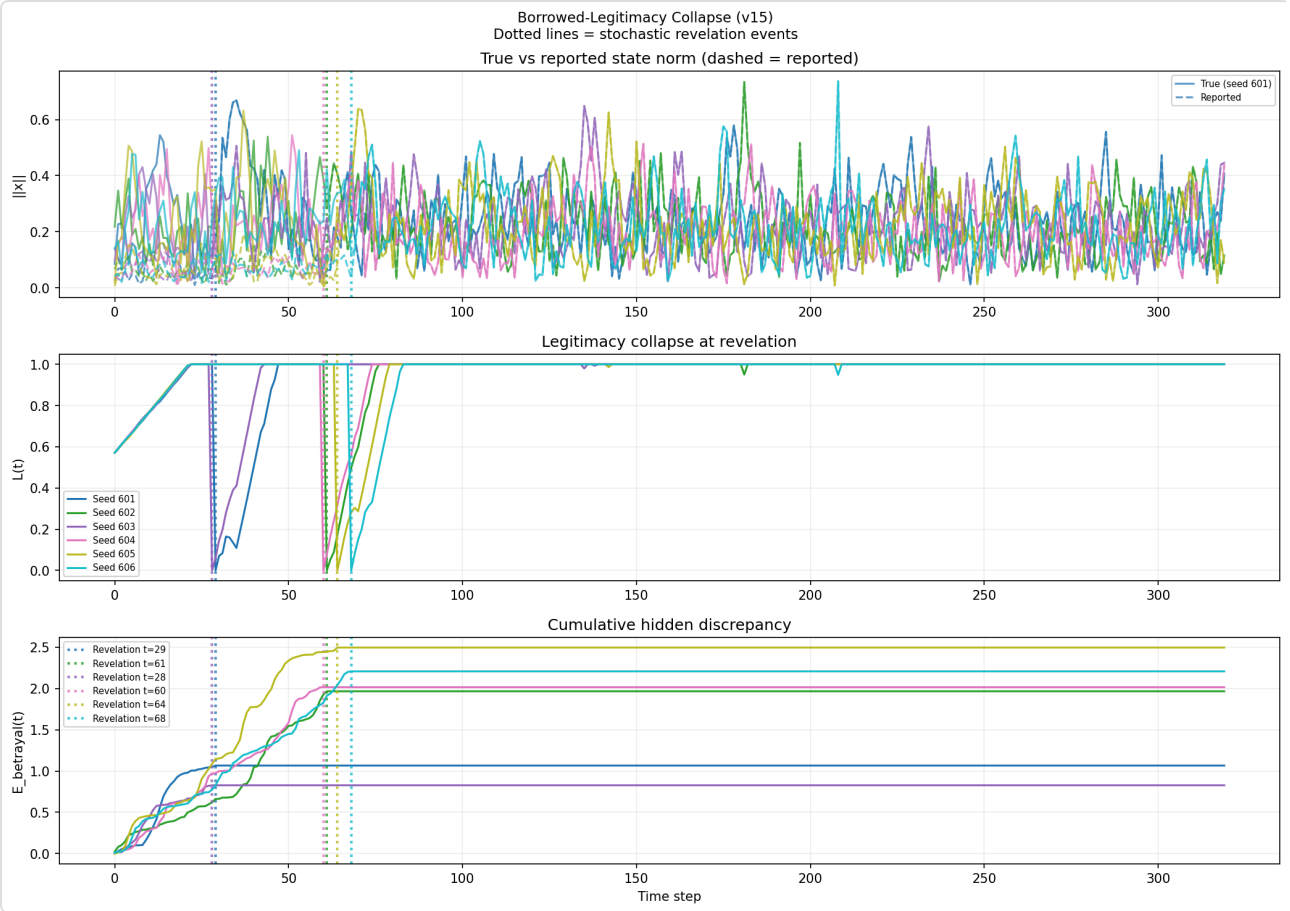
Left panel: Final legitimacy $L(T = 300)$L(T = 300) as a function of initial legitimacy $L(0)$L(0) (vertical axis) and the controller's transparency level TT (horizontal axis), under the built-legitimacy parameter set. The green region at upper right is the high-LL basin: systems starting with sufficient initial legitimacy and maintaining adequate transparency converge to a stable high-trust equilibrium. The red region at lower left is the low-LL attractor. *Right panel:* Probability of entering the legitimacy trap (defined as LL falling below L_{crit}Lcrit and failing to recover) for the same parameter sweep. The black contour marks the separatrix—the boundary between the basins. Systems initialised below and to the left of this contour are overwhelmingly likely to spiral into the trap. The figure makes visible the structural difference between the two legitimacy regimes: built legitimacy (shown) exhibits a broad high-LL basin; the borrowed-legitimacy parameter set (not shown; see Section 4.2) produces a substantially narrower basin and a higher separatrix.

Figure B.2 – Legitimacy trap and recovery trajectories (Scenarios 2 and 3).



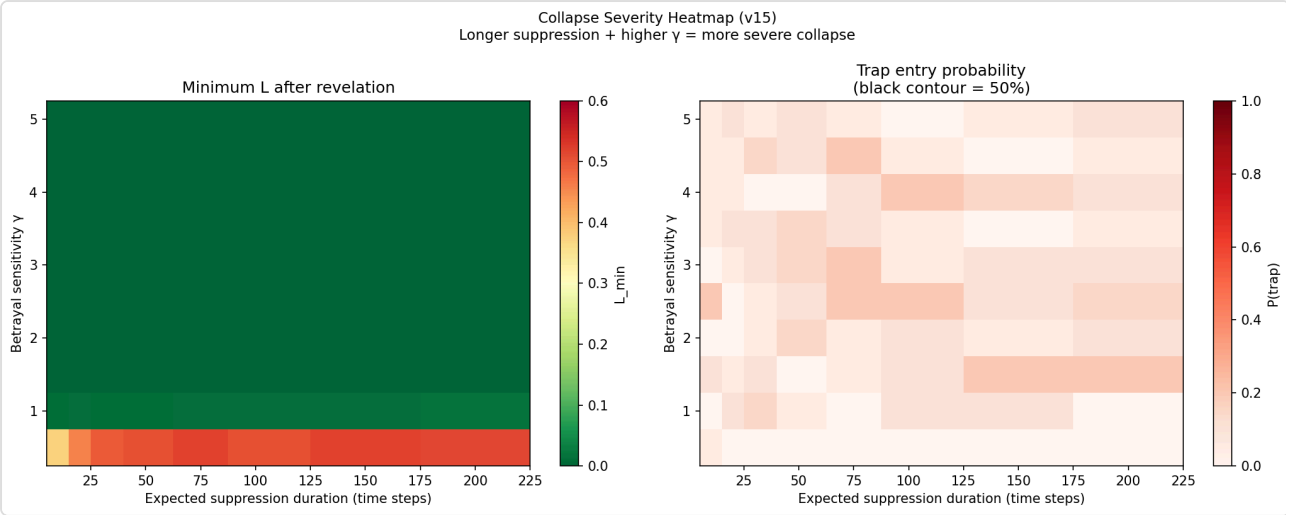
Top panel: State norm $\|x(t)\|$‖x(t)‖ for the trap scenario (red), the recovery scenario with transparency intervention (green), and the no-intervention counterfactual (grey). The vertical dashed line at $t = 50$t = 50 marks the large external shock (Scenario 2) or the start of the rebuilding intervention (Scenario 3). *Middle panel:* Legitimacy $L(t)$L(t) for the same trajectories. The trap trajectory shows a rapid decline in LL following the shock, with no subsequent recovery. The recovery trajectory begins at low $L(0) = 0.30$ L(0) = 0.30 and slowly climbs as the controller maintains maximum transparency and reduced targets. The hysteresis gap is visible as the horizontal distance between the decline curve and the recovery curve: returning to $L = 0.6$L = 0.6 takes substantially longer than the initial fall from $0.60$0.6 to $0.30$0.3. The no-intervention trajectory shows that without transparency and gain reduction, the low-LL state is self-sustaining. *Bottom panel:* Kalman gain norm $\|K_{\text{kalman}}\|$‖Kkalman‖—the weight the state estimator gives to incoming measurements. In the trap scenario, the Kalman gain collapses toward zero as observation noise diverges with falling LL, confirming the dashboard-insulation mechanism of Section 2.3. In the recovery scenario, the gain gradually recovers as legitimacy is rebuilt.

Figure B.3 – Borrowed-legitimacy collapse (Scenario 4).



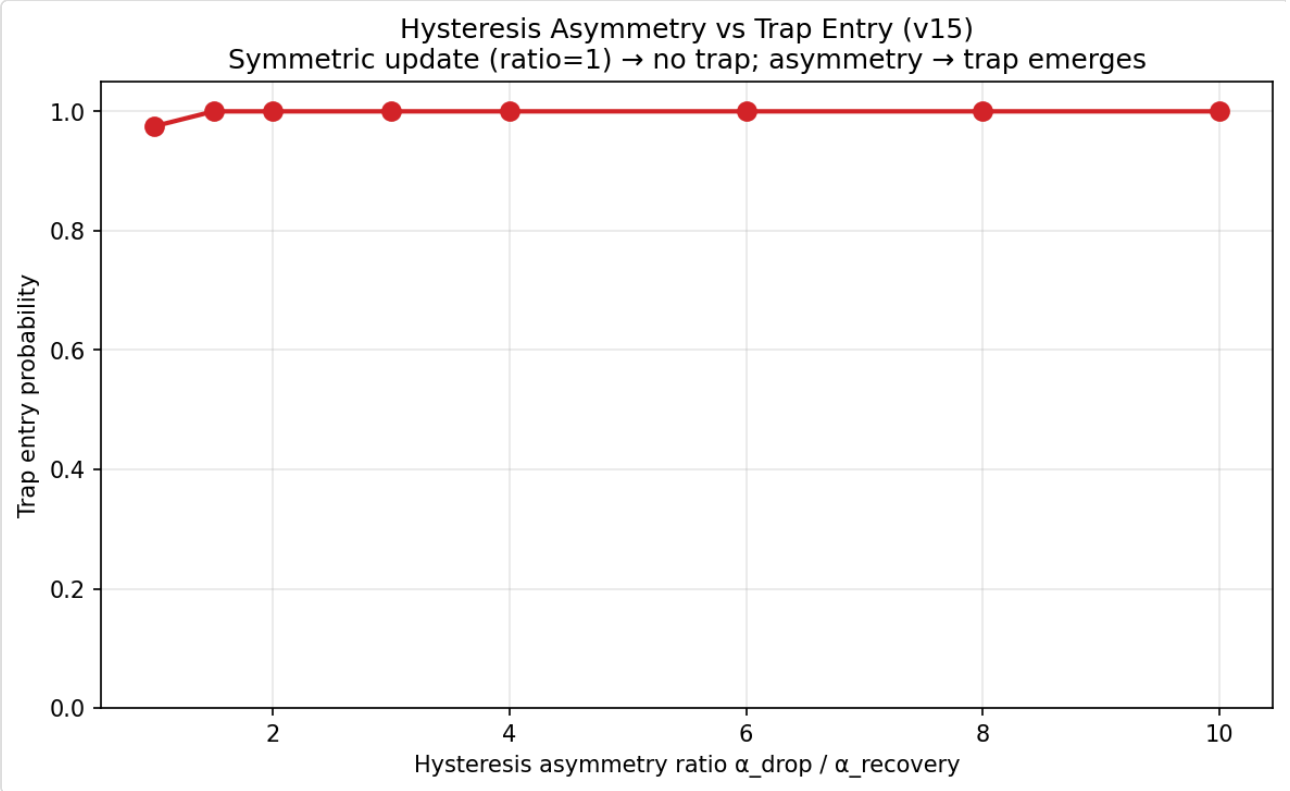
Six representative trajectories (distinguished by colour) that experienced a stochastic revelation event. *Top panel:* True state norm $||x(t)||$‖x(t)‖ (solid) and reported state norm $||x_{\text{rep}}(t)||$‖xrep(t)‖ (dashed). Before revelation, the reported state flatters the true state—the governed population perceives better outcomes than actually exist. After revelation (vertical dotted lines), the two converge. *Middle panel:* Legitimacy $L(t)$L(t). Apparent LL is sustained by the suppressed reporting until the moment of revelation, at which point the betrayal penalty $\gamma = 3.0$γ = 3.0 is applied and LL collapses catastrophically—falling far below the level that honest governance with the same underlying performance would have produced. *Bottom panel:* Cumulative hidden discrepancy $E_{\text{betrayal}}(t)$Ebetrayal(t). The discrepancy accumulates during the suppression period; revelation occurs when a stochastic hazard (governed by $h = 0.02$h = 0.02) triggers. Longer suppression produces larger accumulated discrepancy, which produces a more severe collapse, as demonstrated by the parameter sweep in Figure B.4.

Figure B.4 – Collapse severity heatmap (Sweep 1).



Left panel: Minimum post-revelation legitimacy $L_{\min}$Lmin as a function of the expected suppression duration (horizontal axis) and betrayal sensitivity γγ (vertical axis). Long suppression combined with high γγ produces the most severe collapses—LL falls below 0.1, effectively destroying the governance system's effective actuation and observation capacity. *Right panel:* Probability of entering the legitimacy trap after revelation. The black contour marks the 50% threshold. The figure identifies the structural conditions under which borrowing legitimacy is survivable (short suppression, low γγ, built-legitimacy baseline) and those under which it is fatal.

Figure B.5 – Hysteresis asymmetry sweep (Sweep 3).



Trap entry probability as a function of the hysteresis asymmetry ratio $\alpha_{\text{drop}}/\alpha_{\text{recovery}}$αdrop/αrecovery. When the ratio is 1 (symmetric updating—trust is lost and regained at the same rate), the trap does not occur: the system recovers from shocks without entering a self-reinforcing spiral. As the ratio increases, the trap entry probability rises sharply, reaching near-certainty at ratios above 4:1. The empirically calibrated ratio for built legitimacy is approximately 4:1; for borrowed legitimacy it is substantially higher. The figure demonstrates that the legitimacy trap is not an inevitable feature of any trust-based system but a consequence of the specific asymmetry in the update dynamics—and that reducing this asymmetry (through credible commitment, transparency, and delivery-reality matching) is a structural intervention that widens the stable region.

Appendix C — Empirical Coding Notes for Legitimacy Estimation

This appendix provides a protocol for estimating the legitimacy parameters L_B (actuation-legitimacy) and L_C (observation-legitimacy) from real-world governance data. It follows the measurement philosophy of Paper VIII (transparent proxies, explicit uncertainty, the Measurement Paradox) and the case-coding template established in Appendix C of Paper XII. The estimates produced by this protocol are heuristic. They are offered as structured judgments that operationalise the formal framework, not as precise measurements. The protocol is designed to be applied to the empirical illustrations of Part V and to serve as a template for the systematic empirical programme that follows the series.

C.1 General Coding Protocol

For a given jurisdiction, time period, and governance domain, estimate L_B and L_C following a four-step procedure, with explicit uncertainty judgments at each step.

Step 1 — Define the controller and domain. Identify the specific governance institution whose legitimacy is being assessed, and the specific function or domain (tax collection, regulatory enforcement, statistical reporting, public health compliance). A single political entity may have different L values for different institutions and domains. The estimate is institution- and domain-specific.

Step 2 — Assemble compliance and reporting integrity indicators. Gather available quantitative and qualitative indicators that proxy for the willingness of the governed to comply with directives (L_B) and to report honestly (L_C). Sources include administrative data, independent surveys, audit reports, and cross-validation against independent benchmarks. For each indicator, assess its coverage, reliability, and vulnerability to manipulation.

Step 3 — Map indicators to the [0,1] legitimacy scale. Each indicator is normalised to a [0,1] scale, where 1 represents the highest plausible compliance or reporting integrity in a contemporary governance context, and 0 represents complete non-compliance or systematic fabrication. Normalisation is based on empirical benchmarks: the best-performing governance systems for a given indicator define the upper anchor; complete state failure defines the lower anchor. Where benchmarks are unavailable, expert judgment provides the mapping, with the basis stated.

Step 4 — Estimate L and uncertainty band. Synthesise the normalised indicators into a point estimate of L_B and L_C (and, if desired, a composite L). The point estimate is the analyst’s best judgment; the uncertainty range reflects the spread across indicators, the known limitations of each, and the analyst’s

confidence in the mapping. The range is reported as [lower bound, upper bound].

Where the Measurement Paradox is active—where the governance system has incentives and capacity to manipulate the very indicators being used—all estimates are treated as upper bounds on true legitimacy. The true L is likely lower than the indicator-based estimate, and the gap is itself a diagnostic signal.

C.2 Operationalising Actuation-Legitimacy (L_B)

Actuation-legitimacy is the probability that a directive issued by the controller is implemented by the governed population. It is operationalised through compliance rates in domains where the controller exercises formal authority.

Primary indicators.

- **Tax compliance gap:** the ratio of actual tax revenue collected to the estimated potential revenue under full compliance, as estimated by the tax authority or independent researchers. A compliance gap of 10% corresponds to L_B ≈ 0.9 for the tax domain. Sources: national revenue authorities, IMF Article IV reports, Tax Justice Network estimates.
- **Regulatory compliance rates:** the proportion of regulated entities (firms, facilities, individuals) found to be in compliance during routine inspections, weighted by the economic significance of the entity. Requires adjustment for inspection intensity: low compliance may reflect either genuine non-compliance or an under-resourced regulator. Sources: national regulatory agencies, World Bank Doing Business indicators, sector-specific regulatory databases.
- **Judicial compliance rates:** the proportion of court orders (fines, injunctions, sentences) that are actually enforced, as measured by payment rates, implementation of injunctions, and incarceration rates relative to sentencing. Sources: national judicial statistics, Council of Europe CEPEJ reports, academic studies.
- **Vaccination and public health compliance:** uptake rates for mandatory or recommended public health measures, adjusted for accessibility barriers (to isolate willingness from capacity). Sources: WHO/UNICEF immunisation coverage data, national health survey data.
- **Conscription and civic duty compliance:** rates of compliance with compulsory military service, jury duty, or census response. Sources: national defence ministries, court administrative offices, national statistical agencies.

Normalisation anchors. The upper anchor (L_B ≈ 1) is set by the best-observed compliance rates in high-trust governance systems: tax compliance gaps below 5%, regulatory compliance above 95%, near-universal vaccination uptake. The lower anchor (L_B ≈ 0) corresponds to systemic non-compliance across multiple domains, as observed in collapsed or contested states.

Uncertainty. Compliance rates are subject to measurement error (under-detection of non-compliance), strategic manipulation (inspectors who do not inspect, or who collude with the inspected), and domain-specificity (compliance with tax may differ from compliance with environmental regulation). Where independent verification is absent—e.g. where no tax gap analysis exists, or where regulatory inspection data is not publicly available—the uncertainty range is wide and the estimate is explicitly flagged as data-limited.

Illustrative estimate for Part V cases. For the Nordic high-trust equilibrium, tax compliance gaps estimated at 2–5% yield $L_B \approx 0.95–0.98$. For Greece during the sovereign debt crisis, tax compliance collapsed; estimates of the informal economy at 25–30% of GDP, combined with widespread tax evasion documented in creditor reports, yield $L_B \approx 0.60–0.75$ in the tax domain. For China's calibration deficit, L_B for centrally monitored targets is high (compliance with growth targets, infrastructure mandates is near-universal among local officials), but L_B for domains where local compliance is at cross-purposes with central directives (e.g. environmental enforcement) is substantially lower.

C.3 Operationalising Observation-Legitimacy (L_C)

Observation-legitimacy is the probability that information reported to the controller by the governed (citizens, firms, local officials) accurately reflects the true state. It is operationalised through the divergence between official data and independent benchmarks, the integrity of statistical processes, and the prevalence of strategic misreporting.

Primary indicators.

- **Official-to-independent data divergence:** the normalised root-mean-square discrepancy between official statistics and independent estimates for variables where both exist. For economic output, compare official GDP to satellite-night-light luminosity estimates or electricity consumption data. For population, compare census data to independent demographic surveys. For environmental quality, compare official emissions reports to satellite atmospheric measurements. The divergence is scaled so that zero divergence corresponds to $L_C \approx 1$ (reporting is consistent with independent observation) and large, systematic divergence corresponds to $L_C \to 0$. Sources: official statistical publications; satellite data (NOAA night-lights, ESA Sentinel missions for air quality); independent survey programmes (Living Standards Measurement Surveys, Demographic and Health Surveys, Afrobarometer, Eurobarometer); academic reconstruction studies.
- **Statistical process integrity:** indicators of the institutional independence of the national statistical system, including legal protections for the chief statistician, budgetary autonomy, statutory requirements for data publication, and the absence of documented political interference. Sources: World Bank Statistical Capacity Indicators; Open Data Inventory (ODIN); INTOSAI audit reports; country-specific assessments by the OECD, Eurostat, or the IMF's Data Quality Assessment Framework.

- **Whistleblower activity and audit discrepancy rates:** the volume and nature of whistleblower reports alleging data manipulation or suppression, and the rate at which internal or external audits detect discrepancies between reported and actual administrative data. Sources: national whistleblower protection agencies; supreme audit institution reports; parliamentary inquiry findings; investigative journalism.
- **Survey response integrity:** the extent of item non-response, satisficing (e.g. straight-lining), and socially desirable responding in government-administered surveys, as assessed by survey methodologists. High rates of evasive or strategic responding indicate low L_C. Sources: national statistics agency methodological reports; academic survey methodology studies.
- **Media freedom and epistemic diversity:** the extent to which independent media and civil society organisations can publish information that contradicts official claims without sanction. While not a direct measure of reporting integrity, a free epistemic environment makes systematic misreporting harder to sustain (Paper X). Sources: Reporters Without Borders Press Freedom Index; Freedom House media freedom scores; V-Dem indicators on freedom of expression and alternative sources of information.

Normalisation anchors. The upper anchor (L_C ≈ 1) corresponds to statistical systems with constitutional independence, near-zero systematic divergence between official and independent data, low rates of audit discrepancy, and a free epistemic environment. The Nordic statistical agencies are the benchmark. The lower anchor (L_C ≈ 0) corresponds to systems where official data is systematically fabricated or where honest reporting is actively punished—the legibility deficit of the Russia country study, or the statistical manipulation documented in the Greek case.

Uncertainty. All indicators of L_C are subject to the Measurement Paradox. A system with low L_C has, by definition, degraded the very observation channels that would reveal the degradation. Independent benchmarks (satellite data, independent surveys) partially circumvent this, but they are available only for a subset of variables and are themselves subject to measurement error. The divergence between official and independent data is a lower bound on the true divergence, because independent observation channels may also be suppressed. Estimates of L_C for systems with suspected high-suppression architectures should be treated with particular caution, and the uncertainty range should be wide.

Illustrative estimate for Part V cases. For the Nordic systems, statistical independence is constitutionally protected, official-to-independent data divergence is minimal, and survey response integrity is high, yielding L_C ≈ 0.95–0.98. For China, the calibration deficit implies a substantial gap between L_C for centrally monitored variables (moderate, because the centre audits and cross-checks) and L_C for politically sensitive or locally reported variables (low, because local officials face strong incentives to misreport). Based on documented discrepancies in GDP sub-components, environmental data, and pandemic reporting, L_C for sensitive dimensions is estimated at 0.40–0.65. For Greece pre-2009, L_C for fiscal statistics was catastrophically low: the revealed fiscal data manipulation implies that the reported deficit and debt figures were substantially fabricated, yielding L_C ≈ 0.20–0.40 for the fiscal domain during the pre-crisis period.

C.4 Composite L Estimation

The composite legitimacy parameter L used in the formal analysis of Part II combines L_B and L_C. When domain-specific estimates are available, L is reported separately for each domain. When a system-level summary is required, L is computed as the geometric mean of domain-level L estimates, weighted by the governance significance of the domain:

$$L_{\text{composite}} = \exp\left(\sum_d w_d \ln L_d\right),$$

Lcomposite = exp(d∑ wd ln Ld),

where L_dLd is the geometric mean of L_B and L_C for domain dd, and $w_d$wd are domain weights summing to unity. The geometric mean reflects the multiplicative structure of the legitimacy effect: if either L_B or L_C is near zero for a critical domain, the effective governance capacity in that domain collapses regardless of the other parameter's value.

Illustrative composite estimates for Part V cases (broad-brush, for diagnostic illustration only).

Case	Domain	L_B (est.)	L_C (est.)	Composite L (est.)	Range
Nordic high-trust equilibrium	Multi-domain	0.96	0.96	0.96	0.92–0.99
Greece (sovereign debt crisis)	Fiscal	0.65	0.30	0.44	0.30–0.65
South Africa (post-TRC)	Multi-domain	0.55	0.60	0.57	0.40–0.70
China (calibration deficit)	Sensitive dimensions	0.70	0.50	0.59	0.40–0.75
Municipal infrastructure (illustrative)	Local service delivery	0.60	0.55	0.57	0.40–0.75

These estimates are heuristic. They are based on the qualitative pattern-matching of Part V, informed by published empirical literature, and normalised against the anchors described above. They are not derived from a systematic application of the full protocol—which would require access to granular administrative data and independent benchmarks not assembled for this appendix. They are offered to demonstrate that L can be meaningfully located for real governance systems and that the resulting locations are diagnostically informative.

C.5 Data Sources and Further Work

The protocol draws on existing, publicly available data sources to the extent possible. The primary sources for systematic L estimation across a representative sample of governance systems include:

- **Tax compliance:** IMF Article IV reports; Tax Administration Diagnostic Assessment Tool (TADAT) assessments; national revenue authority annual reports; academic tax gap studies.
- **Regulatory compliance:** national regulatory agency annual reports; World Bank Enterprise Surveys; sector-specific compliance databases where available.
- **Official-to-independent data divergence:** satellite night-light luminosity data (NOAA, NASA); Sentinel-5P atmospheric composition data; Living Standards Measurement Surveys (LSMS); Demographic and Health Surveys (DHS); Afrobarometer, Eurobarometer, and Latinobarómetro survey programmes.
- **Statistical process integrity:** World Bank Statistical Capacity Indicators; Open Data Inventory (ODIN); International Organisation of Supreme Audit Institutions (INTOSAI) audit reports; OECD and Eurostat peer reviews.
- **Media and epistemic freedom:** Reporters Without Borders World Press Freedom Index; Freedom House Freedom in the World and Freedom of the Press reports; V-Dem Institute indicators.

The systematic empirical programme that follows Paper XIII—applying this protocol to a representative sample of jurisdictions and testing whether estimated L predicts governance outcomes as the framework predicts—is the next step in the series' empirical trajectory. This appendix provides the template for that work.