Corrigibility as a Structural Precondition for Digital Public Infrastructure: A Cybernetic Framework*

A Preprint

Anivar A Aravind
Independent Researcher
Bengaluru, India
ping@anivar.net

April 2026
Revised: 11 July 2026

Abstract

Digital Public Infrastructure is evaluated using aspirational criteria: interoperability, inclusion, openness, scale. These criteria do not establish whether systemic errors can be corrected by affected participants. This paper defines corrigibility as the minimal structural precondition for reversible public infrastructure. Five jointly necessary conditions form a closed corrective loop: EXIT, CODE, AUDIT, GOVERN, FORK. Failure of any condition opens the loop, rendering correction discretionary. Using control-topology mapping grounded in Ashby's Law of Requisite Variety, the paper formalizes corrigibility as an architectural property: partial compliance is structurally equivalent to open-loop operation. Applied analysis demonstrates discriminative power against existing infrastructures. The five conditions are stated as an invariant of public infrastructure rather than of the technology that implements it. A companion paper carries the same invariant to learned and agentic systems, where only the verification machinery changes. Corrigibility does not guarantee fairness. It guarantees reversibility.

Keywords Corrigibility · Digital Public Infrastructure · Cybernetics · Feedback Control · Requisite Variety · Infrastructure Accountability · India Stack

Executive Summary

Can the system be corrected when it goes wrong? The criteria by which Digital Public Infrastructure (DPI) is currently evaluated — openness, interoperability, inclusion, scale — do not answer that question.

Corrigibility is the structural ability of affected participants to detect, contest, and override systemic error. Without this property, digital infrastructure operates in open-loop: errors may be acknowledged, but correction remains discretionary.

Five architectural constraints are required for corrigibility:

1. Reversible participation (**EXIT**)
2. Inspectable logic (**CODE**)
3. Independent verification (**AUDIT**)
4. Binding participatory constraint (**GOVERN**)
5. Independent reproducibility (**FORK**)

These constraints operate across observability, participation, constraint, and replacement layers. The absence of any one constraint creates structural asymmetry. The framework provides testable criteria for evaluating real-world DPI systems.

*SSRN: doi.org/10.2139/ssrn.6059075

How to Read This Paper. The argument proceeds in three parts: public systems require bounded error propagation (the invariant); five conditions close the corrective loop (the architecture); these conditions discriminate between real systems (the evaluation). Case studies illustrate discriminative power, not exhaustive institutional assessment.

1 Introduction

Digital Public Infrastructure is not merely a collection of software stacks; it is the encoding of political arrangements into technical artifacts [Winner, 1980]. Access to essential survival services (food distribution, banking, healthcare, mobility) is increasingly mediated through these systems. Because these systems apply fixed rules to the infinite variety of human life, they will inevitably misclassify, exclude, or fail specific users. This is not a malfunction; it is a mathematical certainty when finite rule sets encounter human diversity.

The critical question for a digital polity is not whether errors occur, but whether the architecture permits those affected to **detect, correct, and reverse** them before harm becomes permanent.

Current policy discourse defines DPI through aspirational language. The G20 New Delhi Declaration (2023) and the UN/UNDP framework [United Nations Development Programme, 2024] describe systems that "should be secure" or "can be built on open standards." These definitions exclude almost nothing. They articulate intent, not condition. Under such broad definitions, a system that traps users in a coercive loop can be labeled "public infrastructure" simply because it operates at scale.

This paper defines corrigibility not as a moral preference, but as a structural property essential for stability. The central argument is that **unverifiable constraint is structurally indistinguishable from absolute operator control.** Therefore, public legitimacy requires more than policy assertions; it requires a cryptographic chain of custody and a verifiable capacity for correction. The normative grounding for this structural definition — non-domination as the criterion for legitimate public infrastructure — is developed at the opening of Section 3.

Contributions. This paper formalizes the structural preconditions for public digital infrastructure, transitioning the discourse from normative aspirations to falsifiable engineering constraints. Specifically, the contributions are threefold:

1. **A Synthesis of Corrigibility**: This framework distills principles from cybernetics, commons governance, and free software into five verifiable structural tests (EXIT, CODE, AUDIT, GOVERN, FORK).
2. **Control-Theoretic Formalization**: A control-theoretic structural model (Appendix B) shows that partial compliance with these tests is structurally equivalent to open-loop operation.
3. **Empirical Vulnerability Assessment**: The framework evaluates major national infrastructures (e.g., Aadhaar, UPI), formalizing the resource barriers to structural accountability.

Paper Structure. Section 2 diagnoses why current DPI definitions fail. Section 3 establishes the theoretical foundations from cybernetics, commons governance, and free software. Section 4 derives the five structural tests. Section 5 immediately grounds these tests in empirical evaluation of existing infrastructure, establishing concrete stakes before proceeding to dynamics. Section 6 analyzes the political economy of incorrigibility: temporal dynamics, correction velocity, essential services, and the sovereignty–scale–neutrality tension. Section 7 addresses methodological foundations and implementation pathways. The appendices formalize the control-theoretic model and provide machine-readable schemas for automated assessment. A glossary of key terms appears in Appendix F.

1.1 Corrigibility: Minimal Structural Definition

Corrigibility is the property of a system in which affected participants can detect, contest, and structurally override systemic error. It is not equivalent to transparency, accountability, openness, or auditability in isolation. A system may publish documentation, provide APIs, or maintain oversight boards and yet remain structurally incorrigible if participants lack binding corrective leverage.

Formally, a digital public infrastructure is structurally corrigible if and only if it satisfies five jointly necessary conditions — sufficient only in the narrow sense of structural access to correction — that together close the feedback loop between system behavior and participant correction. These conditions operate across distinct control layers: observability, participation, constraint, and replacement. The failure of any one condition converts the system into an open-loop structure in which error may accumulate without guaranteed structural correction.

The remainder of this paper defines these conditions and demonstrates their necessity.

Verification Tiers. Corrigibility can be assessed at three operational tiers: (1) *Presence*: laws, bodies, and policies exist on paper; (2) *Behavior*: controls execute under stress, audits produce consequences; (3) *Proof*: trust is continuously testable, authority is scoped and revocable, claims are machine-verifiable, failures are bounded. Most DPI deployments satisfy Presence. Few reach Behavior. Almost none achieve Proof. The five conditions define what must be verified. The tiers define how rigorously.

2 The Problem: Definitions Without Structure

The accepted definitions of Digital Public Infrastructure share a common defect: they describe desired outcomes without specifying the architectural preconditions required to achieve them. This section dissects the authoritative definitions to demonstrate their structural inadequacy.

2.1 The Definitional Landscape

The G20 New Delhi Leaders' Declaration (2023) established the international consensus:

> "a set of shared digital systems that should be secure and interoperable, and can be built on open standards and specifications to deliver and provide equitable access to public and/or private services at societal scale" [Group of Twenty (G20), 2023].

The UNDP framework [United Nations Development Programme, 2024] elaborates:

> "DPI refers to digital solutions that enable basic functions essential for public and private service delivery, including digital identity, payments, and data exchange. These solutions can be built as open-source, with open standards and specifications, and should incorporate principles of inclusion, security, and privacy by design."

The World Bank's identification principles add:

> "Establish a trusted—unique, secure, and accurate—identity" and "[c]reate a responsive and interoperable platform," toward identification systems that are "inclusive, trusted, accountable" [World Bank, 2021].

2.2 Anatomizing the Aspirations

These definitions deploy a consistent vocabulary: *secure*, *interoperable*, *open*, *inclusive*, *accountable*. Examined structurally, each term collapses under scrutiny.

"Secure." Secure for whom? A system can be cryptographically robust against external attackers while remaining structurally impervious to those it governs. The CIDR (Central Identities Data Repository) in Aadhaar is "secure" only in the operator's sense. UIDAI asserts that the repository resists breaches, and no outside party can test the assertion. It is not secure in the sense that citizens can verify its operation or contest its determinations. Security without accountability is a fortified monopoly.

"Interoperable." Interoperable with what? A system that speaks standard protocols to upstream services while maintaining proprietary control over downstream users achieves interoperability for operators, not subjects. UPI's API specifications are interoperable; citizens cannot interoperate with NPCI's switching logic. The interface is public; the machine is private.

"Open Standards." Open standards are not open execution. As Simon Phipps of Sun Microsystems observed, "FOSS serves as the canary in the coalmine for the word 'open'. Standards are truly open when they can be implemented without fear as free software in an open source community" [Phipps, 2007]. PDF is an open ISO standard; Adobe Acrobat's execution of it is proprietary. A system can speak a public language while its internal logic remains invisible. Similarly, a payment rail can implement open APIs while its routing algorithms, fee structures, and dispute resolution mechanisms remain black boxes. Publishing a specification does not make a system inspectable. The critical question for any "open standard" claim is: *which open standard?* If the specification URL cannot be shared publicly, or if the standard cannot be implemented in free software without legal encumbrance, the openness is performative.

"Inclusive." Inclusive under what conditions? A system that enrolls everyone achieves enrollment inclusivity. But enrollment is the beginning, not the end. A system that enrolls everyone while providing no mechanism for contestation or exit achieves inclusive capture. The question is not "Can everyone get in?" but "Can anyone get out?" Mandatory enrollment with no exit is not inclusion; it is mandatory lock-in.

"Accountable." Accountable to whom, through what mechanism, with what enforcement? The definitions invoke accountability as an adjective, not a structure. They specify no verification procedure, no audit right, no correction channel. Accountability without mechanism is rhetoric.

2.3 The Four Structural Gaps

The definitional inadequacy manifests as four distinct capture vectors:

First, Open Standards are not Open Execution. A proprietary system can speak a standard language while remaining entirely opaque in its internal logic. (Addressed by the **CODE** test, Section 4.)

Second, Legal Frameworks do not guarantee Technical Rights. A system that requires a court order to correct a database error effectively denies correction to the majority of its users. Legal remedy operates on bureaucratic timescales that cannot match infrastructure execution speed. (Addressed by the correction velocity inequality, Section 6.1.)

Third, Aspiration is not Compliance. Definitions that rely on what a system "can" do or "should" be offer no resistance to authoritarian drift. Modal language creates no enforceable constraint. Every term in the G20 definition is advisory ("should be secure," "can be built on open standards"). Advisory language excludes nothing. (Addressed by the **GOVERN** test, Section 4.)

Fourth, Scale is not Accountability. A system that serves a billion users achieves universality, but universal deployment does not confer public legitimacy. Population-scale operation may simply mean population-scale capture. The "public" in such definitions refers to the number of subjects, not the structure of control. (Addressed by the essential services problem, Section 6.7.)

2.4 Structural Theater and Open-Washing

The absence of structural criteria enables what we term **Open-Washing**: the invocation of openness, interoperability, or digital sovereignty as reputational signals without satisfying reproducibility or governance conditions.

This is not a theoretical concern. Mozilla warned in 2017 that Aadhaar's claims of openness were misleading: "The development was not open, the source code is not open" [Baker and Gadgil, 2017]. In 2020, Mozilla and the Internet Society cautioned that India's National Open Digital Ecosystems (NODE) framework risked institutionalizing open-washing by leaving "the definition of 'open' vague and at the complete discretion of individual implementers," enabling "closed ecosystems that are only open in appearance while being closed in practice" [Mozilla Foundation, 2020].

In the absence of EXIT, CODE, AUDIT, GOVERN, and FORK guarantees, claims of openness are merely descriptive properties of implementation rather than enforceable constraints on power. A system may publish APIs, release partial source code, or adopt open standards while remaining structurally incorrigible.

Open-washing therefore constitutes a category error: conflating technical transparency with structural accountability. The corrigibility framework distinguishes *symbolic openness* from *constitutional openness* by requiring that affected participants retain binding pathways for contestation and reproduction.

2.5 The Definitional Thesis

Without hard structural requirements, "Public" becomes a mere descriptor of scale rather than a mode of accountability. Just as classification systems become invisible as they gain acceptance [Bowker and Star, 1999], incorrigible infrastructure recedes into the background, becoming unchallengeable precisely as it becomes essential.

Remark 2.1 (The Definitional Problem)**.** *Current DPI policy specifies **intent, not conditions**. It articulates what systems should achieve without specifying what they must structurally provide. The framework proposed in this paper converts aspirational adjectives into verifiable predicates. It asks not "Is this system described as open?" but "Can independent parties inspect its operation?" Not "Is this system intended to be accountable?" but "Can those it affects trigger correction?"*

This danger is acute as digital infrastructure expands from simple databases to complex orchestration layers. When systems become opaque by design (whether through proprietary code, closed APIs, or undocumented decision logic), the gap between definitional aspiration and structural reality widens further. A companion paper [Aravind, 2026] extends this framework to learned systems (AI/ML) where additional verification challenges arise.

2.6 Related Work

The DPI literature spans policy frameworks, technical architectures, and critical analyses, but lacks a unified structural standard for accountability.

Policy Frameworks. The G20 framework [Group of Twenty (G20), 2023] and UNDP definitions [United Nations Development Programme, 2024] establish aspirational principles (openness, inclusion, interoperability) without verification criteria. The World Bank's Identification for Development (ID4D) initiative emphasizes functional identity but does not specify architectural requirements for user exit or independent audit. These frameworks describe *what* DPI should achieve without specifying *how* to verify achievement.

Measurement Frameworks. The most comprehensive global assessment, covering 210 countries across digital identity, payments, and data exchange systems [Fetter et al., 2025], measures six attributes: interoperability, oversight, privacy, inclusion, adoption, and coordination. Yet the framework explicitly acknowledges that "governance measurements identify the prevalence of legal frameworks and oversight bodies but cannot assess enforcement capacity." The variables measure *presence*: does an oversight body exist, does a data protection act exist, does an interoperability policy exist. They do not measure *function*: can citizens invoke these mechanisms to correct systemic error? The findings are stark: only 3% of payment systems meet privacy-related variables; only 12% meet inclusion variables; only 50% of digital identity systems meet interoperability requirements. Even these low figures overstate structural accountability, because presence of a legal framework does not establish citizen capacity to activate it. A system may have an "audit mechanism" (UCL variable) while failing AUDIT (this framework) because no independent party can verify execution. A system may have "participation conditions" while failing EXIT because no functional alternative exists. Terminological alignment masks structural divergence; in the terms of Section 1.1, these are Tier-1 (Presence) measurements presented as if they established Tier-2 (Behavior). The same schema now governs deployment finance: the World Bank's Global DPI Program reports reach and enrollment — eighty countries, 176 million users of new or enhanced services in FY25 — as its results, with no corrective-capacity metric among them [World Bank, 2026a].

Critical Analyses. Scholarship on India Stack has documented exclusion harms [Khera, 2019] and constitutional tensions [Bhatia, 2019]. However, these critiques remain case-specific rather than providing generalizable tests applicable across jurisdictions and technologies.

Control Theory and Governance. Cybernetic approaches to governance date to Ashby [Ashby, 1956] and Beer's viable systems model, but have not been systematically applied to digital infrastructure accountability. Lessig's "code is law" [Lessig, 2006] establishes that architecture constrains behavior, but does not specify which architectural properties are necessary for democratic legitimacy.

Software Freedom. The free software tradition [Stallman, 2002] provides operational tests (inspect, modify, redistribute) but focuses on code artifacts rather than infrastructure systems that include data, compute, and network effects.

Gap and Contribution. No existing framework provides falsifiable, technology-agnostic tests for infrastructure accountability that (1) derive from formal stability theory, (2) apply to deterministic systems at population scale, and (3) yield machine-verifiable assessments. This paper addresses that gap by synthesizing cybernetics, commons governance, and software freedom into five structural tests with formal model and empirical application.

3 Theoretical Foundations

Normative Anchor. This framework adopts *non-domination* — the structural absence of arbitrary power — as the normative criterion for legitimate public infrastructure, drawing on the analytical core of Pettit's republican tradition [Pettit, 1997]. Alternative frameworks reach similar conclusions through different routes: Habermas's deliberative theory [Habermas, 1996] centers communicative legitimacy; Sen's capability approach [Sen, 1999] centers substantive freedoms; Mouffe's agonistic democracy [Mouffe, 2000] centers the productive maintenance of conflict. We center non-domination because its focus on structural protection provides the most exact isomorphism with

system architecture: a digital system is public to the extent that those it governs possess the structural capacity to override its arbitrary application. The five corrigibility tests operationalize the engineering conditions under which this arbitrariness is bounded. While the framework's mechanics are compatible with overlapping democratic traditions, its foundational metric is the structural distribution of power rather than the optimization of administrative utility.

To operationalize this defense against arbitrary power, we do not invent new ethical categories. Instead, we triangulate the mechanical requirements of non-domination from three independent operational traditions: Cybernetics (to map the necessary feedback loops), Commons Governance (to structure the constraint), and Free Software (to legally guarantee the right to inspect and reproduce). Whether viewed through physics, political economy, or code, the stability of a system depends on the capacity of its subjects to provide corrective feedback.

3.1 Cybernetics and Requisite Variety

In cybernetics, the relationship between a controller and its environment is governed by Ashby's Law of Requisite Variety [Ashby, 1956]. The law states that for a system to be stable, the number of control states must meet or exceed the number of environmental states:

$$V(\text{controller}) \geq V(\text{disturbance}) \tag{1}$$

In DPI, the "disturbance" is the boundless diversity of the human population. No pre-programmed rule set can match this variety. Stability, therefore, relies on feedback channels that transmit error signals from the population back into the system's behavior. If these channels (specifically the ability to exit or correct) are blocked, the controller's variety effectively collapses. The system loses the capacity to regulate its own impact, leading to accumulated errors that manifest as either instability or systematic exclusion.

Human variety includes technical and social edge cases that deterministic rule sets cannot exhaustively enumerate: differences in biometric capture conditions (e.g., manual laborers with worn fingerprints, elderly individuals with fading iris patterns), non-standard household and family structures that diverge from canonical registries, multilingual and script variance that invalidates literal string matches, and intermittent or low-bandwidth contexts where real-time digital verification is impossible. These concrete phenomena instantiate Ashby's theoretical argument: controllers lacking channels for diverse corrective signals will systematically misclassify and exclude.

3.2 The Commons and Constitutional Constraint

Elinor Ostrom's work on commons governance [Ostrom, 1990] establishes that sustainable systems require monitors who are accountable to the users and collective-choice arrangements that allow users to modify the rules. A system in which rule-making is the exclusive preserve of the operator acts as an enclosure, not a commons.

Critically, Ostrom identifies that governance is not merely about access; it is about graduated sanctions and low-cost conflict resolution. Applied to digital infrastructure, this implies that users must possess **Constitutive Power**: the ability to set binding limits on the system's behavior. If a system's constraints are merely advisory or can be overridden by the operator at will, the system is fundamentally ungoverned.

3.3 Free Software and the Right to Reproduce

The Free Software movement [Stallman, 2002] contributes the mechanical means of enforcement. Stallman's "Four Freedoms" are often misread as a philosophy of sharing, but they are functionally a philosophy of power. The freedom to study (Inspectability) ensures power is visible; the freedoms to modify and distribute (Reproduction) ensure that power is not a monopoly.

This tradition clarifies that openness is not a passive property (reading code) but an active one (forking infrastructure). Transparency without the power to act on it is functionally inert. The ability to reproduce the system independent of its original steward is the only ultimate check against operator capture.

3.4 Failure Case: Open-Loop Digital Infrastructure

Consider a national identity infrastructure that publishes APIs and technical documentation but provides no practical exit mechanism, restricts audit access to operator-approved entities, and centralizes governance authority without binding participant constraint. Suppose a classification error affects a subset of users.

Users may complain. Journalists may report. Regulators may inquire. However, if no participant can independently verify the error (AUDIT), exit without disproportionate penalty (EXIT), impose binding modification (GOVERN), or reproduce a functionally equivalent alternative (FORK), then the system remains structurally unchanged. Error may be acknowledged yet persist.

Such a system is informationally open but structurally closed. The feedback loop between harm detection and systemic correction is incomplete. The infrastructure operates in open-loop mode with respect to affected participants.

Corrigibility requires closing this loop.

4 The Structural Conditions of Corrigibility

Five architectural constraints are required for structural corrigibility. These conditions operate across distinct control layers and form a weakest-link system: the failure of any one condition collapses corrective authority from guaranteed to discretionary.

4.1 Observability Layer: CODE and AUDIT

CODE requires inspectability of the operative logic determining system behavior. **AUDIT** requires that independent actors can verify system outcomes without operator discretion.

Without CODE, internal rule structure is opaque. Without AUDIT, error detection is operator-controlled. Observability is necessary but insufficient for corrigibility.

4.2 Participation Layer: EXIT

EXIT requires that participants can withdraw from the system without disproportionate penalty. Exit functions as an error signal. Where exit is infeasible, the system is insulated from negative feedback and may externalize harm.

4.3 Constraint Layer: GOVERN

GOVERN requires that affected participants possess binding mechanisms to modify system rules. Advisory participation is insufficient. Corrigibility requires enforceable constraint, not consultation.

4.4 Replacement Layer: FORK

FORK requires that a functionally equivalent system can be independently reproduced. Forkability ensures that corrective governance cannot be permanently blocked by centralized control.

Together, these conditions form a closed-loop structure. Figure 1 stacks the four control layers; Figure 2 renders the same architecture as a continuous feedback diagram. The five conditions are jointly necessary for structural corrigibility. They are sufficient in a narrow architectural sense: if satisfied, corrective mechanisms remain structurally accessible to affected participants. This sufficiency concerns structural access to correction, not optimal outcomes, fairness, or epistemic quality.

Architectural conditions are necessary but not complete. Durable legitimacy also requires machine-verifiable proof: cryptographic attestations, anchored logs, and revocation and change histories that make claims about authority, constraints, and remedial actions demonstrably true at decision time.

4.5 Test 1: EXIT (Reversibility of Participation)

Definition 4.1 (EXIT). *A system S satisfies **Reversibility of Participation** if, for any user in state $s \in S$, there exists a transition to a non-participatory state s_0 such that the penalty $\pi(s \to s_0)$ is strictly bounded: $\pi(s \to s_0) < \tau_{exit}$, where τ_{exit} is a threshold below which the penalty does not constitute material exclusion from essential services. For essential systems, where no bounded-penalty s_0 can exist (Proposition 6.2), the test is dischargeable in exactly one other way: verified Functional Exit Equivalence (Section 6.7.2), which recreates the error-signal strength literal exit would have supplied. The test remains binary — bounded-penalty exit, or verified FEE, or failure.*

The foundational condition of any corrigible system is the **Reversibility of Participation**. Irrevocable consent functions structurally as mandatory lock-in. For an infrastructure to operate as a public utility rather than a coercive

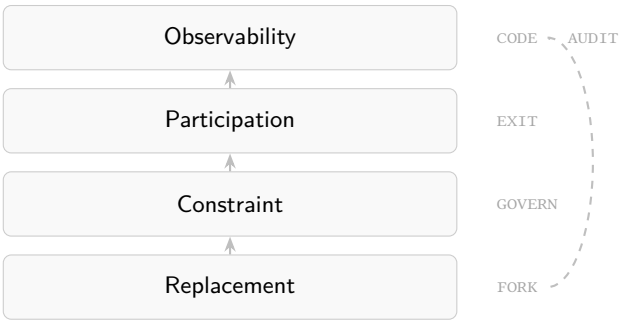

Figure 1: **Closed-Loop Corrigibility Architecture.** Four control layers form a closed feedback loop. Removal of any layer converts the system to open-loop, permitting unbounded error accumulation.

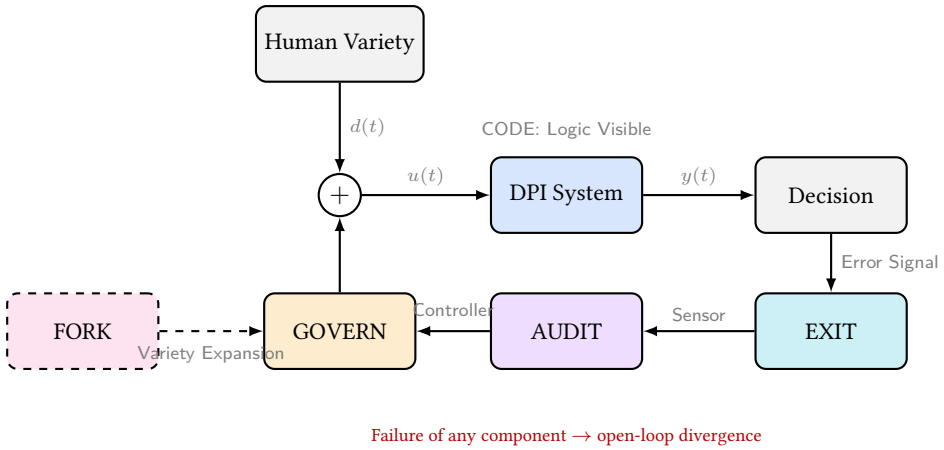

Figure 2: **Corrigibility as a Closed-Loop Control Architecture.** Digital Public Infrastructure is modeled as a feedback control system subject to continuous human and environmental disturbance. Corrigibility requires that error signals (EXIT) remain detectable, that system behavior be intelligible (CODE), that sensing mechanisms be transparent (AUDIT), that binding corrective authority exist (GOVERN), and that structural reproduction be feasible (FORK). Failure of any single component collapses the loop into open-loop operation. Under non-zero integrating disturbance, error diverges unboundedly (Null-Feedback Instability, Proposition B.1; Appendix B). Corrigibility is therefore a structural property of loop integrity, not a matter of operator intent.

instrument, users must possess the capacity to withdraw from it without incurring disproportionate penalty or material exclusion.

In practice, many systems operate as a ratchet. They are easy to enter but functional imperatives make them impossible to leave. **Aadhaar** exemplifies this failure: opting out results in exclusion from subsidized food rations under the upheld Section 7 linkage, and — although *Puttaswamy II* (2018) struck down mandatory linking for bank accounts and SIM cards — de facto insistence on Aadhaar at the point of service persists across banking and connectivity [Supreme Court of India, 2018, Khera, 2019]. When a system becomes a prerequisite for existence, refusal results in a disproportionate survival penalty rather than functioning as feedback. **UPI** partially mitigates this because cash remains legal tender, but network effects increasingly marginalize non-participants. **Linux**, by contrast, imposes no survival penalty: a user can migrate to BSD or Windows without losing access to essential services.

From a cybernetic perspective, EXIT constitutes the primary negative feedback loop. Blocking this channel silences the user's error signal. This blockade overloads the political layer with demands that the technical layer has suppressed.

This dynamic was formalized by Hirschman [Hirschman, 1970] in his analysis of organizational decline. Hirschman distinguished two correction mechanisms: exit (withdrawal from participation) and voice (attempts to change the system from within). His central finding is that these mechanisms exist in productive tension, not as substitutes.

Exit, in Hirschman's formulation, is "neat," "impersonal," and "indirect": one either exits or one does not. This binary quality makes exit a high-bandwidth error signal. When users leave, the system receives unambiguous information

that something has failed. Voice, by contrast, is costly and conditioned on the influence and bargaining power that participants can bring to bear. Voice is gradual, political, and requires sustained effort.

Critically, Hirschman demonstrated that exit without voice produces abandonment rather than correction: quality-conscious users flee rather than fight. But voice without exit produces capture: complaints become cosmetic (see Section 6.4.1) because the system faces no credible threat of user departure. The effectiveness of voice depends on the credibility of exit.

This framework illuminates why mandatory enrollment is structurally pathological. By eliminating exit, a system does not merely suppress one feedback channel. It degrades voice itself. When citizens cannot leave, their complaints lack credibility. The operator can absorb infinite grievance without modifying system behavior ($\partial S/\partial G = 0$). Hirschman's insight formalizes what cybernetics predicts: blocking the error signal does not eliminate the error; it eliminates the system's capacity to respond.

4.6 Test 2: CODE (Inspectability of Logic)

Definition 4.2 (CODE)**.** *A system S satisfies **Inspectability of Logic** if, for any decision function $f : X \to Y$ that determines user outcomes, there exists a publicly accessible artifact A_f such that A_f fully specifies the computation of $f(x)$ for all inputs $x \in X$.*

Users who cannot leave a system must possess the ability to inspect it. **Inspectability of Logic** requires that the rules determining outcomes are visible to those they govern. Power in a digital system resides in execution. Lessig [Lessig, 2006] established that software architecture functions as regulation: code "determines what people can or cannot do in the first place," unlike legal regulation which sets rules for behavior and retroactively punishes non-compliance. This is regulation without appeal. If the execution path remains hidden, that power is absolute; it is not merely unaccountable but structurally unobservable to those it governs.

Transparency specifications or open standards are insufficient if the executing artifact remains a black box. **Aadhaar's** CIDR (Central Identities Data Repository) is proprietary: the matching algorithm that determines identity cannot be examined. **UPI** publishes API specifications, but the NPCI switching logic remains closed. **Let's Encrypt** demonstrates the alternative: its CA software is fully open source, allowing anyone to audit the certificate issuance logic that secures web traffic.

4.7 Test 3: AUDIT (Independent Verification)

Definition 4.3 (AUDIT)**.** *A system S satisfies **Independent Verification** if any external party P, without requiring operator authorization, can measure the system's error rate ϵ_S, verify compliance with stated specifications, and document failure modes in production environments. Formally: $\forall P : \textit{Access}(P, \epsilon_S) = \textit{true}$ without $\textit{Authorize}(\textit{Operator}, P)$.*

While inspection enables understanding, **Independent Verification** enables truth. This functions as the sensor in the cybernetic feedback loop. A system is verifiable only if external, permissionless actors can test its behavior, measure its error rates, and document its failure modes within production environments.

Remark 4.1 (The Auditability Distinction)**.** *The framework does not equate auditability with direct participation. AUDIT $\neq$ "everyone audits"; AUDIT $=$ "no one can prevent auditing." **Journalism analogy:** Not everyone investigates corruption. Democracy still depends on the possibility of investigation. The problem today is not that auditors do not exist; it is that operators can legally block them. The tradeoff is contestable expertise over unchallengeable authority, a defensible democratic position.*

Ostrom's design principles require that monitors be accountable to appropriators, not external authorities [Ostrom, 1990]. Applied to DPI: audit capacity must be exercisable by affected communities or their designated agents, not delegated exclusively to operator-approved auditors or captured regulatory bodies.

Intermediary-Discretion Surface. The AUDIT surface extends to every rung of the implementation chain, not only to the system operator. In practice, the layer between the operator and the affected person — ration dealers, enrollment-camp operators, local-office clerks, vendor integration partners — exercises delegated discretion that determines whether the system's stated behavior reaches the citizen. Aggregate error-rate metrics (ϵ_S system-wide) can certify AUDIT as passing for the operator while a particular implementation rung sustains a locally captured regime invisible to system-level measurement. The audit surface must therefore include: (i) per-node or per-agent variance in denial, exception, and manual-override rates, with statistical outlier detection as a mandatory audit output; (ii) a disclosure requirement that the operator enumerate every role holding delegated discretion over subject outcomes, with each role's documented revocation and appeal path; and (iii) Rule A.9 enforcement-history evidence

applied to sanctions against intermediaries, not only against the operator. An audit that measures only the system's aggregate behavior has not measured the layer through which the governed actually encounter it.

Incorrigible systems typically capture this function. **UPI** permits RBI regulatory audits, but transaction-level verification requires NPCI authorization; independent researchers cannot measure failure rates or dispute resolution times. **Proprietary cloud services** (AWS, Azure, GCP) publish compliance certifications while blocking independent verification of infrastructure behavior at scale. **Bitcoin** provides the counter-model: every transaction is independently verifiable by any node, and consensus is cryptographic proof. Similarly, **Let's Encrypt** publishes Certificate Transparency logs, making every issued certificate publicly auditable in real time.

The structural function of opacity is sensor failure. A system whose error rates cannot be independently measured cannot be regulated. The controller receives no signal about divergence from intended behavior. Under Ashby's Law, such a system will drift until external forces impose correction. Those forces, operating outside the technical loop, will impose correction through channels the system cannot anticipate or absorb. Pasquale [2015] describes this as a "one-way mirror" relationship, but the cybernetic translation is precise: asymmetric observability breaks the feedback loop at the sensing stage.

4.8 Test 4: GOVERN (Constitutive Constraint)

Definition 4.4 (GOVERN)**.** *A system S satisfies **Constitutive Constraint** if there exists a governance function $G : Rules \to Rules'$ such that (1) G is accessible to affected parties, not solely the operator; (2) G is binding, meaning the operator cannot override G's output; and (3) there exists a cryptographic or legal chain of custody linking system behavior to G's constraints.*

Verification identifies the error, while **Constitutive Constraint** enforces the correction. This framework distinguishes structural governance from administrative functions like consultation, multi-stakeholder meetings, or feedback forms.

Governance exists only when binding constraints limit system behavior in ways the operator cannot override. **Aadhaar** is governed by UIDAI fiat: the same entity that operates the system sets its rules. **UPI** has NPCI governance, but NPCI is majority-owned by participating banks, creating structural conflicts. **Let's Encrypt** shows what binding constraint looks like: the ISRG charter legally binds the organization to its public benefit mission, with community board representation. **Bitcoin's** BIP process demonstrates governance through proven fork history; contentious changes have been rejected by the network, proving the community can override developer intent. **PostgreSQL's** permissive license and distributed core-team governance — no single corporate copyright holder, no contributor license agreement — prevent any single entity from capturing the project.

4.8.1 The Post-Execution Fallacy

A critical failure in modern policy involves relying on exogenous or post-hoc remedies to substitute for structural constraints. Mechanisms such as courts, ombudsmen, and grievance redressal officers operate on "bureaucratic time," measured in months or years. Infrastructure operates on "digital time," measured in milliseconds. A system that executes a harmful action, such as wrongful deletion of a beneficiary, and relies on a court order to reverse it six months later is structurally ungoverned for that duration. True governance must function within the system's execution loop to block prohibited states before they manifest.

Consequently, claims of governance must be evidentiary rather than declarative. They require a signed cryptographic chain of custody linking the operator's authority to an external, non-overrideable instrument.

4.8.2 Strategic Openness vs Structural Accountability

Open artifacts do not guarantee open governance. Systems can deploy *asymmetric openness*: releasing code or weights for ecosystem adoption while retaining governance control. This pattern (detailed in Section 5) satisfies CODE while failing GOVERN. The GOVERN test requires bidirectional accountability: not merely that citizens can observe the system, but that they can correct it. A system that publishes its code while reserving all governance decisions to the operator has satisfied CODE but failed GOVERN.

A related confusion conflates governing one's own data with governing the system itself. Data protection frameworks (GDPR, CCPA) grant users rights to access, correct, delete, and port their data. These rights are necessary but insufficient: they allow users to manage information within rules the operator defines, not to change those rules. The GOVERN test asks: can affected parties change the rules, not merely operate within them?

4.9 Test 5: FORK (Independent Reproduction)

Definition 4.5 (FORK). *A system S satisfies **Independent Reproduction** if an independent party can instantiate a functionally equivalent system S' such that (1) no legal prohibition prevents S''s deployment; (2) the artifacts required for reproduction (code, schemas, protocols) are publicly available; and (3) user state U_S is portable: $U_S \to U_{S'}$. For deterministic infrastructure, natural barriers (capital, network effects) do not disqualify FORK; only constructed barriers (legal prohibition, regulatory monopoly, deliberate technical enclosure) do. For learned systems, the companion paper shows that reproduction cost itself enters the test: resource asymmetry beyond a policy threshold (Compute and Data Capture) fails FORK regardless of licensing terms [Aravind, 2026].*

The ultimate check on power is the ability to recreate it. **Independent Reproduction** determines whether the ecosystem allows the infrastructure itself to be separated from its current steward. This requirement distinguishes structural resilience from simple market competition.

Remark 4.2 (Credible Replaceability). *The paper does not define FORK as "routine divergence in production." It defines FORK as **credible replaceability in extremis**. **Constitutional analogy:** The US Constitution is forkable (amendable, secession historically possible). That does not mean every county runs a different constitution. The possibility disciplines the center. **The stable equilibrium:** Latent forkability + strong coordination incentives. Email works not because everyone forks SMTP, but because anyone could. **Key asymmetry:** FORK is a threat, not a preference. No one wants fragmentation. But irreversibility without exit is worse.*

The existence of competitors, such as choosing between two proprietary cloud providers, provides an "Exit" option but not a "Fork." Shifting providers allows a user to escape a specific operator, but it forces them to abandon their history, identity, and operational logic. This action is substitution, not reproduction.

4.9.1 Natural vs Constructed Barriers

A critical distinction must be drawn between barriers to forking that are **natural** (arising from coordination costs, network effects, or capital requirements) and those that are **constructed** (arising from legal prohibition, regulatory monopoly, or deliberate technical enclosure).

- **Natural barriers**: Gmail dominates email not because forking SMTP is illegal, but because Google invested in spam filtering and UX. Anyone *can* run their own mail server. Network effects and capital create friction, but not prohibition.
- **Constructed barriers**: NPCI is the *only permitted* instant payment switch in India [Reserve Bank of India, 2026]. The barrier is regulatory, not economic. Even with infinite capital, you cannot legally compete with UPI's core switching function.
- **Grant-conditioned barriers**: Sri Lanka's MOSIP implementation is contractually restricted to donor-country system integrators, despite MOSIP's open-source license [Biometric Update, 2025]. The barrier is neither regulatory nor technical but financial: grant conditions foreclose supplier choice, demonstrating how "open" DPI exports can create new dependencies.

The FORK test fails only when constructed barriers prevent reproduction. Natural barriers reduce the *likelihood* of forking; constructed barriers eliminate its *possibility*. A system with high natural barriers but no constructed barriers (Linux, email) remains structurally corrigible. A system with low natural barriers but constructed prohibition (hypothetically: open-source software banned by law) would fail FORK despite technical simplicity.

4.9.2 State Portability as a Precondition for Fork

A critical structural requirement: code forkability is insufficient if **user state** is non-transferable. In traditional software (Linux, PostgreSQL), the artifact being forked *is* the system. In Digital Public Infrastructure, the artifact is merely the skeleton; the system includes the user's credentials, transaction history, social graph, and accumulated trust relationships.

Let U_S denote the user state graph in system S, and S' a forked alternative. A fork is **functionally meaningful** only if:

$$U_S \to U_{S'} \quad \text{(state portability)} \tag{2}$$

Open code with locked data produces a hollow fork. The legal right to reproduce an identity system is meaningless if citizens cannot port their credentials, attestations, and service eligibility to the new instance.

Switching Cost and Portability. The practical barrier to exit and fork can be assessed through **switching cost**: the aggregate friction of data export, service downtime, credential re-establishment, network loss, and legal barriers. When switching cost to all plausible alternatives exceeds a policy threshold, EXIT and FORK functionally fail regardless of formal rights. Aadhaar exhibits near-maximal switching cost: no data export exists, service denial during transition is total, and legal mandate forecloses alternatives. UPI retains lower switching cost because cash and cards remain legal. The formal operationalization appears in Appendix B.

Technical Requirements for Portability. To ensure meaningful forkability, DPI systems must implement: (1) standardized export bundles with user-owned credentials, (2) interoperability APIs to accept ported data, and (3) statutory guarantees requiring public services to accept verified portable credentials. This is the logic of PSD2 in banking: portability mandates were intended to make bank-switching economically viable.

A system passes the FORK test only if the community possesses the legal rights, technical artifacts, economic means, **and state portability guarantees** to spawn a functionally equivalent instance. License choice determines whether forkability persists across derivatives: copyleft licenses (GPL, AGPL) ensure modifications remain forkable; permissive licenses (MIT, Apache) allow derivatives to be enclosed. For network-deployed DPI, AGPL-style requirements prevent operators from capturing modified code behind service interfaces. **Aadhaar** fails completely: no entity outside UIDAI can reproduce the identity infrastructure. **OpenSearch** demonstrates what successful forking looks like. When Elastic changed its license, AWS forked Elasticsearch and the community continued development independently. **Linux** has been forked thousands of times (Android, Chrome OS, countless distributions). **AWS S3** presents partial forkability: the API specification is public and MinIO provides a compatible implementation, but the operational scale, global infrastructure, and integration ecosystem cannot be reproduced by most organizations. Table 1 illustrates the distinction across representative cases: substitution (EXIT) and reproduction (FORK) are independent properties, and a system can satisfy one while failing the other.

System	EXIT	FORK	Structural Reality
Cloudflare	PASS	FAIL	Can switch to Fastly, cannot reproduce
AWS EC2	PASS	FAIL	Can migrate to Azure, logic/state locked
Aadhaar	FAIL	FAIL	Cannot refuse, cannot reproduce
Linux	PASS	PASS	Can leave, can fork
AWS S3	PASS	PARTIAL	API open (MinIO), scale unreproducible

Table 1: Differentiation between EXIT (substitution) and FORK (reproduction). EXIT requires only that an alternative service exists; FORK requires that the system itself can be independently reproduced. Many systems satisfy EXIT through substitution while failing FORK through irreproducibility, and the distinction matters under centralized-vendor lock-in.

4.10 Contestation of Representational Categories in AI-Mediated DPI

The five tests presented above were derived for deterministic infrastructure: ledgers, registries, and rule-based engines. As DPI increasingly incorporates learned components, the GOVERN test acquires a second contestation surface that the deterministic case does not require. This subsection identifies that surface; the formal constraints on the learned-system side are developed in the companion paper [Aravind, 2026].

When learned systems are deployed within DPI, the categories they use to classify citizens ("eligibility," "risk," "fraud," "vulnerability," "priority") cease to be mere statistical properties of a model. They become **consequential administrative determinations**. A citizen flagged as "high-risk" by an AI-mediated welfare system experiences that label as legal fact: payments suspended, services revoked, additional verification demanded. The label has the same operational force as a determination from a deterministic registry, even though it emerged from a probabilistic computation over an undisclosed feature space.

The structural problem is that affected communities must possess the capacity to *review and contest the representational categories themselves*, not just dispute individual execution decisions. Representation in AI-mediated DPI differs fundamentally from representation in deterministic administrative systems: the categories are **emergent rather than deliberately designed**. In a deterministic registry, "eligibility" is defined by a rule the legislature wrote; the contestation surface is the rule's text. In a learned system, "eligibility" is defined by the joint distribution of training data, loss function, and inference pipeline; the contestation surface is the schema, the data, and the model's operative representation simultaneously.

Three structural consequences follow:

1. **Per-decision appeal is necessary but insufficient.** A citizen who successfully appeals one classification has not contested the category that produced it. The next million decisions will be drawn from the same category. Adequate contestation requires standing to challenge the schema, not just the application of the rule.
2. **Disclosure obligations extend to the operative representation.** Publishing model weights does not satisfy GOVERN if the categorical structure that emerges at inference time is undisclosed. The state retains nominal authority over the rule but loses substantive authority over the categories that animate it.
3. **Standing must include the affected class, not only the operator.** Because the schema captures the population's classification surface, contestation rights cannot be confined to procurement disputes between the state and the vendor. Affected populations must hold standing to challenge the representational schema directly.

Remark 4.3 (Extension of GOVERN to Ontological Choices)**.** *The GOVERN test extends to ontological choices when those choices produce binding classifications. An AI-mediated DPI deployment satisfies GOVERN only if affected populations can challenge the categorical schema, not merely dispute individual outputs. The formal constraints required to render this contestation operative (LWD-R disclosure, operative-representation transparency, action-boundary auditability) are developed in the companion paper's analysis of learned-system corrigibility [Aravind, 2026].*

Collective-Standing Precondition (Ostrom Principle 7). Ostrom's seventh design principle, the minimal recognition of the right to organize, is a precondition for the mechanism families through which GOVERN operates at population scale: class-action vehicles, deliberative forums, and juridical standing all presuppose that affected parties can constitute themselves as a collective actor, with rights to associate, aggregate claims, and fund representation. The framework's GOVERN test verifies the channel; Principle 7 verifies whether the subject that uses the channel can be constituted. Where the right to organize is suppressed — by association law, administrative harassment, cost barriers that foreclose legal aggregation, or structural atomization — the formal availability of GOVERN mechanisms does not translate into operative corrective authority, and the loop is open in fact even when the channel is structurally present. GOVERN determinations therefore carry the implicit precondition that this associational capacity is legally protected and historically exercised at the relevant stratum (Rule A.9 evidence standard). Where this precondition fails, the audit must record GOVERN as failing at that stratum and document the associational constraint as the cause. Individual grievance pathways are not substitutes; they address individual cases without generating the collective error signal that governance correction requires.

This bridge clause completes the deterministic framework's account of the GOVERN test for the AI era: the test does not change, but its evidentiary burden does. The companion paper takes up the verification machinery in detail and develops a strict-interpretability firewall: when the operative representation of a learned component cannot be disclosed to the standard required for auditable reproduction, the system fails CODE for high-stakes deployments and should not, consistent with this framework's legitimacy criteria, be adopted as Epistemic Public Infrastructure in rights-affecting tiers [Aravind, 2026].

4.11 The Topology of Necessity

Five tests describe the anatomy of a stable control loop rather than a random assortment of normative virtues (see Figure 2).

- **EXIT** provides the Error Signal.
- **AUDIT** acts as the Sensor.
- **CODE** ensures Transmission Clarity.
- **GOVERN** serves as the Actuator.
- **FORK** provides Selection Pressure over the entire loop.

Failure in any single dimension breaks the control loop at a specific physical point. The result is not partial infrastructure but **Open Loop Control**. An open-loop system executes directives without the capacity to sense deviation, interpret error, or apply corrective force.

Failure-Dominated Structure. Corrigibility is failure-dominated. If any single constraint collapses, corrective authority becomes contingent rather than guaranteed. This weakest-link structure mirrors distributed systems reliability: system safety is bounded by the least reliable component. Under human variety and environmental change, open-loop control drifts toward failure.

Corrigibility closes the loop. Only when all five tests are satisfied does a system possess a complete feedback topology capable of detecting divergence, transmitting intelligible signals, executing correction, and replacing failed control logic when necessary. Without this closure, error correction is structurally discretionary, and bounded-error operation cannot be guaranteed.

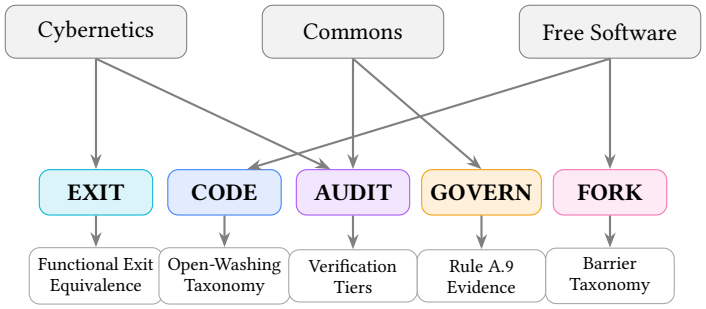

Figure 3: The framework at a glance. Three independent theoretical traditions (top) derive the five corrigibility tests (middle), and the tests in turn derive their verification instruments (bottom): Functional Exit Equivalence (Section 6.7.2), the open-washing taxonomy, verification tiers, the Rule A.9 evidence standard (Appendix D), and the constructed-versus-natural barrier taxonomy, each shown under the test it most directly serves. Determination is evaluated at the least-resourced stratum (Remark B.1) and in both the inward and outward exercise of each test.

4.12 Structural Symmetry of Power

The five tests, while derived from distinct intellectual traditions (Figure 3), share a common organizing principle. Each enforces a constraint on power concentration at a different layer of the system. Together, they form a single power invariant:

Layer	Test	Function (Cybernetic)
Participation	EXIT	Reversibility of participation
Observability	CODE	Inspectability of rules
Observability	AUDIT	Verifiability of error
Constraint	GOVERN	Constraint on controller variance
Replacement	FORK	Selection pressure via reproduction

Each test enforces the same principle: *No actor exercising power may be the sole judge of its continuation, scope, or correction.*

A system without EXIT is coercive. A system without GOVERN is arbitrary. A system without FORK is captive.

4.13 The Dual Exercise of the Tests

Each test admits two exercises, distinguished by who holds the corrective capacity. The **inward exercise** belongs to parties at or inside the operator's institutional perimeter — the operator's own controls, its auditors, its regulator. The **outward exercise** belongs to the subject the system decides about: the participant whose error signal the loop exists to carry. The five definitions in this section are stated in the outward voice — refusal, inspection, verification, constraint, and reproduction are capacities of affected participants. The verification apparatus that discharges them (manifests, audit rules, evidence standards; Appendices A and D) operates largely in the inward voice. The gap between the two voices is where corrigibility claims most often fail in practice: a deployment can satisfy every test through inward instruments and remain unreachable from the outside.

The columns are not alternatives, and the inward column is not a deficiency. Enforcement machinery, independent audit, and regulatory teeth are how correction is executed, and nothing reaches the subject without them. The failure mode is substitution, treating the inward column as the whole of the test. Consider a system whose operator can halt it, whose auditors can verify it, and whose regulator can bind it, while its subjects cannot refuse it, learn the rule that decided their case, confirm the record, contest the category, or carry their state elsewhere. That system is governed by

Table 2: The dual exercise of each test. Inward instruments enforce; outward instruments carry the error signal the enforcement exists to serve. Under the weakest-link principle, a test discharged only inward has been verified for the operator, not for the governed — the gradient quantifier (Remark B.1) restated at the instrument level.

Test	Inward exercise (operator, auditor, regulator)	Outward exercise (the governed)
EXIT	Halt, suspension, deprovisioning of the system's own components	Refusal without disproportionate penalty (Definition 4.1); Functional Exit Equivalence for essential services (Section 6.7.2)
CODE	Source access for audit teams; reproducible builds	Rules legible to the person they decide: the operative rule in a subject's case is identifiable and stated
AUDIT	Permissionless verification by external parties (Rule A.10 chain of custody)	Notice that a decision occurred, and a record the subject can verify against the lived event [Aravind, 2026], under purpose-bound data duties (Rule A.11)
GOVERN	Binding charter with enforcement history (Rule A.9)	Contestation with binding effect, including the categorical schema (Section 4.10); collective standing as precondition
FORK	Legal rights, artifacts, and resources for independent reproduction	State portability of the subject's own credentials and history (Equation 2); acceptance of the ported record

its operator and incorrigible to the people it operates on. The inward column is how the operator governs the system. The outward column is how the governed correct the operator. Corrigibility is the second, standing on the first.

One deployed mechanism already exercises a test in the outward voice at population scale, and the concession is stated here because it sharpens the framework rather than weakening it. In the digital-wallet paradigm, credential presentation executes through the subject: the holder determines when, where, and how credentials are shared, and refusal-to-present is a subject-side act that operates ex ante, at machine speed, inside the execution loop [World Bank, 2026c]. In-loop subject authority is therefore demonstrated as buildable at population scale for exactly one verb, disclose. No deployed system extends the same architectural position to contest, correct, revoke, or halt. The demonstration converts each remaining outward omission from a claim of technical infeasibility into an architectural choice, and the Post-Execution Fallacy (Section 4.8.1) should be read with this concession attached: the operator's objection that in-loop subject authority cannot be implemented at scale is no longer available.

4.14 Layer-Decomposed Evaluation

Complex infrastructure frequently stratifies into distinct functional layers, each with independent governance, transparency, and reproducibility properties. In agentic deployments the stratification acquires a second axis, delegation depth, where one system's harness contains another's; the companion paper extends this decomposition to nested delegation [Aravind, 2026]. The five tests must be applied to each layer separately when:

1. Different entities control different layers (protocol versus implementation versus trust framework)
2. Transparency varies across layers (open protocol, closed switching logic)
3. Forkability differs by layer (code is forkable, network effects are not)

4.14.1 Propagation Rule

A system that passes a test at one layer but fails at another does not achieve "partial" compliance. It fails the test. The evaluation proceeds from the layer closest to user impact upward; failure at any layer propagates to the whole.

Formally: Let $L_1, L_2, \ldots, L_n$ be the layers of a system, ordered by proximity to user impact. For test T:

$$T_{\text{system}} = \bigwedge_{i=1}^{n} T(L_i) \tag{3}$$

The system passes T if and only if every layer passes T. The full corrigibility determination extends this across all five tests:

$$\text{Corrigible}(S) = \bigwedge_{t \in \mathcal{T}} \bigwedge_{i=1}^{n} T_t(L_i) \quad \text{where } \mathcal{T} = \{\text{EXIT, CODE, AUDIT, GOVERN, FORK}\} \tag{4}$$

This double conjunction formalizes the weakest-link principle: failure of any test at any layer propagates to the system determination.

4.14.2 Application to Identity Infrastructure

Table 3 illustrates how layer decomposition applies to identity infrastructure: distinct architectural strata exhibit characteristically different failure modes that cannot be aggregated into a single system-wide judgment.

Layer	Example Components	Typical Failure Mode
Protocol	W3C DID, VC Data Model	Often passes CODE
Implementation	Specific resolver, wallet software	May fail CODE (proprietary)
Credential	Issuer policies, attribute schemas	Often fails GOVERN (unilateral issuer)
Trust	Relying party acceptance, trusted lists	Often fails FORK (cannot reproduce acceptance)

Table 3: Identity infrastructure layer decomposition by vertical stratum. The DID-specific functional decomposition of the same stack is given in Table 7.

The UPI payment rail exhibits similar layering: the protocol specification (NPCI-published) passes CODE at the protocol layer, but the switching logic (NPCI-operated) fails CODE at the implementation layer. Since the implementation layer is closer to user impact, the system fails CODE.

4.14.3 The Weakest-Layer Principle

This propagation rule has a counterintuitive but structurally necessary consequence: a system's corrigibility status equals its weakest layer across any test dimension. Strength at one layer cannot compensate for failure at another.

This principle prevents a common evasion: claiming corrigibility by publishing peripheral artifacts while retaining control over the layers that determine outcomes. Open-washing (Section 5.4) typically exploits layer confusion. Releasing SDKs (passing CODE at the interface layer) while keeping core logic proprietary (failing CODE at the execution layer) does not produce partial corrigibility. It produces total failure at the execution layer, which propagates to the system determination.

5 Empirical Evaluation

The preceding sections established what corrigibility means and why it matters. Theory, however, must confront practice. Before analyzing *why* systems fail these tests (the political economy, Section 6), we first demonstrate *that* they fail and document the human consequences. This section applies the five tests to real-world infrastructure, establishing the concrete stakes that motivate the analysis to follow. (The stress test of the same framework against learned systems is the companion paper [Aravind, 2026].)

To demonstrate the framework's application, we evaluate systems across three categories: government infrastructure promoted as "DPI," platform infrastructure claiming "openness," and infrastructure that satisfies all five tests.

A Note on Utility. These evaluations assess structural accountability, not functional utility. A system can be simultaneously useful and incorrigible. Aadhaar may enable efficient service delivery; this does not establish that it permits correction by those it affects. The framework evaluates governance architecture, not operational performance. Utility and accountability are orthogonal dimensions: neither implies the other.

5.1 Government Infrastructure: The Trap of Partial Compliance

Government-operated systems currently designated as Digital Public Infrastructure frequently exhibit a dangerous failure mode: they achieve partial compliance on technical tests (EXIT, CODE, AUDIT) while failing structurally on governance tests (GOVERN, FORK). By capturing the monopoly on execution, they render the feedback loop inoperable.

Table 4: **Evaluation of Government Infrastructure.** Centralized execution leads to consistent structural failure despite "open standards" rhetoric. PARTIAL is a diagnostic annotation only; in the binary determination it resolves to FAIL under the weakest-link principle (Equation 4, Section 7.1).

(a) Aadhaar (India): Mandatory ID

Test	Result	Evidence
EXIT	×	Disproportionate penalty: statutory PDS linkage; de facto banking/SIM insistence post-*Puttaswamy II* [Supreme Court of India, 2018, Khera, 2019].
CODE	×	Execution logic opaque: proprietary matcher, closed algorithms.
AUDIT	×	External testing barred: error measurement prohibited without UIDAI authorization [Parliament of India, 2016].
GOVERN	×	No binding constraint: UIDAI is both operator and regulator.
FORK	×	No legal/technical/economic means: total centralized capture.

(b) UPI (India): Payment Rail

Test	Result	Evidence
EXIT	PARTIAL	Cash exists but network effects coerce.
CODE	PARTIAL	Protocol specs open, switching logic closed.
AUDIT	PARTIAL	Regulator audits, public measurement blocked.
GOVERN	×	Banks own NPCI, regulate themselves.
FORK	×	Hub-and-spoke requires NPCI cooperation.

(c) Pix (Brazil): Central Bank Payment

Test	Result	Evidence
EXIT	PARTIAL	Cash/cards exist; digital-only services coerce.
CODE	PARTIAL	API spec public; central routing closed.
AUDIT	PARTIAL	Stats public; independent measurement blocked.
GOVERN	×	BCB sets/enforces rules unilaterally.
FORK	×	Central switch requires BCB.

(d) X-Road (Estonia/Nordic): Data Exchange Layer

Test	Result	Evidence
EXIT	×	Exclusion from government services.
CODE	✓	MIT source on GitHub.
AUDIT	✓	Transaction logs, documented APIs.
GOVERN	✓	NIIS treaty, MIT license irrevocable.
FORK	✓	20+ country deployments prove feasibility.

(e) UK Open Banking: Regulated API Standard

Test	Result	Evidence
EXIT	PARTIAL	Can revoke API consent; cannot exit banking.
CODE	×	Bank implementations proprietary.
AUDIT	✓	FCA oversight, TPP certification.
GOVERN	✓	CMA order, PSD2 regulation.
FORK	PARTIAL	Rights exist; means require regulatory approval.

Pattern: Pix (Brazil) illustrates how measurement frameworks mask structural failure. Global DPI assessments score Pix as the highest-aligned payment system in Latin America, meeting all eight payment-system variables spanning the six measured attributes [Fetter et al., 2025]. Yet Pix fails both GOVERN (Banco Central do Brasil sets and enforces rules unilaterally, with no binding citizen-corrective mechanism) and FORK (the central switch cannot be reproduced without BCB authorization). The variables measured—interoperability policy, participation conditions, oversight body—capture *presence*, not *function*. A citizen wrongly excluded from Pix has no structural path to correction beyond petitioning the same authority that excluded them. High measurement scores coexist with closed corrective loops.

Other government systems exhibit different failure patterns. X-Road (4/5) fails only EXIT: citizens cannot opt out of government data exchange, but the system passes CODE, AUDIT, GOVERN, and FORK [Nordic Institute for Interoperability Solutions, 2026]. UK Open Banking (2/5) passes AUDIT and GOVERN but fails CODE: the API *specification* is public, but the bank *implementations* are proprietary. Specification $\neq$ Code. You can see the interface; you cannot inspect the execution.

5.2 Platform Infrastructure: The Illusion of Openness

In the private sector, systems often leverage "open source" branding while retaining centralized control. This decoupling of artifact transparency from governance reality creates a distinct failure pattern: the artifacts are public; the decisions are not. Table 5 applies the five tests across cloud, mobile-platform, and messaging examples to make the pattern visible.

Table 5: **Evaluation of Private/Platform Infrastructure.** Technical openness (Source Code/APIs) does not guarantee structural accountability. PARTIAL is a diagnostic annotation; in the binary determination it resolves to FAIL (Equation 4).

(a) Cloud Infrastructure (AWS, Azure, GCP)

Test	Result	Evidence
EXIT	✓	Can migrate workloads to competing providers.
CODE	PARTIAL	APIs documented; internal routing/pricing opaque.
AUDIT	PARTIAL	Billing auditable; infrastructure logic unverifiable.
GOVERN	×	Provider sets terms, regions, compliance unilaterally.
FORK	×	Scale, global infrastructure unreproducible by states.

(b) Android (AOSP + GMS)

Test	Result	Evidence
EXIT	PARTIAL	Device portable; ecosystem lock-in creates friction.
CODE	PARTIAL	AOSP open; Play Services/Identity closed.
AUDIT	PARTIAL	OS auditable; Google Services opaque.
GOVERN	×	Google sets roadmap/certification unilaterally.
FORK	PARTIAL	LineageOS exists; app ecosystem requires GMS.

(c) Signal Protocol

Test	Result	Evidence
EXIT	✓	Can switch to alternatives freely.
CODE	✓	Client and server source available (GPL/AGPL).
AUDIT	✓	Cryptographic properties independently verifiable.
GOVERN	×	Foundation board sets direction unilaterally.
FORK	×	State non-portable: no federation or graph export, so $U_S \not\to U_{S'}$ fails clause (3).

Pattern: Asymmetric Openness. These systems deploy strategic openness: releasing artifacts for ecosystem capture while retaining governance. This satisfies CODE (inspection possible) while failing GOVERN (correction blocked). The framework distinguishes artifact transparency (can you see the code?) from governance accountability (can you correct the system?). Open weights, open protocols, and open standards can each mask closed correction loops. "Open weights" $\neq$ "open." "Open source" $\neq$ "accountable." The artifact is open. The correction loop is closed. Seeing the machine does not govern it.

5.3 Corrigible Infrastructure

For contrast, Table 6 summarizes eighteen systems that pass all five tests. A striking pattern emerges: these systems are predominantly non-essential infrastructure. No individual's survival depends on accessing Linux or Let's Encrypt.

This correlation between corrigibility and non-essentiality is not coincidental; it reflects a structural tension analyzed later in this paper.

System	Steward	Score	Why Corrigible
Linux Kernel	Linux Foundation	**5/5**	Fully reproducible. Full triad (legal, technical, economic) exists. Independent distributions prove forkability.
Let's Encrypt	ISRG	**5/5**	Open source CA, IETF standards, nonprofit 501(c)(3)
Wikipedia	Wikimedia Foundation	**5/5**	CC BY-SA content, public edit history, community RfC governance
Matrix Protocol	Matrix.org Foundation	**5/5**	Apache 2.0, federated by design, self-hostable
Bluesky (AT Protocol)	Bluesky PBC	**5/5**	MIT/Apache; portable DID identity and self-hostable PDS make the fork threat credible, which is what binds PBC direction-setting (contrast Signal's FORK fail: state non-portability, so no latent-fork discipline)
PostgreSQL	PGDG	**5/5**	BSD-like license, public mailing lists, major forks exist
IPFS	Protocol Labs	**5/5**	MIT/Apache, public IPIP process, no token governance
Bitcoin	Decentralized	**5/5**	MIT source, public BIP process, proven forks (BCH, BSV)
Kubernetes	CNCF	**5/5**	Apache 2.0, KEP process, distributions (OpenShift, k3s)
Firefox	Mozilla Foundation	**5/5**	MPL 2.0, nonprofit manifesto, forks (LibreWolf, Tor Browser)
Apache HTTP	Apache Foundation	**5/5**	Apache 2.0 license, ASF governance, powers a substantial share of the web
Apache Kafka	Apache Foundation	**5/5**	Apache 2.0, KIP process public, Fortune 500 standard
OpenSearch	Linux Foundation	**5/5**	Forked from Elasticsearch; Apache 2.0, multi-vendor governance
Valkey	Linux Foundation	**5/5**	Forked from Redis; BSD license, community governance restored
Hyperledger	LF Decentralized Trust	**5/5**	Apache 2.0 framework; institutional deployments require separate evaluation
LibreOffice	Document Foundation	**5/5**	Forked from OpenOffice; LGPL/MPL, nonprofit governance
MariaDB	MariaDB Foundation	**5/5**	Forked from MySQL; GPL, foundation governance
Eclipse IDE	Eclipse Foundation	**5/5**	EPL 2.0, 380+ member organizations, committer-led governance

Table 6: Infrastructure systems satisfying all five corrigibility tests. Scores are screening-tier determinations (Tier 1–2: presence and behavior); certification-tier GOVERN verdicts additionally require the Rule A.9 enforcement-history evidence, which Bitcoin's contentious-fork record and Kubernetes's multi-vendor overrides exemplify and which several rows have not yet accumulated. The set is predominantly non-essential infrastructure, the pattern the essentiality analysis of Section 6.7 explains.

Fork as Correction: OpenSearch, Valkey, LibreOffice, and MariaDB demonstrate Ashby's Law in practice. Each emerged when governance variety collapsed: Oracle's Sun acquisition (2010) triggered LibreOffice and MariaDB; unilateral license changes triggered OpenSearch (2021) and Valkey (2024). The ecosystem restored requisite variety through forking.

The framework is governance-agnostic: corporate participation is not disqualifying. Kubernetes (CNCF), Hyperledger, and Let's Encrypt all have significant corporate involvement while passing all tests. The test is whether governance permits correction, not who participates.

License vs. Governance: Redis Ltd. added AGPLv3 in 2025, restoring open licensing. But license is orthogonal to governance. *License* determines what you may do with code; *governance* determines who decides what it becomes. Redis passes one; Valkey passes both.

International Counterexample: Estonia's X-Road. Estonia's X-Road architecture provides a counterexample demonstrating that scale does not necessitate centralized capture. Its federated data exchange model distributes control across institutional nodes, preserving CODE transparency (fully open-source, MIT license) and enabling institutional FORK at the implementation layer. While EXIT remains limited for certain core registries, the federated topology significantly reduces concentration coupling compared to monolithic architectures. X-Road has been adopted by Finland and Iceland, and deployed in more than twenty-five countries (Table 4), illustrating that corrigibility properties are architectural rather than cultural. The framework does not require cultural homogeneity or small population size; it requires federated governance that prevents single-point capture. (X-Road's full determination remains a fail on EXIT — Section 6.7.3; it appears here as the strongest near-miss among evaluated government systems, not as a pass.)

5.4 The Taxonomy of Open-Washing

Systems that fail these structural tests frequently employ vocabulary to obscure their closure. The recurring patterns reduce to three strategic categories, each instantiated by several named tactics:

1. **Symbolic-openness washing.** Substituting technical or diplomatic markers for governance accountability. Instances include *Standard-Washing* (adopting open technical standards such as ISO/W3C as proxy for governance accountability), *Open-Washing* (releasing peripheral SDKs while keeping core execution logic proprietary), *Multilateral-Washing* (using G20/World Bank endorsements to substitute for technical auditability), and *DID-Washing* (deploying decentralized identifiers without decentralized governance; see Section 5.4.1).
2. **Coerced-legitimacy washing.** Grounding legitimacy in security narratives or coerced consent. Instances include *Safety-Washing* (justifying opacity by claiming inspection poses a security risk) and *Consent-Washing* (obtaining formal consent under conditions of economic coercion).
3. **Substantive-substitution washing.** Invoking orthogonal public goods to legitimize structural changes that extend beyond the stated purpose. The principal instance is *Climate-Washing* (using environmental rhetoric to legitimize movement, payment, or surveillance changes; see Section 5.4.2).

5.4.1 DID-Washing: The Layer Confusion Problem

Decentralized Identifiers (DIDs) present a distinct challenge to corrigibility evaluation because they separate technical decentralization from governance decentralization across multiple architectural layers. The W3C DID Core 1.0 specification [World Wide Web Consortium, 2022] defines DIDs as "globally unique identifiers" that "do not require a centralized registration authority." This technical claim is accurate. However, a system can implement DIDs at the identifier layer while retaining centralized control at every layer that determines actual utility.

Identity infrastructure comprises at least four functional layers, each with independent governance (Table 7):

Table 7: DID infrastructure layers and typical corrigibility failures, by function. The vertical-strata decomposition of the identity stack is Table 3.

Layer	Function	Common Failure
Identifier	DID generation, resolution	Often passes CODE (open specs)
Credential	Issuer policies, attribute schemas	Often fails GOVERN (unilateral issuer)
Wallet	Storage, presentation, consent	Often fails FORK (vendor lock-in)
Reliance	Verifier acceptance, trust frameworks	Often fails EXIT (no alternative verifiers)

A system may achieve decentralization at the identifier layer while remaining fully centralized at the credential, wallet, or reliance layers. Evaluating only the DID specification produces a false positive; evaluating the complete stack reveals the actual governance topology.

Example: EU Digital Identity Wallet. Even where a wallet framework adopts decentralized identifiers at the identifier layer, the layers above it can remain fully centralized. In the EU Digital Identity Wallet framework [European Union, 2024] — whose architecture reference framework centers certified wallet applications and trusted lists rather than decentralized trust anchors — the credential layer is governed by member state authorities with unilateral issuance power; the wallet layer specifies certified applications with vendor approval requirements; and the reliance layer mandates acceptance by specified service providers. Whatever the identifier layer's degree of decentralization, the governance layer is centralized state control mediated through a federated but non-forkable architecture.

The framework's response: evaluate each layer independently, then apply the Weakest-Layer Principle. A system's corrigibility equals its weakest layer. DID-washing exploits the gap between technical specification and governance reality.

5.4.2 Climate-Washing: Environmental Rhetoric as Capture Vector

Climate policy has become a vector for infrastructure capture. The rhetorical pattern is consistent: invoke environmental necessity to justify surveillance, mandate digital-only interfaces, or eliminate cash. This section documents specific instances where climate framing masks structural changes that the framework identifies as incorrigibility failures.

Movement Legibility. Traffic-management schemes enforced through automatic number-plate recognition make private movement machine-legible as a side effect of congestion or emissions policy: the camera network built to fine through-journeys is, structurally, movement-tracking infrastructure whose retention, access, and reuse rules are set by the operator. The climate framing (reduce car trips) can obscure the governance change (movement becomes recorded and reviewable). Under the framework, the question is not the traffic policy but the corrigibility of the legibility infrastructure it installs: who can audit the retention, and who can contest a record.

Cash Elimination. Cash-reduction advocacy sometimes invokes the carbon footprint of physical money. Whatever the environmental accounting, the governance effect is uncontested: elimination of cash creates mandatory digital payment infrastructure with no EXIT option.

Carbon Tracking. Personal carbon accounting systems propose linking identity infrastructure to consumption tracking. The stated purpose is behavioral nudging for sustainability. The structural effect is comprehensive consumption surveillance with no technical or legal mechanism for opting out while maintaining normal economic participation.

The framework's response: environmental goals do not constitute exemptions from corrigibility requirements. A surveillance system does not become corrigible because its stated purpose is ecological. The tests evaluate architecture, not intentions.

6 The Political Economy of Digital Infrastructure

The empirical evaluation reveals a troubling pattern: systems that serve essential functions consistently fail the five tests, while systems that pass are predominantly non-essential. This section analyzes *why* this pattern exists. The answer lies in the political economy of infrastructure: the temporal dynamics of correction, the structural tension between essentiality and accountability, and the sovereignty–scale–neutrality tension facing states that attempt to build sovereign digital infrastructure at scale.

6.1 The Dynamics of Incorrigibility

The five tests define the static architecture of a corrigible system. However, infrastructure exists in time. When we apply these structural requirements to dynamic systems, three fundamental implications emerge regarding stability, energy conservation, and the rule of law. These principles explain why centralized grievance redressal mechanisms (the standard answer of bureaucracies to error) are mathematically insufficient to guarantee stability.

6.2 The Harm Accumulation Principle

A critical clarification is required regarding the framework's claims about incorrigible systems:

> **Principle (Harm Accumulation):** Incorrigible systems accumulate harm until correction becomes catastrophic rather than incremental. The framework does NOT claim "incorrigibility leads to collapse." It claims that **corrigibility is a precondition for a system to be meaningfully public.**

This is ontological classification, not historical prediction. North Korea is stable, incorrigible, and not public in any meaningful sense, confirming rather than refuting the framework. Tyrannies can persist indefinitely while crushing their subjects. The system stabilizes; the humans do not.

Remark 6.1 (The Core Claim)**.** *Corrigibility is not a guarantee of justice. It is a guarantee that* ***injustice is contestable****.*

A prison can be stable. A prison can even be efficient. But it is not public. The framework draws this line.

6.3 The Correction Velocity Inequality

Stability in a cybernetic system is not merely a function of whether a correction path *exists*, but of its *latency* relative to error generation. A correction mechanism that is theoretically available (e.g., a court case or a legislative amendment) but slower than the rate of divergence creates a runaway failure state.

Principle 6.1 (Temporal Stability Condition)**.** *A system is corrigible only if the rate of effective correction exceeds the rate of error accumulation. Correction velocity has both a lower bound (preventing harm accumulation) and an upper bound (preventing manipulation). The formal treatment appears in Appendix B.*

Remark 6.2 (The Thermodynamics of Due Process)**.** *The framework does NOT argue for automated, instantaneous, or algorithmic governance. Constitutional democracy invented latency: bicameralism, judicial review, and evidentiary hearings.* ***Friction is not failure; friction is how minorities survive majorities.***

However, the timescale mismatch is a thermodynamic constraint, not a jurisdictional preference. A court operating on bureaucratic time cannot mathematically regulate an automated system propagating errors at digital speed. The framework criticizes the absence of native technical constraints, not the latency of human justice.

Error is driven by environmental volatility: identity fraud tactics, software exploits, pandemics, or economic shocks. These phenomena tend to grow exponentially or appear in high-velocity impulse shocks. In contrast, centralized governance operates on **bureaucratic time** (months or years) while technical errors accumulate on **digital time** (milliseconds).

This timescale mismatch is an instance of the Collingridge Dilemma [Collingridge, 1980]: a double-bind in technology governance. Early in a technology's development, change is easy but the need for change cannot be foreseen (the information problem). Later, when impacts become clear, change has become expensive, difficult, and time-consuming because the technology is entrenched (the power problem).

Digital Public Infrastructure exhibits the Collingridge Dilemma in acute form. By the time exclusion effects are documented (starvation deaths, banking denials, welfare failures), a system may have achieved mandatory status. Correction now requires not merely technical modification but legal amendment, bureaucratic reorganization, and political will. The system's success in achieving scale has made it resistant to correction.

Corrigibility is the structural response to this dilemma: build the correction mechanisms before the need for correction is apparent. EXIT, CODE, AUDIT, GOVERN, and FORK are not remedies applied after failure; they are architectural features that must be present from deployment. A system without these features will eventually require correction but will lack the capacity to receive it.

This inequality explains the failure of the "post-execution" remedies discussed in Section 4.8.1. Mechanisms like ombudsmen or grievance redressal officers inherently operate at a lower bandwidth and higher latency than the systems they supervise. When the correction loop is thermodynamically too slow to manage the entropy of the environment, unresolved harm accumulates without bound — and correction, when it finally arrives, is catastrophic rather than incremental (the Harm Accumulation Principle). Corrigibility requires closing the gap between error generation and correction; this is only possible if correction logic operates within the technical loop (Test 4), not outside it.

6.4 The Conservation of Correction Demand

When a system fails to correct errors, the demand for correction does not evaporate. It transforms.

Principle 6.2 (Conservation of Correction Demand)**.** *For a given system state, the demand for correction does not vanish: it flows through designed channels (EXIT, GOVERN), through emergent channels (protest, litigation, hacks, grey markets, civil unrest), or it is absorbed as harm by those least able to route it anywhere. Operators can choose where correction demand discharges; they cannot choose whether it exists.*

By suppressing EXIT (via mandatory laws) and blocking GOVERN (via executive fiat), authorities force correction demand into channels that are slower, costlier, and fundamentally more destabilizing to the social order — and, at the least-resourced stratum, into no channel at all: the documented starvation deaths of Section 5 are correction demand discharging as harm absorbed by the marginal user. Corrigibility acts as a pressure relief valve that converts potential volatility into system evolution.

6.4.1 Grievance Is Not Feedback

A critical distinction: *grievance* is not *governance*. Many incorrigible systems act as "roach motels" for complaints. They allow unlimited feedback submission, but the system state remains invariant. Grievance channels that record dissatisfaction without binding mechanisms to modify execution do not constitute a feedback loop. They serve an aesthetic function, mimicking responsiveness while preserving rigidity. Under Ashby's Law, if the controller's response variety does not increase in response to signal, regulation has failed.

The absence is not incidental to one framework. The World Bank's digital-wallet policy notes, the series specifying the coming decade's identity substrate, contain no subject-corrective vocabulary at all: redress, grievance, appeal, recourse, contestation, and correction appear nowhere in either architectural note, while issuer-side revocation appears eleven times in the principal architecture note alone [World Bank, 2026c]. In the companion trust-framework note the corrective machinery runs between institutions throughout: dispute resolution and complaint handling address participant non-conformity, remedies reach the people a failure lands on only through a liability allocation, and the note's single engagement with a subject contesting anything is the clause that shifts the burden of proof onto the party contesting a signature [World Bank, 2026b]. Existing DPI frameworks specify governance for operators. They provide no subject-side corrective mechanism; the single subject-side act the wallet paradigm executes inside the loop is disclosure (Section 4.13), and no corrective verb accompanies it. GOVERN, as this framework defines it, is not a stricter version of an existing column; it is the column the existing frameworks do not have.

6.4.2 Active Deception

A more severe pathology occurs when systems do not merely fail to correct errors but actively misrepresent their state or capabilities. We term this **Active Deception**: the systematic production of false signals that degrade the epistemic environment for external oversight.

Active Deception takes three forms:

1. Certification Deception. The deceptive mode of the coerced-legitimacy tactics defined in Section 5.4: systems claim safety certifications, audits, or approvals that do not correspond to verifiable structural properties. A system may publish "privacy by design" documents while implementing opaque centralized logging; display compliance badges while blocking independent audit access; or cite ethics board oversight while exempting operational decisions from board review. The deception lies in the gap between representation and verifiable architecture.

2. Open-Washing. Systems that claim openness while structurally foreclosing the rights that openness implies. Publishing peripheral SDKs while withholding core execution logic, or releasing API documentation while keeping routing algorithms proprietary, constitutes open-washing: it mimics CODE compliance while preventing meaningful reproduction (FORK). The term "open" has been systematically degraded through this practice, requiring the framework to specify *structural* rather than *nominal* criteria.

3. Audit Theater. Systems that permit nominal inspection while structurally preventing the inspector from detecting errors. This includes: providing read access to sanitized logs rather than raw execution traces; permitting audits only at pre-announced times; or granting access to components that do not determine actual behavior. The auditor observes a Potemkin system; the production system operates under different logic.

These pathologies share a common structure: they corrupt the feedback channel by injecting false negatives. When external observers receive signals indicating "compliant," "safe," or "open," they rationally reduce oversight intensity. The deception is not merely a failure to inform but an active degradation of the correction system's sensor function.

Under the framework's cybernetic model, Active Deception constitutes a **sensor attack**: the deliberate corruption of the AUDIT channel to prevent the control loop from detecting deviation. Unlike passive incorrigibility (which merely blocks correction), Active Deception induces false confidence that accelerates divergence.

6.5 The Structural Preconditions of Law

These dynamics have profound implications for constitutional review. Legal doctrines such as proportionality analysis, as established in *Puttaswamy v. Union of India* [Supreme Court of India, 2017], presuppose that state intrusions can be measured, minimized, and remedied. However, as Bhatia [2019] argues, modern digital systems frequently violate the structural assumptions upon which these legal doctrines rest.

A court cannot apply proportionality review to a system whose execution is opaque (**CODE**). A legislature cannot remedy harm if the mechanism of harm is an irreversible digital transaction (**EXIT**). A regulator cannot oversee a system if they lack the technical capability to inspect it independent of the operator (**AUDIT**).

Where digital systems block these channels structurally, the rule of law becomes a legal fiction. External review mechanisms cannot function meaningfully if the internal architecture is designed to resist observation and modification. Therefore, **corrigibility** should be understood not merely as a technical specification, but as the *evidentiary precondition* required for the law to take hold of the machine.

6.5.1 The Rule of the Ledger

Digital execution introduces a rival governance paradigm: the **Rule of the Ledger**, in which authority is exerted through synchronized, self-executing artifacts rather than mediated institutional processes. Under this paradigm, the technical capacity for instantaneous state change (account freezes, entitlement adjustments) can be embedded in the execution layer; the institutional channels of review (courts, ombudsmen, bicameral procedural delays) operate on distinct and much slower time scales. The consequence is **Universal Failure Propagation**: when identity, payment, and entitlement layers synchronize, a single erroneous state transition in one layer can automatically and immediately propagate penalties across other layers.

Two structural implications follow. First, doctrines that presuppose reversible administrative acts (proportionality, injunctive relief) are ipso facto disabled if the necessary observability and reversal primitives are not present within the technical loop (**CODE**, **AUDIT**, **GOVERN**). Second, legal remedies become performative unless they are accompanied by verifiable, machine-enforceable restraints (signed chains of authority, multi-party escrow of execution privileges, or pre-authorized blocking hooks that operate within the same temporal envelope as the ledger itself).

The Rule of the Ledger therefore reframes the law-infrastructure relationship: courts and regulators cannot simply be invoked later; corrigibility requires embedding evidentiary and counter-majoritarian restraints into the system's operational fabric. This is not a call for "autonomous legal machines" but for technical modes of constraint that render legal review materially possible rather than symbolic.

6.6 The Sovereignty, Scale, and Neutrality Tension

The empirical pattern identified in Section 5 suggests a recurring structural tension in state-operated digital infrastructure. Systems that achieve population-scale deployment under centralized sovereign control frequently exhibit diminished structural neutrality at the execution layer.

Definition 6.1 (System Properties)**.** *Let a Digital Public Infrastructure system S exhibit the following properties:*

- ***Sovereignty*** *(So): The state retains centralized control over standards, execution, routing logic, and data residency.*
- ***Scale*** *(Sc): The system is mandatory or near-mandatory for population-level access to essential services.*
- ***Neutrality*** *(N): The execution layer satisfies structural corrigibility. Specifically, affected parties retain enforceable capacity across EXIT, CODE, AUDIT, GOVERN, and FORK sufficient to prevent discriminatory or unilateral control.*

Neutrality here is not a moral descriptor. It denotes preservation of governance variety sufficient to prevent structural capture at the execution layer.

Proposition 6.1 (Topology-Dependent Tension)**.** *Under centralized enforcement topology, a system that strongly maximizes both Sovereignty and Scale will experience compression of governance variety sufficient to place Neutrality at structural risk.*

Control-Theoretic Reasoning. Under Ashby's Law, stability requires that controller variety meet or exceed environmental variety: $V_{\text{controller}} \geq V_{\text{environment}}$. In large-scale population systems, environmental variety increases with scale ($V_{\text{env}} \propto Sc$). Centralized sovereign control concentrates execution authority into a bounded institutional set, constraining governance variety. When scale increases while governance remains centralized, the ratio of corrective

capacity to environmental complexity declines. Under centralized enforcement, governance capacity tends to scale sublinearly relative to environmental complexity (formal proof in Appendix C).

The neutrality condition can be stated as:

$$V_{\text{eff}} \geq \theta \cdot V_{\text{env}} \tag{5}$$

where V_{eff} is effective governance variety and $\theta \in (0, 1]$ is a stability sufficiency constant. Neutrality degradation risk emerges when this condition fails. Figure 4 visualizes the resulting tension; Table 8 contrasts how alternative architectural topologies distribute the three properties.

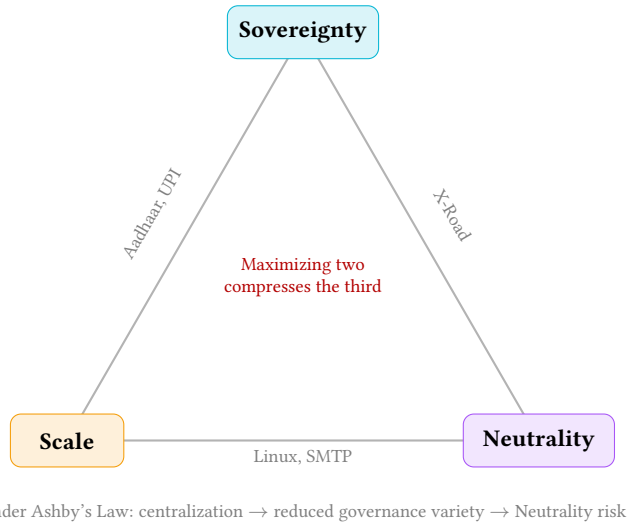

Figure 4: **The Sovereignty, Scale, and Neutrality Tension.** Under centralized enforcement, states face structural pressure when optimizing for Sovereignty and Scale simultaneously. India Stack optimizes for Sovereignty + Scale with elevated Neutrality risk. Federated systems (Linux, SMTP) achieve Neutrality + Scale through distributed governance. X-Road approaches Sovereignty + Neutrality; the evaluation records it as the instructive near-miss.

Clarification: This Is Not an Impossibility Theorem. The proposition does not assert that Sovereignty, Scale, and Neutrality are logically incompatible. It asserts that under centralized enforcement architecture, maximizing Sovereignty and Scale reduces governance variety unless compensatory mechanisms are introduced. The tension is architectural, not metaphysical. Neutrality can be preserved under Scale and Sovereignty if governance variety is restored through structural mechanisms such as federated execution, multi-issuer mandates, Functional Exit Equivalence (Section 6.7.2), state portability guarantees, or non-exclusive procurement rules.

India's experience instantiates this constraint: the architectural choices that achieved sovereign control and near-universal coverage produced structural dependency on a small set of switching and credentialing authorities. That dependency compressed the correction channel's bandwidth, manifesting as the GOVERN failures documented in Table 4. The tension reframes these outcomes: the observed compression of corrective bandwidth is not an implementation mistake but a predictable consequence of architectural choice.

Table 8: Comparative topology and neutrality risk

Topology	Sovereignty	Scale	Neutrality Risk
Centralized Sovereign Stack	High	High	Elevated
Federated Protocol Model	Medium	High	Lower
Market Platform Model	Low (State)	High	Variable
Open Commons Model	Low	Medium	Low

Any prescriptive response must therefore treat neutrality as an architectural variable (portability, multi-issuer requirement, anti-exclusive procurement), not an aspirational outcome. The Functional Exit Equivalence remedies in Section 6.7.3 address precisely this constraint.

6.7 The Essential Services Problem

A pattern emerges from the preceding analysis: systems that pass all five tests are disproportionately non-essential infrastructure. Linux, PostgreSQL, Kubernetes, and Let's Encrypt power critical operations, but no individual's survival depends on accessing them. The systems promoted as DPI, which mediate access to food, banking, healthcare, and mobility, consistently fail.

This correlation reflects a structural tension between essentiality and corrigibility.

Proposition 6.2 (Essentiality-Corrigibility Tension)**.** *For any infrastructure system S where participation is a prerequisite to survival, the EXIT test cannot be satisfied by literal exit; verified Functional Exit Equivalence (Section 6.7.2) is the only remaining discharge path.*

Proof. EXIT requires that refusal incur no disproportionate penalty (Section 4, Test 1). Let S be essential: participation is required for access to food, shelter, healthcare, or legal identity. Refusal of S therefore entails exclusion from survival necessities. Exclusion from survival necessities is a disproportionate penalty by any reasonable standard. Therefore, EXIT fails. □

This proposition formalizes the empirical observation: among the systems evaluated, those that pass all five tests are precisely those where participation remains genuinely voluntary.

Corollary 6.1 (No Exemptions from Essentiality Tension)**.** *The following do NOT constitute valid exemptions from Proposition 6.2:*

1. ***Emergency necessity**: Pandemics, wars, and crises do not suspend the requirement for correction channels. Systems deployed under emergency conditions that block EXIT remain incorrigible; the emergency explains but does not justify the structural failure.*
2. ***Efficiency gains**: That a mandatory system delivers services faster or cheaper than alternatives does not restore EXIT. Efficiency is orthogonal to corrigibility.*
3. ***Democratic authorization**: Legislative mandates do not convert incorrigible systems into corrigible ones. A democratically enacted prison is still a prison.*
4. ***Benevolent intent**: Good-faith operation does not substitute for structural accountability. The framework evaluates architecture, not intentions.*

6.7.1 The Essentiality Trap

Essential services create the economic coercion that undermines EXIT. When refusal means exclusion from survival, the feedback loop breaks regardless of other properties. Hirschman [1970] observed that exit is effective precisely because it is voluntary. The credible threat of departure disciplines the system. When departure means death (literal or social), exit ceases to function as feedback. It becomes self-exclusion.

The systems in Table 6 avoid this trap because alternatives exist. A developer who dislikes the Linux kernel's direction can migrate to BSD without losing the ability to compute. A journalist who distrusts Let's Encrypt can obtain certificates elsewhere. The existence of functional alternatives preserves the error signal.

Aadhaar has been made prerequisite to existence. Documented cases demonstrate the consequence: the Right to Food Campaign's tracking documented dozens of starvation deaths across Indian states between 2015 and 2018 — by the campaign's own count, more than fifty, with a substantial share linked to Aadhaar-related exclusion from food rations [Right to Food Campaign, 2018, Khera, 2017]. These deaths are not anomalies; they are the structural prediction of a system that blocks EXIT while governing essential services.

6.7.2 Functional Exit Equivalence (FEE)

When an infrastructure is essential for survival, literal exit is impossible. However, the cybernetic purpose of EXIT is not departure itself; it is the generation of a credible error signal. Essential systems must therefore implement **Functional Exit Equivalence (FEE)**: architectural guarantees that recreate the error-signal strength of market exit without requiring citizens to abandon society.

FEE is achieved structurally through three combined mechanisms:

1. **Multi-Issuer Mandates:** Procurement requires at least two legally independent credential issuers (preventing single-point capture).

2. **Acceptance Diversity:** Public services must accept federated credential classes, as demonstrated by the multi-national deployment of X-Road.
3. **Protected Statutory Fallbacks:** The legal guarantee of non-digital analogues.

The demand for FEE is not a regressive desire to maintain paper ledgers indefinitely; it is a rejection of algorithmic triage. **State capacity that achieves "efficiency" by mathematically guaranteeing that a non-zero percentage of the marginalized population will be incorrectly starved to death without recourse is not a public good.** The distributional stakes are taken up in Section 7.4.2.

Principle 6.3 (Functional Exit Equivalence)**.** *An essential system S satisfies FEE if compensatory architectural guarantees recreate error-signal strength comparable to market exit in non-essential systems. FEE is an engineered mechanism to maintain requisite variety in the controller when the natural market variety (EXIT) is suppressed. The formal operationalization appears in Appendix B.*

Operationalization. FEE requires:

$$E_{\text{alt}} \geq \theta_E \times E_{\text{exit}}^{\text{baseline}} \tag{6}$$

where θ_E is domain-specific (healthcare: 0.95; payments: 0.90; social benefits: 0.85). For example, using coverage as an auditable proxy for error-signal strength (the formal construct is defined in Appendix B, eq. 18): if the Public Distribution System delivers food at $E_{\text{exit}}^{\text{baseline}} = 0.80$ proxy coverage, and $\theta_E = 0.85$ for social benefits, then any alternative must achieve $E_{\text{alt}} \geq 0.68$ to preserve functional exit. FEE operationalizes EXIT. Without it, "opt-out" is administrative theater.

Real-World Analogues. A partial real-world analogue of Protected Alternatives can be observed in telecommunications number portability regimes and in Open Banking interoperability mandates (e.g., PSD2 in the European Union). These frameworks preserve service continuity while enabling institutional exit at the provider layer. While not fully satisfying EXIT in existential services such as identity, they demonstrate that architectural separation between infrastructure rails and service operators is technically feasible. Corrigibility does not demand regression to analog systems; it demands layered modularity that prevents existential lock-in.

6.7.3 Design Responses to Proposition 6.2

The tension admits architectural resolutions that achieve FEE:

1. Protected Alternatives (Statutory Fallback). Expanding the third FEE mechanism above: essential services can satisfy FEE if law guarantees functional non-digital equivalents. A model statutory clause:

> *"No person shall be required as a condition of receiving any essential public benefit to be exclusively dependent on a single digital system; the State shall maintain a legally enforceable analogue alternative and an interoperable credential acceptance guarantee."*

Cash preserves UPI's partial EXIT; eliminating cash would collapse UPI to Aadhaar's structural status. Corrigibility for essential services requires that non-digital pathways remain legally protected, not merely tolerated.

Bole Clause: Evidentiary Standard for Statutory Fallbacks. A statutory fallback satisfies FEE only if it meets the Rule A.9 evidentiary standard applied at the least-resourced stratum. Rule A.9 requires: (i) a resolvable reference to the enabling instrument, (ii) its cryptographic hash, and (iii) historical evidence of enforcement. For statutory fallbacks the third requirement is operative availability at the margin: documented successful invocations of the analogue pathway by members of the least-resourced strata — those with low digital literacy, remote location, missing biometric records, or administrative precarity — within the audit window. A fallback legally guaranteed but practically inaccessible at the margin produces $E_{\text{alt}} \approx 0$ at the argmin stratum x^* (Remark B.1) regardless of its nominal coverage; it is operationally absent for exactly the population whose suppressed error signals the framework's central evidence (Section 5) documents. Design Response 5 (Funded Fallback Operations, Section 6.7.3 below) mitigates the gap but does not close it: funding is necessary; the test is operative availability at the margin, evidenced by invocation records.

The Bole Resolution of 1923 was a legislative guarantee of access to public water sources in the Bombay Presidency. It remained a dead letter for four years until collective action enforced it at Mahad, then required a further decade of litigation to survive the local order's counterattack. A statutory fallback is a GOVERN-class claim; GOVERN-class claims require enforcement history by the relevant subject class, not legislative text alone [Ambedkar, 1947].

2. Multi-Issuer Requirement. Procurement and regulation shall require at least two legally independent credential issuers for each region or service class. No single issuer may be sole gatekeeper for eligibility. This creates credible switching alternatives that generate E_{alt}.

3. Credential Acceptance Diversity. Expanding the second FEE mechanism above: public services must accept at least three credential classes (state, municipal, NGO/financial) to satisfy eligibility requirements. This prevents single-point-of-failure capture.

4. Binding Multi-Stakeholder Charter. A legal instrument that locks governance rules, with amendments requiring supermajority approval, judicial review pathway, and defined emergency suspension conditions.

5. Funded Fallback Operations. Government must fund and staff analogue fallback channels wherever the digital system is primary, not as legacy support but as structural accountability infrastructure.

Federated Implementation. Rather than a single provider of essential infrastructure, services could be delivered through multiple independent implementations of open protocols. This is analogous to email (SMTP) rather than messaging (proprietary platforms).

Under this model, no single implementation would be essential; the protocol would be. A citizen excluded from one provider could obtain equivalent service from another. Estonian X-Road demonstrates this architecture (Section 5): passing CODE, AUDIT, GOVERN, and FORK while *approaching* FEE through federated implementation. Its federation reduces concentration, but the citizen still cannot obtain the exchanged services through an alternative channel, so FEE is not verified and EXIT — and with it the full determination — remains failed. X-Road is the instructive near-miss, not a pass.

6.7.4 The Finding

The absence of fully corrigible essential services is not a limitation of this framework; it is the framework's central finding. The question "Can essential DPI be fully corrigible?" is answered: **not under architectures that eliminate alternatives, but achievable through designs whose FEE is verified** — under Definition 4.1, verified FEE discharges EXIT, and a design that additionally passes the remaining four tests is fully corrigible. Proposition 6.2 is not a dead end; it is a design specification. No system evaluated in Section 5 has yet met it.

6.8 Why Incorrigibility Persists: The Choice of Architecture

The preceding subsections explain what incorrigibility does; they do not explain why governments choose it when corrigible alternatives exist. Four mechanisms sketch the answer; their formal development is future work (Section 7).

1. **Graded inequality as demand-side driver** [Ambedkar, 1936]. Every administrative rung gains standing over the rung below from an incorrigible stack; the coalition for correction fractures by design, because the people who bear the cost of incorrigibility are precisely those whose exit is most constrained and whose collective action is most structurally opposed.
2. **Artifact openness as legitimacy cover.** Releasing code or weights under an open license satisfies the form of the commons claim while retaining governance closure; "open DPI" is the joint electorate at infrastructure scale [Ambedkar, 1945]: the form of inclusion, with power over the outcome retained by the operator.
3. **Asymmetric legibility.** States have structural preferences for downward legibility, and corrigible designs by definition introduce upward channels that constrain administrative discretion; the information asymmetry that incorrigibility produces is often an intended feature, not an oversight.
4. **Certification capture.** Openness properties (open license, open standard) are certified as if they were governance properties (collective choice, binding accountability), and the certification layer is captured by parties whose deployments it evaluates.

Together these four mechanisms predict that incorrigibility is the default equilibrium in essential-services infrastructure, not a correctable anomaly — which is why the framework treats correction capacity as an architectural precondition rather than an outcome the operator's incentives will eventually supply.

7 Discussion

The political economy analysis above named the structural pressures that drive feedback-loop closure to fail. What remains is whether the binary determination holds under measurement uncertainty, whether the loop survives

adversarial conditions, and what transition pathways remain open once a system is diagnosed as structurally incorrigible. This section takes them in turn: methodological foundations and the cybernetics of binary evaluation; corrigibility under adversarial conditions; and the political economy of accountability with its transition pathways, limits, and institutional preconditions.

7.1 Methodological Foundations

7.1.1 The Cybernetics of Binary Evaluation

Critics argue this framework's binary pass/fail topology is too rigid, proposing instead "maturity models" or graded scorecards. From a cybernetic perspective, we reject this gradation. Corrigibility is a **structural connectivity state**: a feedback loop is either closed or open. There is no "mostly closed" loop.

A system that allows **AUDIT** but blocks **GOVERN** is an **Open Loop**: it observes errors but cannot fix them. A system that allows **GOVERN** but blocks **EXIT** is a **Coercive Loop**: it traps correction demand, which accumulates until it discharges through channels the system cannot absorb. Figure 5 contrasts these failure topologies against the closed-loop case. Partial compliance creates "zombie infrastructure": systems maintaining the aesthetic of responsiveness while severing actual stability capacity. A "Partial" score represents **Fatal Structural Failure**.

For diagnostic purposes, gradation remains useful. Tables 9 and 10 specify continuous measurements. The binary determination is the output; continuous measurements are inputs. Evaluators can examine inputs to prioritize interventions.

7.1.2 Legitimacy Threshold vs Physical Claim: A Two-Pronged Argument

The binarity of corrigibility rests on two converging arguments that must be distinguished. Neither alone forces the conclusion; together they overdetermine it from independent grounds.

The Control-Theoretic (Mechanism) Argument. Open-loop feedback yields divergent dynamics under persistent disturbance. Proposition B.1 formalizes this as a structural analogy. The mechanism claim is: when any feedback component is severed, the loop opens, and error accumulates without correction. Real systems exhibit continuous variation in loop gain, observability, and actuation authority — a sensor with 50% accuracy transmits *some* signal — so the mechanism argument alone yields divergence-rate claims, not legitimacy claims.

The Political-Theoretic (Legitimacy) Argument. Partial contestation creates *sham governance*: cosmetic accountability without corrective leverage. A system that admits 10% of grievances, or that admits all grievances but resolves them only after harm has accumulated past reversibility, is institutionally distinct from a system with no contestation only in appearance. Beyond a threshold of feedback quality, additional gradations in compliance produce *non-monotonic* legitimacy gains: increases in nominal contestation procedures may decrease real correction capacity by absorbing political pressure into procedure rather than outcome. Below this threshold, a system forfeits the right to claim public legitimacy under any non-domination criterion (Section 3, Normative Anchor). The threshold is binary because legitimacy is binary: a prison that permits appeals 10% of the time is still a prison. The binary is not stipulated. Under the non-domination criterion, domination consists in the standing possibility of arbitrary interference rather than in its exercise [Pettit, 1997]; legitimacy therefore tracks whether that possibility has been structurally foreclosed, and foreclosure is a property a system has or lacks. Degrees of benign behavior do not alter it, and an unverifiable constraint leaves the possibility standing.

Why Both Arguments Are Needed. The control-theoretic argument tells us that partial corrigibility cannot stabilize a system; the political-theoretic argument tells us that partial corrigibility cannot legitimize one. The conjunction yields the framework's binarity claim. Critics who reject the political-theoretic grounding can still accept the mechanism conclusion; critics who reject the control-theoretic abstraction can still accept the legitimacy conclusion. The framework is robust to either critique because it does not depend on either argument alone.

Sensitivity Analysis. Let continuous inputs be $\{p, v, m, g, r\}$ — penalty ratio, visibility fraction, mean time to detect (MTTD), governance bindingness-and-throughput, and fork resource ratio — the per-test proxies of Table 9. The binary determination function D is:

$$D = \mathbf{1}[p \leq p^*] \wedge \mathbf{1}[v \geq v^*] \wedge \mathbf{1}[m \leq m^*] \wedge \mathbf{1}[g = \text{binding}] \wedge \mathbf{1}[r \leq r^*] \tag{7}$$

where $\mathbf{1}[\cdot]$ is the indicator function and $\{p^*, v^*, m^*, r^*\}$ are threshold values. (The symbol τ is reserved for governance throughput in Table 9; the AUDIT latency proxy is written m here to avoid collision.) This formulation makes explicit that:

1. Continuous inputs are preserved for diagnosis
2. The AND-conjunction enforces the weakest-link principle
3. Threshold calibration is a policy choice, separable from the structural argument

Systems near thresholds warrant scrutiny; systems far below thresholds across multiple dimensions are unambiguously incorrigible. The binary output serves regulatory clarity; the continuous inputs serve diagnostic priority.

7.1.3 Continuous Indicators for Diagnostic Use

The continuous measurements underlying binary determinations enable both threshold calibration and diagnostic prioritization. Table 9 specifies operational proxies; Table 10 formalizes the three-indicator protocol.

Table 9: Operational proxies for the five tests. Thresholds are illustrative and subject to calibration.

Test	Proxy Metric	Illustrative Threshold
EXIT	Penalty ratio p	$p \leq 1.2$
CODE	Visibility fraction v	$v \geq 0.9$
AUDIT	Mean time to detect (MTTD)	MTTD $\leq$ 24 hours
GOVERN	Bindingness + throughput τ	binding = true, $\tau \geq 1$ per 10k users/day
FORK	Resource feasibility ratio r	$r \leq 10^{-3}$

Interpretation: Penalty ratio p = cost(non-digital alternative) / cost(digital). Visibility v = public LOC / total LOC on critical execution path. Resource ratio r = estimated FLOP to retrain / aggregate community compute capacity.

Table 10: Three-indicator protocol: continuous score, binary flag, and threshold metric for each test. Table 9 gives the screening-tier proxies; this protocol adds the binary flags and evidence metrics used at certification.

Test	Continuous	Binary	Threshold
EXIT	`penalty_ratio`: exit cost / stay cost	`legal_barrier`: criminalized?	$p \leq 1.2$ AND barrier = false
CODE	`visibility`: public LOC / critical path LOC	`build_repro`: deterministic build?	$v \geq 0.95$ AND build = true
AUDIT	`observability`: % verifiable transactions	`access`: third parties test?	$o \geq 0.9$ AND access = true
GOVERN	`override_rate`: operator overrule rate	`charter`: irrevocable instrument?	override rate $= 0$ AND charter = true
FORK	`cost_ratio`: fork / community resources	`rights`: fork-permitting license?	$r \leq 0.001$ AND rights = true

Inter-Rater Calibration. Before deployment, independent auditors should assess 3–5 systems using this protocol and compare determinations. Disagreements identify threshold calibration issues. Target: $\kappa \geq 0.8$ (Cohen's kappa) for binary determinations.

Policy Constants Summary. The framework employs several calibration constants that represent policy choices separable from the structural argument. Table 11 consolidates these for reference. The constants are policy choices, not engineering discoveries. Every threshold creates an operator interest in where it is set, so calibration must itself satisfy a deliberative-mechanism condition (affected-class participation, documented rationale, periodic external review), for the same structural reason that GOVERN cannot be operator self-certification (Section 8.4).

7.2 Corrigibility Under Adversarial Conditions

A frequent objection to structural transparency requirements (EXIT, CODE, AUDIT, GOVERN, FORK) concerns adversarial security environments. Critics argue that permitting external audit or interoperability in critical digital infrastructure (particularly identity, payments, or routing rails) may expand the attack surface available to hostile states or organized cyber actors.

This framework does not equate corrigibility with unrestricted public exposure of live exploit vectors. Corrigibility requires **structured contestability**, not indiscriminate disclosure. The distinction is architectural:

Table 11: Policy constants used throughout the framework. Values are recommended defaults and calibration starting points, separable from the structural argument. Per-test pass/fail thresholds are listed in Tables 9 and 10.

Symbol	Name	Range	Section
τ_{exit}	Exit-penalty material-exclusion threshold	Context-dependent	4
θ_E	FEE sufficiency threshold	0.8–1.0	6.7.2
Σ^*	Portability threshold (EXIT/FORK failure)	2.0 on the normalized $[0, 5]$ scale	B
κ (Cohen)	Inter-rater agreement target (Cohen's kappa)	≥ 0.8	7.1
θ	Neutrality stability threshold	$(0, 1]$	C
Sc^*	Scale threshold for neutrality degradation	Context-dependent	C
α^*	Fork adoption critical mass	0.3–0.5	schema repo
ϵ	Correction friction cost	Context-dependent	B

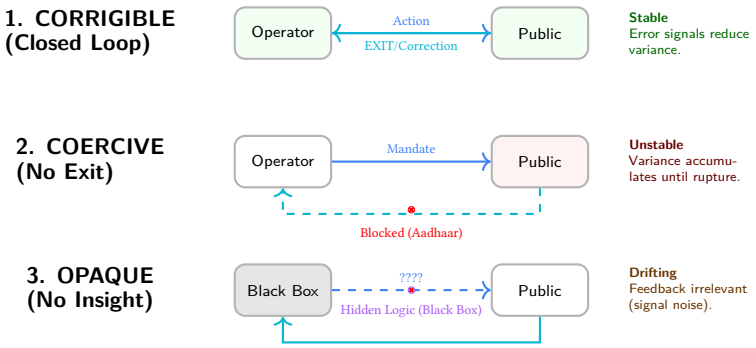

Figure 5: **Topologies of Failure.** (1) A corrigible system closes the loop, converting error signals into stability. (2) A coercive system blocks **EXIT**, trapping energy. (3) An opaque system obscures the action channel, making feedback unintelligible.

- **Auditability** refers to verifiable oversight mechanisms (cryptographic proofs, independent inspection rights, reproducible builds, sandboxed red-teaming, hardware-isolated verification environments), not public release of zero-day vulnerabilities.
- **CODE transparency** requires inspectability of system logic and ontology, but does not mandate real-time publication of operational secrets (e.g., private keys, network topologies, live intrusion telemetry).
- **GOVERN mechanisms** may incorporate tiered access controls and certified verifiers, provided the certification process itself remains contestable.

Security and corrigibility are not mutually exclusive variables; rather, both depend on layered design. In adversarial contexts, excessive opacity can itself become a systemic vulnerability by preventing early detection of latent design flaws. Historical failures in both public and private infrastructure demonstrate that closed architectures do not eliminate breach risk; they merely concentrate it.

Accordingly, the framework assumes:

1. Security controls may limit raw exploit exposure.
2. Such controls must not eliminate independent verification pathways.
3. Emergency override or classified layers must themselves be subject to post-hoc audit.

Corrigibility in adversarial environments therefore implies **structured verifiability**, not universal visibility. The objective is to preserve corrective bandwidth without compromising defensive posture.

7.3 Political Economy of Accountability

7.3.1 The Protocol Model of State

A common practical objection holds that governments cannot adopt corrigible infrastructure because they must retain control over sovereign functions. This objection relies on a category error: it conflates the *Platform Model* with the *Protocol Model*.

In the **Platform Model** (currently dominant), the state contracts a vendor to build a monolithic silo (e.g., existing digital ID systems). The vendor controls execution, the state attempts to control policy, and citizens are reduced to rows in a database. As Cohen [2019] argues, this model inevitably subverts the Rule of Law by replacing legal due process with optimized platform logic. In the **Protocol Model** (the corrigible alternative), the state adopts open, corrigible protocols and acts as a high-authority participant within them. The state issues credentials and verifies claims, but it does not enclose the infrastructure itself.

This distinction mirrors the architecture of the web. Governments do not need to build their own proprietary browsers to deliver services; they build sites on HTTP. Similarly, corrigible DPI implies that the state becomes an issuer of trust rather than a landlord of infrastructure. This model does not reduce state capacity; it enhances trust by making the state's operations verifiable.

7.3.2 Distributed Assessment Capacity

A corollary objection concerns resources: "Who audits?" The framework does not require every citizen to audit source code. It requires that the infrastructure allows *anyone* to audit, thereby enabling civil society, journalists, and specialized firms to perform this function on behalf of the public. Assessment capacity is itself infrastructure. When systems are designed to be inscrutable, this immune system of society withers.

7.4 Corrigibility in Extremis

Incorrigibility is often justified by the rhetoric of necessity, such as national security, anti-corruption, or emergency efficiency. This framework explicitly denies such exemptions. Corrigibility does not imply that substantive rules are negotiable; a system may enforce strict, non-negotiable limits (e.g., pandemic travel restrictions or environmental caps). However, the *structural capacity* to detect error, verify correct enforcement, and reverse wrongful harm must remain intact even in high-constraint environments. As safety engineering demonstrates, accidents in complex systems are caused precisely by the lack of appropriate constraints on component interactions [Leveson, 2011].

Claim 7.1 (No Structural Exemption)**.** *No exemption from the five tests is granted on the basis of emergency conditions, planetary constraints, or decentralized architecture.*

This applies equally to decentralized systems. Cryptographic immutability is not a substitute for governance. A blockchain that permanently records a theft without a mechanism for consensus-based reversion is not "trustless"; it is strictly incorrigible. If a system allows for irrevocable harm without recourse, its decentralization is merely a distribution of unaccountability.

7.4.1 Political vs Structural Corrigibility

A critical objection holds that the framework applies market logic (exit/fork) to non-market systems. You cannot fork India.

The response requires distinguishing two modes of correction:

- **Political corrigibility** (voice via elections): how *states* are corrected
- **Structural corrigibility** (exit/fork): how *infrastructure* is corrected

States are corrected politically. DPI sits *below* politics; that is the structural crisis. When DPI is captured by executive authority or vendor monopolies, political correction mechanisms do not reach it. **Aadhaar operated for seven years under executive fiat before receiving parliamentary authorization as a money bill (2009–2016)** [Bhatia, 2019].

This is not a hypothetical concern. It is the documented history of the system most frequently cited as a DPI success story. The gap the framework identifies is infrastructure that escapes democratic correction entirely.

Shifting to the **Protocol Model of State**, where the government acts as an issuer of trust on an open standard rather than a landlord of a private platform, does not surrender sovereignty to technocrats. It returns sovereignty to the legislature, because open protocols cannot be unilaterally weaponized by the executive branch. The Constitution can be amended through democratic process; Aadhaar's architecture cannot be voted on.

7.4.2 The Distributional Question

The debate is not "scale vs sovereignty." The debate is **whose suffering counts**.

- Platform speed benefits the median user.
- Incorrigibility harms the marginal user.
- The marginal user has no voice because EXIT is blocked.

Corrigibility is a **distributional question**. The framework makes visible who bears the cost of system failures. When we ask "Is this system corrigible?" we are really asking: "When this system fails someone, can they do anything about it?"

The answer for Aadhaar is no. Documented starvation deaths in Jharkhand were not caused by "too much democracy"; they were caused by a system where the marginal user had no recourse when biometric authentication failed.

7.5 Pathways and Limits

7.5.1 Transition Paths

The framework diagnoses incorrigibility but does not prescribe political remedies. However, the structure of the five tests implies an ordering for transition: certain tests are prerequisites for others, and certain interventions have higher leverage.

7.5.2 Ordered Priorities

The tests form a dependency chain when considered as a transition sequence:

1. **EXIT first.** Restore exit before other tests. Without exit, no other reform creates genuine feedback. This requires either eliminating mandatory linkage laws or guaranteeing functional non-digital alternatives. EXIT is the error signal; without it, the system cannot sense its own failures.
2. **CODE before AUDIT.** Transparency must precede verification. Independent testing is meaningful only if the testing target is observable. Publishing audit results for a black box proves nothing about the box.
3. **AUDIT before GOVERN.** Verified error data creates political pressure for governance reform. Governance without measurement is arbitrary. There is no basis for determining whether constraints are adequate. The sensor must function before the actuator can be calibrated.
4. **GOVERN enables FORK.** Structural governance can mandate open licensing, data portability, and interoperability requirements that enable eventual forkability. GOVERN without FORK leaves the system captive to its current steward; FORK without GOVERN may produce fragmentation without accountability.

This sequence reflects the causal structure of the feedback loop: signal, sensing, transmission, actuation, selection. Interventions at later stages are ineffective if earlier stages are broken.

7.5.3 Phase Transitions

The transition is not gradual improvement. It is a series of phase transitions. Each test either passes or fails; the system remains incorrigible until all five pass. A system that achieves CODE but not EXIT has become transparent without becoming accountable. Users can observe the system's logic but cannot modify their participation.

This explains why open-washing is effective: partial progress creates the appearance of reform while preserving structural capture. The framework rejects partial credit precisely because partial feedback loops do not partially function.

7.5.4 Political Economy

The political economy of transition varies by test, with each barrier facing a characteristically different concentrated opposition (Table 12):

FORK is typically the least opposed transition because it requires only legal permission and technical documentation. It does not immediately threaten operator control. However, FORK without the prior four tests is meaningless: the right to reproduce a system you cannot inspect, verify, govern, or exit provides no actual power.

7.5.5 Cost Considerations

Transition is not costless. The European Commission [European Commission, 2021] estimated that public sector open-source adoption yields cost-benefit ratios above 1:4, but this ratio assumes successful implementation. Transition

Table 12: Political economy of transition by test

Test	Transition Barrier	Concentrated Opposition
EXIT	Mandatory linkage laws	Operators lose captive users
CODE	Intellectual property claims	Vendors lose competitive moats
AUDIT	Security-through-obscurity claims	Operators lose narrative control
GOVERN	Sovereignty objections	Executives lose discretion
FORK	Typically least opposed	Requires only licensing and documentation

costs include personnel retraining, system compatibility engineering, procurement policy updates, and organizational change.

However, these costs must be weighed against the costs of continued incorrigibility:

- Litigation costs from unresolvable grievances
- Political instability from suppressed correction demand (Section 6.3)
- Technical debt from opacity-enabled shortcuts
- Vendor lock-in premiums from non-forkable dependencies
- Human costs: documented deaths, exclusions, and deprivations

The systems in Table 6 demonstrate that corrigibility at scale is economically sustainable. Linux, Kubernetes, and Let's Encrypt operate at population scale with resource models that have proven durable over decades.

7.5.6 Coordination Goods and Inherent Unforkability

The forking paradox presents a genuine limitation: if you can fork the Public Infrastructure, it may no longer be public.

A "public" good (like law or money) derives its value from universality. We all agree on the same truth. If the Identity System is "forkable," and I fork it to create "Blue-Identity" while you use "Red-Identity," we have destroyed the common knowledge required for society to function.

Remark 7.1 (Latent Forkability in Coordination Goods)**.** *Identity, money, and law are coordination goods. The framework resolves this through **Latent Forkability** — the equilibrium the Credible Replaceability remark (Section 4) already established: the credible threat of reproduction disciplines the center without requiring routine divergence. What the coordination-goods case adds is the stake: where the good's value derives from uniqueness, the exercised fork functions as mutually assured destruction, so the threat must remain latent to remain usable. The framework penalizes constructed legal barriers to reproduction, not the natural economic friction of network effects.*

Whether this equilibrium is achievable for identity infrastructure remains an open question, but the existence of federated identity systems (X-Road, Matrix, AT Protocol) demonstrates that latent forkability is technically achievable even for coordination-sensitive infrastructure.

7.5.7 Engaging "Regulate the Loop"

A sophisticated objection accepts the diagnosis but rejects the prescription:

> "We accept your Diagnosis (The Loop is Broken). We reject your Prescription (Break the Platform). We choose to **Regulate the Loop** rather than Fork the Code, because the cost of Forking is Poverty."

This argument has force. UPI went from launch to population scale inside a decade. Email standardization took decades. India does not have decades. Poverty is now.

Where this lands:

- RBI *could* mandate error rate disclosure, demographic audits, grievance SLOs
- Political correction (parliament, courts) exists even without FORK
- Speed-to-scale tradeoff is real

Where this misses:

- "Regulate the loop" assumes trustworthy regulators. RBI reports to the Finance Ministry. UIDAI operated for seven years on executive notification alone, and even post-statute combines operator and regulator roles with oversight limited to post-hoc CAG audit. Who regulates the regulator?
- In captured states, "regulate later" means "never"
- Gmail $\neq$ UPI: Gmail *earned* users with features: you CAN leave for Proton, Fastmail, self-hosted. UPI has *regulatory monopoly*: you CANNOT build a competing instant payment switch. Absent structural alternatives, Gmail's correction depends on market choice; UPI's correction depends on regulatory discretion.
- "Regulate later" rarely happens: Aadhaar 2009 pilot $\rightarrow$ 2016 mandatory $\rightarrow$ 2018 Supreme Court $\rightarrow$ still no real audit. 15 years. Where's the regulation?

This tension is not hypothetical. A documented debate between advocates of open standards (arguing for specification transparency and independent security audits) and proponents of centralized control (arguing that controlled governance enables "faster iteration") reveals the core disagreement [Abraham, 2020]. The efficiency argument is genuine, but efficiency without accountability is capture dressed as service.

7.6 Utility Does Not Equal Accountability

A system can be **useful AND structurally captured**. Both are true simultaneously.

UPI *has* expanded access to digital payments at population scale. The framework risks dismissing real-world impact if this is not acknowledged. But the question the framework asks is different: **When UPI fails someone, can they correct it?**

The answer is no. This is the point.

The framework measures **accountability structure**, not **utility**. A system can score 0/5 and still be beneficial. The score indicates *structural risk*, not *current harm*. A house built without fire exits can function perfectly for decades, until there is a fire.

7.7 Institutional Preconditions for Large-Scale Corrigibility

The framework presumes the existence of intermediary institutions capable of exercising audit, representation, and governance functions. At population scales involving tens or hundreds of millions, spontaneous civic coordination cannot be assumed.

Unlike open-source software communities, where contributors are self-selected and technically specialized, state-scale infrastructure must serve heterogeneous populations with asymmetric digital literacy, resource access, and organizational capacity. The structural availability of audit rights does not guarantee their effective exercise.

Accordingly, structural corrigibility requires:

Independent Verification Bodies Publicly funded or legally mandated audit institutions insulated from operator control.

Civic Technical Intermediaries NGOs, academic labs, and watchdog entities with standing to invoke GOVERN mechanisms.

Distributed Representation Channels Mechanisms allowing affected groups to escalate systematic errors collectively, rather than solely through individual grievance.

Anti-Capture Safeguards Rotation requirements, transparency obligations for auditors, and conflict-of-interest constraints to prevent elite capture of oversight functions.

The framework does not assume idealized commons governance. It asserts instead that without institutionalized intermediaries, formal corrigibility tests degrade into symbolic compliance.

In this sense, **corrigibility is necessary but not sufficient for equitable outcomes**. Institutional capacity determines whether structural safeguards translate into practical accountability.

8 Conclusion and Transition Paths

Digital Public Infrastructure is a claim of political legitimacy. Such legitimacy cannot be derived from self-declarations, "open standards" labeling, or multilateral endorsements. It requires **Corrigibility**: the capacity of the governed to observe the rules, verify their execution, limit their reach, and reproduce the infrastructure to escape its capture.

Five architectural constraints verify this capacity. The analysis demonstrates that systems currently touted as DPI models (Aadhaar, UPI) fail these tests, revealing them to be administrative enclosures rather than public goods. Conversely, the systems which actually run the world (Linux, Kubernetes, Let's Encrypt) pass them structurally.

8.1 The Roadmap Out

The structure of the five tests implies an ordering for transition (detailed in Section 7.5): EXIT → CODE → AUDIT → GOVERN → FORK. This sequence reflects the causal structure of the feedback loop. Interventions at later stages are ineffective if earlier stages are broken.

8.2 The Core Finding

The choice is not between chaos and control. It is between **Open Loop** systems that drift toward fragility and authoritarianism, and **Closed Loop** systems that stabilize themselves through the correction of those they serve.

> *Systems that cannot be corrected by those they affect will eventually be corrected by forces they cannot control.*

8.3 Summary of Topological Risk

The "physics" of digital governance demonstrates that certain architectures structurally increase the risk of capture independent of operator intent. The transition to platform-mediated governance does not eliminate the need for the Rule of Law; it necessitates its translation into the **Rule of the Ledger** (Section 6.5) — and, in the agentic setting the companion paper takes up, the Rule of the Workflow [Aravind, 2026]. Table 13 maps the principal failure modes and structural remedies across identity, payment, and data-exchange layers.

Table 13: Failure modes and structural remedies by infrastructure layer

Layer	Primary Failure Mode	Capture Type	Structural Remedy
Identity	Dependency	Administrative	FEE / Portability
Payment	Lock-In	Transactional	Multi-Rail / Interoperability
Data Exchange	Siloing	Coordination	Open Standards / APIs

Corrigibility is the structural response to the Collingridge Dilemma (Section 6.3): architectural constraints must be established before entrenchment makes correction impossible.

8.4 Limitations

1. **Methodological scope: binary determinations.** The binary determination answers "Does this system merit the claim of public infrastructure?" rather than "Which constraint should reformers address first?"; the continuous metrics in Tables 9 and 10 serve diagnostic prioritization, while the binary output serves the legitimacy threshold. The control-theoretic formalization (Proposition B.1) is labeled in-text as a structural analogy rather than a predictive dynamics claim, and the framework evaluates legitimacy rather than stability: incorrigible systems may persist indefinitely without being legitimate.
2. **Solution validation gaps: FORK, FEE, and prescriptive limits.** FORK may be structurally unachievable for essential state infrastructure: legal monopoly, network effects, data non-portability, and decade-long trust bootstrapping are barriers technical openness alone cannot overcome, and the framework provides no demonstrated example of essential state identity infrastructure disciplined by an actually executed fork. Functional Exit Equivalence (FEE) is proposed as the resolution for the essentiality-corrigibility tension, but no essential identity infrastructure has implemented full FEE; partial analogues (telecom number portability, PSD2 banking APIs) operate in domains with weaker network effects. The framework is therefore primarily diagnostic; the prescriptive elements (FEE, open licensing, multi-stakeholder governance) remain theoretical constructs requiring empirical validation.
3. **Evaluation interpretive limits.** GOVERN requires historical evidence of enforcement and is therefore more interpretation-dependent than the technical tests. The correction-velocity critique targets external bureaucracies (courts, ombudsmen) rather than all deliberative processes; Bitcoin's BIP process and other community-paced governance remain valid governance under the framework.

4. **Network effects are not formally modeled.** Network effects can make formally forkable systems practically unforkable (e.g., Signal's open code but unfederated network). The framework treats this as natural economic friction rather than constructed legal barrier, but does not provide a quantitative model for when network friction crosses into capture.
5. **Empirical scope limitations.** The primary case studies (Aadhaar, UPI) derive from India Stack; Pix and X-Road receive five-test evaluations in Section 5, but systematic validation across further national DPI implementations (Singapore NDI, Kenya M-Pesa, the EU Digital Identity Wallet) is required before universal claims are established. The political economy of why governments choose incorrigible designs is sketched as four mechanisms in Section 6.8; its formal development remains future work.
6. **The framework encodes specific values.** This framework prioritizes structural accountability through exit mechanisms (the capacity to leave or reproduce) over voice mechanisms (political participation, democratic oversight), reflecting a specific intellectual tradition (Hirschman's exit/voice distinction, Stallman's free software principles, commons governance theory). Alternative frameworks prioritizing voice have independent value. The framework is not ideologically neutral; it makes explicit what it values.

8.5 Future Work

Extensions warranting investigation: (1) **Empirical velocity measurement**, quantifying when systems become ungovernable; (2) **International comparison**, testing generality across X-Road, Singapore NDI, EU Digital Identity, Brazil Pix; (3) **Political economy of architecture choice**, developing formally the four mechanisms sketched in Section 6.8; (4) **Transition cost modeling**; (5) **Legal doctrine integration**, operationalizing corrigibility for courts; (6) **Network effect formalization**; (7) **FEE pilot design**, specifying implementable FEE mechanisms for identity infrastructure. A companion paper [Aravind, 2026] extends this framework to learned systems (AI/ML) where behavior is determined by training data rather than source code.

Liability and Enforcement. Architectural corrigibility must be complemented by clear liability and enforcement semantics: for any failure mode, this framework's supply-chain roles (principal, integrator, operator, trustee) must be mappable onto the value-chain roles the law actually distinguishes — provider, deployer, importer, distributor, authorised representative under the EU AI Act [European Parliament and Council of the European Union, 2024] — and to mandated mitigation steps, so that evidence produces enforceable consequences rather than forensic narrative alone.

Adversarial Robustness. This framework assumes good-faith actors operating within institutional constraints. It does not address adversarial manipulation: coordinated gaming of audit mechanisms, governance capture through procedural compliance, or fork-and-poison attacks. Extension to adversarial settings requires additional machinery beyond this paper's scope.

Methodological Note. The five-test structure presented here did not emerge from theoretical deduction; it hardened across successive revisions through adversarial engagement with live systems. Each version closed a specific loophole: self-attestation gave way to independent assessment; partial-credit "maturity models" gave way to strict binary determination; advisory governance was disqualified in favor of non-overrideable constraint; substitution (EXIT) was distinguished from reproduction (FORK); and decentralized identifiers were prevented from laundering centralized governance via layer-decomposed evaluation. The framework's audit trail—version history, schema changelogs, and the rules that govern protocol revision—is maintained in the schema repository.

Core Thesis

Corrigibility defines the boundary between reversible and irreversible infrastructure. It does not guarantee justice. It does not optimize administration. It guarantees that corrective authority remains structurally available to those the system governs. Digital infrastructure aspires to durability. Corrigibility ensures durability does not become capture. Public systems are either structurally reversible or structurally irreversible. Corrigibility marks that boundary.

Acknowledgments

This work draws on the author's lived practice across the three traditions it triangulates: two decades of free-software contribution and engagement with population-scale public digital systems, commons-governance work originating at Moving Republic (https://movingrepublic.org), and a decade of sustained architectural critique during

the adversarial phases of Aadhaar, UPI, and national health-data deployments—alongside long-standing study of cybernetics and the political theory of power. Errors and infelicities remain the author's own.

License

Data and Code Availability

The complete framework, including JSON Schemas and assessment tooling, is available at:

- **Schema Repository**: https://github.com/anivar/corrigibility-schema
- **Protocol Documentation**: PROTOCOL.md, AGENTS.md, docs/ — test and analysis documentation covering EXIT, CODE, AUDIT, GOVERN, FORK, layer decomposition, variety drift, and action boundaries; assessments emit audit.json following Protocol 3.1
- **DPI Schemas**: schema/dpi/infrastructure.json, schema/dpi/audit.json
- **EPI Schemas**: schema/epi/infrastructure.json, schema/epi/audit.json (see companion paper [Aravind, 2026])

Reading the Appendices

The appendices that follow ground the framework's claims in formal machinery. Appendix A develops the adversarial-manifest logic that converts policy assertions into machine-readable claims. Appendix B formalizes the structural model and proves the corrigibility invariant. Appendix C proves the neutrality-compression result for centralized enforcement at scale. Appendix D states the verification and enforcement rules (Rule A.9 on binding authority, Rule A.10 on chain of custody, and the tiered identity-assurance model) that operationalize GOVERN and AUDIT. Appendix E addresses the strongest anticipated objections. Appendix F consolidates terminology. Full JSON schemas, the protocol's version history, and reference implementations are maintained as a versioned external artifact in the schema repository (see Data and Code Availability above); the rules and definitions on which the paper's argument actually depends remain in the appendices below.

A Architecture of Accountability

To move beyond qualitative debate, this framework establishes a formal logic for accountability. The machine-readable artifacts described below serve not merely as data formats, but as a *digital constitutional binding* that separates operator claims from auditor verification. By enforcing specific data structures, we convert policy rhetoric into falsifiable assertions of system state.

A.1 Logic: The Adversarial Manifests

The architecture relies on a separation of concerns mirroring the adversarial legal process. It employs two distinct document types (see Figure 6):

1. **infrastructure.json (The Operator's Affidavit):** A signed declaration by the system operator asserting specific operational properties, governance commitments, and functional limitations.
2. **audit.json (The Auditor's Finding):** An independent, cryptographically signed evaluation that tests the operator's claims against observable reality.

This architecture makes contradictions machine-readable. If an operator claims a governance mechanism is "binding" in their affidavit, but the auditor's observation marks the `user_authority` field as "advisory," the mismatch serves as mathematically verifiable proof of governance failure.

Schema Encoding. The full normative JSON schemas implementing the affidavit/finding distinction, with field-by-field anchors back to the paper sections that justify them and structural enforcement of Rule A.9 (Appendix D) via JSON Schema conditionals, are maintained as a versioned artifact in the schema repository (https://github.com/anivar/corrigibility-schema).

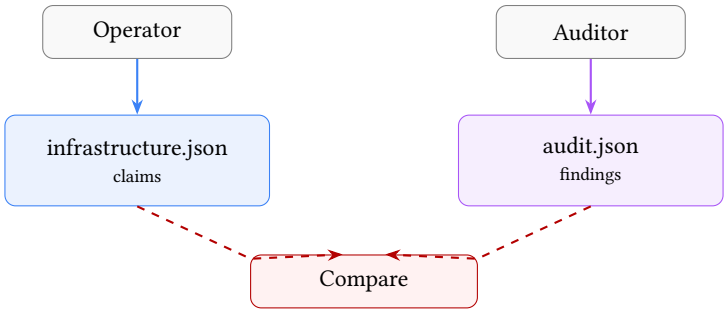

Figure 6: Adversarial Verification Architecture. Operators declare; Auditors verify. The structural gap constitutes the audit finding.

A.2 Interpretation: The Non-Authoritativeness of Determination

A critical logic gate in this framework is the derivation of status.

Claim A.1 (Derived, Non-Authoritative Determination)**.** *The* `determination` *object in* `audit.json` *is a derived convenience field. It is* ***not authoritative****.*

Any value asserted in `determination` MUST be mechanically recomputed from the `tests` object by evaluators or courts. If an auditor asserts `corrigible: true` but the `tests` object contains a failure, the manifest is strictly invalid. This rule ensures that no actor (operator, auditor, or regulator) can "declare" a system corrigible by fiat; status exists only as the computed sum of observable structural properties.

B Formal Structural Model

Executive Summary (Non-Technical Overview). This appendix formalizes two core claims of the paper using control theory as a structural model:

1. **Null-Feedback Instability:** If any component of the feedback path is structurally severed, the system operates as open-loop and will diverge under persistent disturbance.
2. **Correction Velocity Inequality:** If the rate of digital error propagation exceeds the rate of institutional correction, post-hoc remedies are insufficient to prevent systemic failure.

This formalization illustrates structural relationships between the five tests and feedback system properties. It does not claim that DPI literally behaves as a linear time-invariant system; the model is pedagogical, clarifying *one half* of why partial compliance represents structural failure rather than partial success. The other half is the political-theoretic argument developed in Section 7.1.2: even where partial feedback transmits *some* corrective signal, partial contestation produces sham governance under a non-domination criterion. The binarity of corrigibility rests on the conjunction of the control-theoretic and political-theoretic arguments; this appendix supplies only the former.

To formalize the assertion that "partial corrigibility is a fatal structural failure," we model Digital Public Infrastructure as a standard feedback control system subject to environmental disturbance.

B.1 Open-Loop Instability Model

Let a DPI system state $y(t)$ (actual impact) be expected to track a reference target $r(t)$ (intended policy/rights), subject to environmental disturbance $d(t)$ (attacks, changing demographics, fraud).

The error signal (grievance) is defined as:

$$e(t) = r(t) - y(t) \tag{8}$$

In a closed-loop system, the corrective action $u(t)$ (GOVERN) is a function of the detected error via a feedback controller K (GOVERN logic) and a sensor transfer function H (AUDIT/EXIT mechanism):

$$u(t) = K(t) \cdot H(e(t)) \tag{9}$$

The plant dynamics (DPI operation) are given by:

$$\dot{y}(t) = u(t) + d(t) \tag{10}$$

Proposition B.1 (Null-Feedback Instability — Structural Analogy)**.** *If any component of the feedback path is structurally broken, specifically if Sensor Transparency $H = 0$ (Failed AUDIT) or Actuator Coupling $K = 0$ (Failed GOVERN), the system effectively operates as Open Loop. Under Open Loop conditions with non-zero integrating disturbance $d(t)$, the error $e(t)$ diverges unboundedly.*

Proof. We treat this as a structural analogy rather than a stability theorem for any specified DPI implementation; the formal apparatus illustrates the consequence of broken loop closure under generic disturbance, not the predicted dynamics of a particular system.

If tests **AUDIT** or **EXIT** fail, $H = 0$. If test **GOVERN** fails, $K = 0$. In either case, the effective corrective feedback is $u_{\text{fb}} = 0$. The system evolution depends only on initial intent u_{open} and disturbance:

$$y(t) = \int_0^t (u_{\text{open}}(\tau) + d(\tau))d\tau \tag{11}$$

Given that environmental variety $d(t)$ is stochastic and non-zero (Ashby's Law), and without negative feedback to compensate:

$$\lim_{t\to\infty} |r(t) - y(t)| \to \infty \tag{12}$$

Thus, structural stability is unattainable without the closure of the feedback loop. "Partial" corrigibility (where $K = 0$ or $H = 0$) yields the same asymptotic divergence profile as fully open-loop operation. □

Corollary B.1 (Binarity of Corrigibility)**.** *Corrigibility is a binary structural property, on two clauses. Mechanism clause: if EXIT, AUDIT, or GOVERN fails, a loop-gain term vanishes, the loop opens, and systemic error cannot be structurally bounded. Verifiability clause: if CODE or FORK fails, correction becomes unverifiable or unenforceable in extremis — the plant is unintelligible or irreplaceable — so corrective authority becomes discretionary rather than guaranteed. Under either clause the determination is failure.*

Justification: Each test corresponds to a necessary component of the corrective apparatus (Figure 7). The corollary rests on the conjunction of two independent arguments developed in Section 7.1.2. *Mechanism (this appendix):* from Proposition B.1, since $L(s) = K(s) \cdot P(s) \cdot H(s)$, if any multiplicative term is zero the loop gain vanishes and partial compliance does not preserve closure. *Legitimacy (main text):* partial contestation produces *sham governance* — accountability with non-monotonic returns where additional procedural gestures may absorb political pressure without yielding correction — so even where partial feedback transmits some signal, the system fails the non-domination criterion (Section 3, Normative Anchor). The mechanism argument alone yields divergence-rate claims; the legitimacy argument alone yields institutional-status claims; the conjunction yields binarity.

Remark B.1 (The Gradient Quantifier: Corrigibility at the Margin)**.** *The Corollary is evaluated at the bottom of the gradient. Proposition B.1 treats H and K as system constants. In population-scale infrastructure, they are distributions over participant position x in the social hierarchy: $H(x)$ is the probability that a grievance from position x generates a registered error signal, and $K(x)$ is the probability that a registered error triggers binding correction visible to x. Both are lowest at the least-resourced stratum — depressed by digital literacy, language, connectivity, time, administrative precarity, and fear of retaliation.*

Loop closure must hold for all participants, not on average:

$$\textit{LoopClosed}(S) \iff \forall x \in \mathcal{X} : H(x) > 0 \ \wedge \ K(x) > 0 \tag{13}$$

The binding determination is taken at $x^* = \arg\min_x \big[H(x) \cdot K(x)\big]$, not at the population mean. A system that closes the loop for median users while $H(x^*) = 0$ or $K(x^*) = 0$ at the margin satisfies the formal model for the comfortable majority and fails it entirely for those the framework's own evidence (Section 5) shows are most harmed by incorrigibility.

Operationally: proxy metrics and audit sampling frames must be reported as distributions. Pass/fail determination is taken at a specified lower quantile of the affected population (recommended: p_{10} stratum, or the stratum documented as least-resourced in the operator's disclosure), not at the mean or aggregate rate. An audit that samples only median-resource users has not assessed whether the loop is closed; it has assessed whether the loop is closed for a particular sub-population that is structurally unlikely to include the framework's central case.

The closed-loop transfer function from reference to output is:

$$\frac{Y(s)}{R(s)} = \frac{K(s)P(s)}{1 + K(s)P(s)H(s)} \tag{14}$$

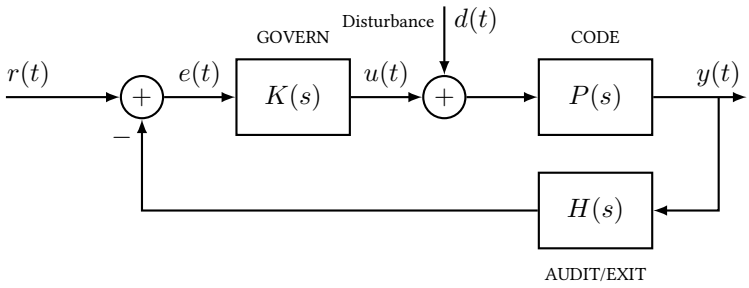

Figure 7: **Control-theoretic model of corrigible infrastructure.** $K(s)$: governance controller. $H(s)$: sensor/audit function. $P(s)$: plant (infrastructure). $d(t)$: environmental disturbance.

The transfer function from disturbance to output is:

$$\frac{Y(s)}{D(s)} = \frac{P(s)}{1 + K(s)P(s)H(s)} \tag{15}$$

Corrigibility requires that each component exists and is non-degenerate:

- EXIT $\Rightarrow e(t) \neq 0$ observable (error signal exists)
- AUDIT $\Rightarrow H(s) \neq 0$ (sensor functional)
- GOVERN $\Rightarrow K(s) \neq 0$ (actuator functional)
- CODE $\Rightarrow P(s)$ is intelligible
- FORK $\Rightarrow P, K, H$ replaceable if convergence fails

Active Deception as Sensor Attack. The pathologies defined in Section 6.4.2 (open-washing, safety-washing, audit theater) fundamentally corrupt the sensor function $H(s)$: they either suppress the true error signal $e(t)$ or inject artificial stability metrics, thereby blinding the controller $K(s)$ to the actual divergence $y(t) - r(t)$. In control-theoretic terms, Active Deception renders $H(s) \to 0$ not by removing the sensor, but by making its output uncorrelated with the true system state. Figure 8 depicts the resulting open-loop topology in which a corrupted or absent $H(s)$ severs the feedback path.

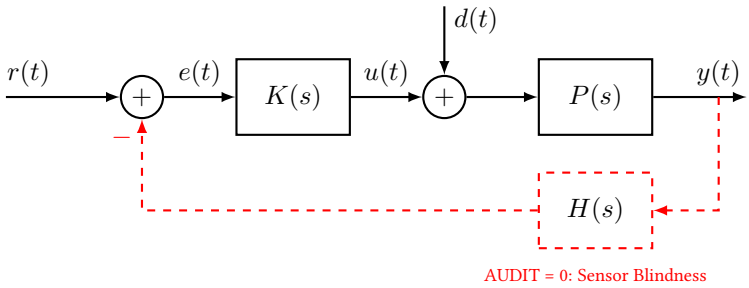

Figure 8: **Open-loop failure mode.** When AUDIT fails ($H = 0$), the feedback path is severed. Equivalent failure occurs if GOVERN ($K = 0$) or EXIT ($e(t) \equiv 0$) is removed.

B.2 The Strict Correction Velocity Inequality

For the system to remain stable not just asymptotically but locally (preventing irreversible harm), the correction rate must dominate the error rate. We formalize this with a friction coefficient.

Condition B.1 (Strict Dynamic Stability). *Let $\epsilon > 0$ represent the "friction cost" of correction (e.g., litigation time, protest cost). A system is stable iff:*

$$\left| \frac{d}{dt} C(t) \right| \geq \left| \frac{d}{dt} E(t) \right| + \epsilon \tag{16}$$

The symmetric upper bound completes the principle stated in Section 6.3: correction faster than it can itself be verified enables manipulated "corrections" to bypass review, so $|dC/dt| \leq V_{verify}$, where V_{verify} is the rate at which corrective actions can be independently checked. Deliberative latency is not friction to be minimized past this bound; it is the verification budget.

Implication: Post-hoc remedies (Courts/Ombudsmen) have a linear or constant rate of operation ($\frac{dC}{dt} = k$). Digital errors propagate at exponential or high-frequency rates ($\frac{dE}{dt} \propto e^t$). Since for any constants there exists a time beyond which the exponential rate exceeds any linear rate — and does so permanently — relying on external courts guarantees eventual, unbounded accumulation of unresolved harm (Figure 9). Where the execution substrate settles irreversibly, the inequality binds at a hard deadline as well as a rate: correction must reach a decision before the substrate makes it final, after which no correction velocity suffices. The correction logic must be intrinsic to the loop.

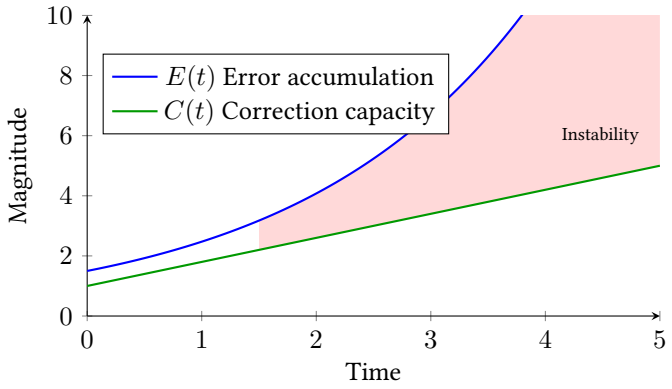

Figure 9: **Correction velocity mismatch.** Error $E(t)$ accumulates faster than correction $C(t)$ in bureaucratic systems. The shaded region represents accumulated unresolved harm.

B.3 Mapping to the Five Tests

Each of the five corrigibility tests corresponds to a distinct component of the closed-loop control structure; the failure mode under each test follows directly from the role of that component (Table 14).

Table 14: Mapping corrigibility tests to control-theoretic components

Test	Control Component	Failure Mode
EXIT	Error Signal	Reference collapse
CODE	Signal Integrity	Noise indistinguishable from intent
AUDIT	Sensor	Blind control
GOVERN	Actuator	No corrective force
FORK	Variety Expansion	Monolithic collapse

B.4 Measurement Procedures for Reproducibility

To enable third-party reproduction of pass/fail determinations, each test requires specific evidence types (Table 15).

Reproducibility Standard. An assessment is reproducible if an independent auditor, given only the evidence artifacts listed in Table 15, reaches the same pass/fail determination. Disagreements must trace to (a) threshold calibration differences or (b) missing evidence, not to ambiguity in the test definitions themselves. Continuous metrics and threshold protocols are provided in Tables 9 and 10 (Section 7.1).

B.5 Restatements of the Instability Result

The proposition above admits several equivalent restatements that may be more familiar to readers from different traditions. None is independently load-bearing; each makes the same instability claim from a different angle (frequency domain, discrete time, cumulative rates).

Table 15: Evidence requirements for reproducible assessment

Test	Required Evidence	Verification Method
EXIT	(1) Legal mandate status, (2) documented alternative pathways, (3) penalty documentation	Survey of legal instruments; cost comparison with alternatives; user testimony
CODE	(1) Source repository URL, (2) build reproducibility proof, (3) execution path coverage	Hash verification; deterministic build; static analysis of critical paths
AUDIT	(1) API documentation, (2) data access agreement, (3) historical error rate publication	Attempt independent measurement; verify data freshness; compare operator vs independent findings
GOVERN	(1) Charter/statute URL with hash, (2) governance action log, (3) enforcement history	Legal analysis; search for overruled operator decisions; verify binding mechanism
FORK	(1) License text, (2) technical artifact availability, (3) resource cost estimate, (4) successful fork examples	License audit; attempt reproduction; cost modeling; historical fork survey

Frequency-domain restatement. For small perturbations around a nominal operating point, the sensitivity function $S(s) = 1/(1 + L(s))$ with loop gain $L(s) = K(s)P(s)H(s)$ characterizes disturbance propagation. Whenever any corrigibility test fails ($K = 0$, $H = 0$, or the error signal is blocked), $L(s) \to 0$ and $S(s) \to 1$: disturbances pass through unattenuated. This is the same instability claim of Proposition B.1, expressed in the frequency domain.

Discrete-time restatement. Consider a discrete-time scalar integrator $y_{k+1} = y_k + d_k - u_k$, where d_k is a disturbance and u_k is the correction. In open loop ($u_k = 0$), $y_N = y_0 + \sum_{k=0}^{N-1} d_k$, and for any non-zero mean disturbance $\mathbb{E}[d_k] = \mu > 0$ the expected state grows linearly: $\mathbb{E}[y_N] = y_0 + N\mu \to \infty$. This provides an elementary, discrete demonstration of the same divergence result.

Friction cost in the discrete model. The friction coefficient ϵ from Condition B.1 manifests here as a delay or reduction in the correction term. If correction requires bureaucratic processing with latency τ time steps and efficiency loss $\eta < 1$, then $u_k = \eta \cdot f(y_{k-\tau}, d_{k-\tau})$ for some feedback policy $f(\cdot)$. When $\tau > 0$ or $\eta < 1$, the effective correction rate is reduced, directly corresponding to the continuous-time friction cost $\epsilon > 0$. The stability condition $|dC/dt| \geq |dE/dt| + \epsilon$ translates to requiring the feedback policy to over-correct relative to the disturbance rate, accounting for these frictional losses.

Cumulative-rate restatement. Letting $E(t)$ and $C(t)$ denote cumulative error and cumulative correction, stability requires $E(t) - C(t) < M$ for some bound M and all $t > 0$, equivalently $\dot{C}(t) \geq \dot{E}(t)$ on average. When correction operates through channels with bounded throughput $\dot{C}_{\max}$ while errors grow super-linearly, a critical time t^* exists beyond which correction can never catch up. This is Condition B.1 restated in cumulative form.

B.6 Functional Exit Equivalence Formalization

An essential system S satisfies FEE if compensatory architectural guarantees recreate the same effective error-signal strength that exit provides in non-essential systems.

Let:

- $E_{\text{exit}}^{\text{baseline}}$ = effective error-signal strength produced by real exit (non-essential case)
- $E_{\text{alt}}(S)$ = effective error-signal strength produced by compensatory mechanisms (essential case)

$$\text{FEE}(S) \iff E_{\text{alt}}(S) \geq \theta_E \cdot E_{\text{exit}}^{\text{baseline}} \tag{17}$$

where θ_E is a policy calibration constant (recommended range: 0.8–1.0).

Operationalizing E. The error-signal strength E can be operationalized as a weighted sum:

$$E = \alpha_1 \cdot \text{effective\_loss} + \alpha_2 \cdot \text{probability\_of\_departure} + \alpha_3 \cdot \text{velocity\_of\_response} + \alpha_4 \cdot \text{legal\_remedy\_speed} \tag{18}$$

where the weights α_i are calibrated via pilot audits. All variables are operationalized such that higher values indicate greater corrective pressure on the operator.

B.7 Switching Cost Formalization

The practical barrier to exit and fork is operationalized through a **switching cost** scalar:

$$\sigma(S \to S') = c_{\text{data}} + c_{\text{downtime}} + c_{\text{reprovisioning}} + c_{\text{network}} + c_{\text{legal}} \tag{19}$$

where each component is normalized to $[0, 1]$:

- c_{data}: cost of exporting and reformatting user data
- c_{downtime}: service interruption during transition
- $c_{\text{reprovisioning}}$: re-establishing credentials and trust relationships
- c_{network}: loss of social connections and network effects
- c_{legal}: regulatory friction and contractual barriers

Portability Threshold. EXIT and FORK functionally fail when switching cost to all plausible alternatives exceeds a policy threshold:

$$\min_{S' \in \mathcal{A}} \sigma(S \to S') > \Sigma^* \implies \text{EXIT/FORK fail} \tag{20}$$

where $\mathcal{A}$ is the set of available alternatives and Σ^* is a calibrated threshold on the normalized $[0, 5]$ scale of σ (e.g., $\Sigma^* = 2.0$). Practical anchors — switching costs exceeding 0.2 of annual disposable income, or 30 days' service disruption — calibrate the *individual* components c_i to their maximal value of 1; they are normalization anchors, not alternative values of Σ^* itself.

Illustrative Application. Aadhaar exhibits $\sigma \approx 4.8$ (near-maximal) because $c_{\text{data}} = 1$ (no export), $c_{\text{downtime}} = 1$ (total service denial), and $c_{\text{legal}} = 1$ (mandate forecloses alternatives). UPI retains $\sigma \approx 1.0$–1.5 because cash and cards remain legal. With $\Sigma^* = 2.0$, Aadhaar categorically fails on the switching-cost criterion, while UPI does not fail on that criterion — though UPI still fails FORK on the constructed-barrier dimension (Section 4: NPCI's regulatory monopoly), since $\sigma < \Sigma^*$ licenses no pass on its own (Equation 20 is a one-way implication).

B.8 Summary of Formal Extensions

1. Linearized restatement ties removal of H, K, or EXIT to vanishing loop gain $L(s)$, exposing unstable process modes.
2. Discrete integrator supplies an elementary divergence demonstration and absorbs Condition B.1 through its latency and efficiency-loss parameters (τ, η), which correspond to the continuous friction cost ϵ.
3. Correction velocity is formalized via Condition B.1 and its cumulative-rate restatement.
4. FEE formalization (eq. 17) operationalizes the principle of essential-services compensation; weights α_i are calibrated externally.
5. Switching-cost formalization (eq. 19, eq. 20) operationalizes EXIT/FORK failure thresholds.

C Neutrality Compression Under Centralized Enforcement

This appendix provides the formal proof supporting Proposition 6.1 (Topology-Dependent Tension).

C.1 Definitions

Let:

- Sc = Scale (population-level integration)
- $V_{\text{env}}(Sc)$ = Environmental variety (heterogeneity of governed interactions)
- $V_{\text{gov}}(Sc)$ = Governance corrective variety

We assume:

1. $V_{\text{env}}(Sc)$ is non-decreasing in Sc (Ashby's Law: environmental complexity grows with population scale)
2. Under centralized enforcement, governance growth is institutionally bounded (bureaucratic capacity constraints)

Formally:

$$V_{\text{gov}}(Sc) = O(Sc^{\alpha}), \quad \alpha < 1 \tag{21}$$

This states that governance corrective capacity grows sublinearly under centralized topology. No specific functional form (logarithmic, square root, etc.) is required. The sublinear constraint reflects institutional reality: adding staff, building appeal mechanisms, and processing edge cases cannot scale linearly with population without structural reform.

C.2 Neutrality Condition

Neutrality does not require full dominance of environmental variety. It requires sufficient corrective bandwidth to prevent systematic routing bias.

Define:

$$V_{\text{gov}} \geq \theta \cdot V_{\text{env}} \tag{22}$$

Where $\theta \in (0, 1]$ is a stability sufficiency constant.

Neutrality degradation risk emerges when:

$$V_{\text{gov}} < \theta \cdot V_{\text{env}} \tag{23}$$

This avoids equating neutrality with perfect stability.

C.3 Proposition (Formal Statement)

Under centralized enforcement topology, there exists a scale threshold Sc^* beyond which neutrality is at risk unless governance growth becomes at least linear in scale.

C.4 Proof

Assume environmental variety grows at least linearly with scale:

$$V_{\text{env}}(Sc) = \Omega(Sc) \tag{24}$$

Under centralized enforcement:

$$V_{\text{gov}}(Sc) = O(Sc^{\alpha}), \quad \alpha < 1 \tag{25}$$

Therefore:

$$\frac{V_{\text{gov}}(Sc)}{V_{\text{env}}(Sc)} \to 0 \quad \text{as} \quad Sc \to \infty \tag{26}$$

Thus there exists Sc^* such that:

$$V_{\text{gov}}(Sc^*) < \theta \cdot V_{\text{env}}(Sc^*) \tag{27}$$

The neutrality condition is violated. □

C.5 Governance Automation Objection

One may argue that governance capacity scales linearly through automated monitoring. However:

If governance automation is implemented within the same centralized topology, then:

- The automation itself requires governance oversight
- External corrective independence does not increase proportionally
- Appeal mechanisms remain institutionally bounded

Thus governance automation does not automatically imply:

$$V_{\text{gov}}(Sc) = \Theta(Sc) \tag{28}$$

unless corrective mechanisms are externally independent and structurally distributed.

This preserves the proposition.

C.6 Falsifiability Clause

The theory would be falsified if a centralized sovereign DPI:

1. Achieves population-scale deployment
2. Maintains governance growth $V_{\text{gov}}(Sc) = \Theta(Sc)$
3. Sustains structural neutrality under high scale

without federated or multi-operator architecture.

Such a system would contradict the sublinear governance constraint. This condition is empirically testable.

C.7 Interpretation

The result does not assert:

- That scale is harmful
- That sovereignty is undesirable
- That neutrality is impossible

It asserts:

- Under centralized enforcement topology, governance corrective capacity must scale at least linearly with environmental complexity to preserve neutrality.
- Sublinear governance growth is a structural prediction of centralized architecture; neutrality degradation risk follows mathematically.

Figure 10 visualizes the compression: as scale rises, environmental variety grows linearly while centralized governance variety grows at most sublinearly, so the neutrality-sufficiency threshold $\theta \cdot V_{\text{env}}$ is crossed at some scale Sc^*.

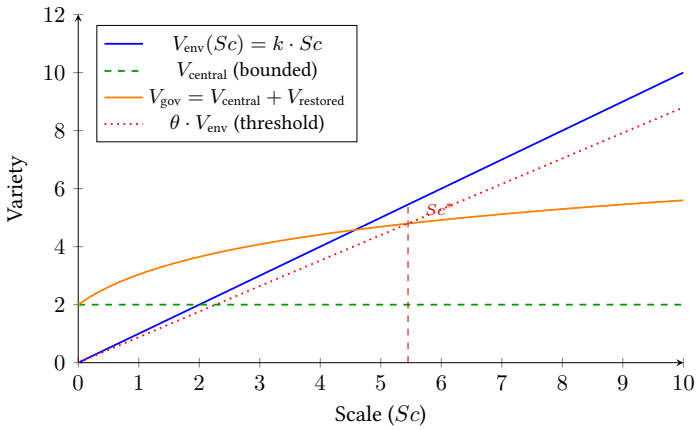

Figure 10: **Governance Variety Compression Under Scale.** As scale increases, environmental variety grows linearly. Under centralized enforcement, governance variety grows sublinearly (logarithmic here); V_{restored} denotes the variety contributed by the compensatory mechanisms of Section 6.6 (federated execution, multi-issuer mandates, FEE, state portability). Neutrality degradation risk emerges when governance variety falls below the threshold $\theta \cdot V_{\text{env}}$ (at Sc^*).

D Verification and Enforcement Rules

To distinguish constitutive governance from performative compliance, the framework states evidentiary rules for who may speak with binding authority and whose audit verdicts count as truth. These rules close the GOVERN and AUDIT loops: without them, the operator's affidavit would float free of the auditor's finding, and the corrective channel would collapse back into narrative.

D.1 Normative Rules

Rule A.9: Operational Definition of Binding Authority To satisfy the GOVERN test, any escalation path claimed as "binding" in the Manifest MUST include: (i) a resolvable reference to the statute or charter (`binding_proof_url`), (ii) a cryptographic hash of that instrument to prevent silent amendment, and (iii) historical evidence of enforcement against the operator. Assertions of authority that rely on post-execution bureaucratic discretion (e.g., ombudsmen, "feedback forms") do not satisfy this rule and must be marked `false`. This rule is what disqualifies advisory bodies from constituting governance, and is the operational test invoked when the main text concludes that systems such as Aadhaar fail GOVERN.

Rule A.10: Chain of Custody Auditor manifests must be canonically serialized (RFC 8785), hashed (SHA-256), and signed (Ed25519) using a valid W3C Decentralized Identifier. Unsigned assessments are treated as schema-invalid. This creates a permanent, falsifiable record of who certified a system, and is what makes an audit verdict independently checkable rather than a narrative claim.

Rule A.11: Auditor Data Duties The AUDIT guarantee ("no one can prevent auditing," Section 4) does not imply unlimited data access by auditors. Audit access must be purpose-bound: auditors access minimized, purpose-specific views of system behavior — population-level aggregates, stratified error rates, sampled records under a defined privacy budget, or replay against pseudonymized logs with subject receipts (see the Subject Receipt Requirement in the companion paper [Aravind, 2026]) as the ground-truth check. The receipt obligation attaches to the rights-affecting determination itself, not to the agentic substrate: deterministic systems, which have no action boundary, discharge it at the effect surface, where the determination becomes a state change in the record the subject must live under. Auditor manifests must declare the data accessed, the purpose, and the retention and deletion schedule for any individual-level records obtained; this declaration is itself governed by Rule A.10 and is auditable. Violations — data accessed beyond stated purpose, retained past the declared window, or shared with parties outside the audit engagement — revoke the auditor's DID standing. "No one can prevent auditing" and "no one can strip-mine the governed" are not in tension; they are the two-sided requirement of the AUDIT test: permissionless access to the operator's behavior, purpose-bounded access to the governed's data. Any audit regime that satisfies the first condition while ignoring the second has extended legibility downward to the very population the test exists to protect, reproducing the asymmetric-observability failure the AUDIT test exists to correct (Section 4).

D.2 Tiered Identity Assurance Model

To prevent "identity washing," where operators use ephemeral keys to rubber-stamp their own systems, the framework employs a tiered assurance model. Verification dashboards should weigh assessments accordingly:

1. **Tier 1: Self-Attested (`did:key`):** Ephemeral identity. Suitable for whistleblowers, individual researchers, or rapid-response checks. *Status: informational.*
2. **Tier 2: Domain-Verified (`did:web`):** Identity cryptographically bound to a specific DNS domain (e.g., `did:web:eff.org`). Attribution to known civil society actors. *Status: verified.*
3. **Tier 3: Institution-Verified:** Identity anchored in transparency logs or institutional trust registries. *Status: regulatory standing.*

This tiering is what allows the AUDIT test to distinguish credible independent verification from operator self-attestation; full schema encodings of the tiers, together with the DID resolution and canonicalization details, are maintained in the schema repository (`https://github.com/anivar/corrigibility-schema`).

E Anticipated Objections and Responses

This appendix addresses the strongest objections to the framework, organized by the discipline from which they originate.

E.1 Control Theory Objections

Objection 1: Binary model ignores continuous feedback dynamics. "Real control systems have continuous loop gain. Partial observability and partial actuation still provide some regulation. Your binary pass/fail ignores systems that are 'mostly corrigible.'"

Response: The objection conflates physical description with normative threshold. We do not dispute that real systems exhibit continuous variation in feedback quality. The binary model is a *legitimacy threshold*, not a physical claim (Section 7). A feedback loop with 50% sensor accuracy transmits some signal; this does not mean the system deserves public trust. The determination function D (Section 7.1) makes explicit that continuous inputs feed a binary output. The threshold is where we draw the line for public legitimacy, separable from the physics.

Objection 2: Where is the Lyapunov stability proof? "You claim systems diverge without feedback. Show the Lyapunov function. Without rigorous stability analysis, this is hand-waving."

Response: Appendix B provides the formal treatment. Proposition B.1 (Null-Feedback Instability — Structural Analogy) shows that when $H = 0$ or $K = 0$, the system operates open-loop and diverges under non-zero disturbance. The discrete integrator example (Section B.5) provides an elementary proof: $y_N = y_0 + \sum d_k \to \infty$ for any non-zero mean disturbance. A full Lyapunov treatment would require specifying the state space for a particular DPI implementation; the framework operates at the architectural level, where the key insight (that open loops diverge) is mathematically trivial.

E.2 Political Science Objections

Objection 3: You ignore legitimate state interests. "Governments have valid reasons for mandatory systems: national security, public health, anti-fraud. Your framework would prohibit pandemic contact tracing."

Response: The framework does not prohibit mandatory systems; it diagnoses their structural properties. Emergency necessity explains but does not justify incorrigibility (Corollary 6.1). A pandemic contact-tracing system can still satisfy CODE (open source), AUDIT (independent testing), GOVERN (binding privacy constraints), and FORK (interoperable protocol). Only EXIT is constrained by mandatory participation. The question is whether the other four tests pass, not whether mandates are ever legitimate. Section 7 explicitly notes that corrigibility is compatible with strict, non-negotiable rules.

Objection 4: The Essentiality-Corrigibility Tension proves your framework is useless. "You prove essential services can never be fully corrigible. This makes your framework irrelevant for the systems that matter most: identity, payments, healthcare."

Response: The proposition identifies a structural constraint, not a reason to abandon analysis. The architectural responses (Section 6.7.3) show two paths: protected alternatives (guarantee non-digital equivalents) or federated implementation (no single provider essential). X-Road demonstrates 4/5 at population scale. The finding that fully corrigible essential services require architectural innovation is the framework's central contribution. Diagnosing why current systems fail is prerequisite to designing better ones.

E.3 Empirical Objections

Objection 5: Your pass/fail determinations are subjective. "Two auditors could reach different conclusions. Without precise, reproducible metrics, this is opinion dressed as science."

Response: Section B.4 specifies evidence requirements for each test. Table 15 lists required artifacts and verification methods. The reproducibility standard states: an assessment is reproducible if an independent auditor, given only the evidence artifacts, reaches the same determination. Disagreements must trace to threshold calibration or missing evidence, not definition ambiguity. Rule A.10 (Appendix D) binds every audit verdict to a canonically signed manifest, making disagreements traceable rather than narrative; the JSON schemas in the schema repository encode the full field set across all five tests. Future work includes inter-rater reliability studies.

E.4 Summary

The framework withstands scrutiny from control theory (binary is threshold, not physics), political science (mandates can coexist with corrigibility on other dimensions), and empirical methodology (reproducible metrics exist). The strongest residual objection (network effects making FORK practically impossible even when legally permitted) is acknowledged in Limitations and proposed for formal treatment in Future Work.

F Glossary of Key Terms

The glossary below is shared verbatim with the companion paper [Aravind, 2026] so that terminology is identical across the pair.

Accountable-Authority Checkpoint
The checkpoint requirement for high-stakes tiers, functional rather than positional: an accountable authority whose grant is fresh, whose liability binds, and whose correction operates within the correction window, whether it reviews each decision, each class of decisions, or the boundary itself.

Action Boundary
The deterministic envelope that wraps stochastic inference in an agent system. Inference outputs (action proposals) must pass through a hard-coded validation layer before execution. The action boundary is the architectural locus of GOVERN in agentic systems.

Action Boundary Protocol
The set of five architectural requirements (deterministic validation, context-window isolation, exception handling, machine-readable specifications, binding constraints) that an agent system's action boundary must satisfy to support corrigibility. Independent of any specific external standard.

Agent System
The tuple $A = (M, H, B, S, T, C, \bar{M})$: model, orchestration harness, action boundary, action specifications, tool definitions, deployment-time configuration, and persistent memory layer. Behavior is a property of the system A, not the model M; each component bears distinct governance requirements and can capture or release sovereignty independently of the others.

AUDIT
Test 3 of corrigibility. Independent verification: third parties can verify that the system's behavior matches its disclosed specification without requiring operator authorization, through tamper-evident logs, statistical bounds, or drift monitoring.

Citizen-Constraint Registry
A channel through which recognized subject-class representatives lodge protective constraints that the validator must evaluate before executing determinations in the certified class. The citizen-side counterpart of the recognized court's halt flag: structural standing at the boundary, not only recourse after the fact.

CODE
Test 2 of corrigibility. Inspectability of the system's execution. For deterministic DPI, this is source-code disclosure; for learned systems (EPI) it is the LWD-R requirement.

Collective-Standing Precondition
Ostrom's seventh design principle applied to GOVERN: the affected population's capacity to organize, aggregate claims, and fund representation must be legally protected and historically exercised at the relevant stratum. Where this precondition fails, GOVERN fails at that stratum.

Compute Capture
A FORK failure mode specific to learned systems. Occurs when the compute cost to retrain a functionally equivalent model (C_{train}) exceeds the compute accessible to non-operator actors ($C_{\text{accessible}}$) by more than a policy-determined multiplier κ.

Corrigibility
The structural capacity of those affected by a system to detect error, signal harm, and trigger correction without incurring material loss or irreversible consequence. Not a moral preference but an architectural stability requirement.

Data Capture
A FORK failure mode parallel to Compute Capture. Occurs when the data required to train a functionally equivalent model (D_{required}) exceeds the data accessible to non-operator actors ($D_{\text{accessible}}$) by more than a policy-determined multiplier κ_D. Encompasses synthetic data, distilled outputs, RLHF traces, reward models, and proprietary evaluation sets.

Deployment Variables
Where the accountable authority sits, how many agents compose the system, how deeply orchestration nests, at what scale it runs, and what species of actor exercises each corrective function. The tests bind over all of them, fixing chain termini, checker independence, and grant freshness rather than geometry. Distinct from the inference-time deployment variables of Operative Representation (quantization, pruning, routing, filtering, and the retrieval pipeline where present), which the R disclosure must enumerate.

Digital Public Infrastructure (DPI)
Shared digital systems — ledgers, registries, payment rails, data exchanges — intended to deliver public services at population scale. The subject of the deterministic-systems paper in this pair.

Dual Exercise of the Tests
The two exercises each test admits: inward, by parties at or inside the operator's institutional perimeter, and outward, by the subjects the system decides about. A test discharged only inward has been verified for the operator, not for the governed.

Effect Surface
The resource that receives an action's effect, where an untyped action becomes typed again. Where the action channel is untyped, boundary enforcement migrates to the effect surface: some deterministic, specification-checkable gate must stand between inference and irreversible effect.

Epistemic Public Infrastructure (EPI)
Public infrastructure whose behavior is generated by learned parameters rather than explicit logic. Includes AI-based identity, eligibility, triage, and orchestration systems. The subject of the learned-systems paper in this pair.

EXIT
Test 1 of corrigibility. Reversibility of participation: affected parties can refuse or leave the system without prohibitive penalty, or verified Functional Exit Equivalence is demonstrated when literal exit is infeasible. For agentic deployments the test decomposes into five layers (Subject, Memory, Workflow, Operator, Bystander); failing any layer fails EXIT.

FORK
Test 5 of corrigibility. Independent reproduction: parties outside operator control hold the legal permission, the public artifacts, and the user-state portability needed to recreate a functionally equivalent instance. Constructed legal barriers to reproduction disqualify; natural economic friction (capital, network effects) does not.

Functional Exit Equivalence (FEE)
Architectural guarantees that recreate the error-signal strength of literal exit when literal exit is impossible (e.g., for monopoly identity systems). FEE is achieved through multi-issuer mandates, credential acceptance diversity, and legally protected statutory fallbacks; verified FEE discharges the EXIT test for essential systems (under agentic deployment, its Subject layer).

GOVERN
Test 4 of corrigibility. Constitutive constraint: affected populations have binding authority over the system, not merely advisory input. For deterministic DPI this is RFC/charter governance; for EPI it is the action boundary protocol combined with citizen-side mechanism families (juridical, distributed/class-action, deliberative).

Governance Denial of Service (GDoS)
Failure mode in agentic systems where action throughput outruns governance correction velocity. Without explicit termination criteria and bounded action specifications, agents loop indefinitely or escalate beyond legitimate authority.

Gradient Quantifier
The evaluation rule that takes pass/fail determinations at the least-resourced stratum of the affected population, the argmin of detection and correction probability, rather than at the mean. An audit that samples only median-resource users has not assessed whether the loop is closed.

Inference vs. Training Forkability
Systems that publish weights but withhold training data or code achieve only inference forkability: the model can be run and fine-tuned, not independently reproduced or corrected. Training forkability, third-party reproduction of the training run itself, is what satisfies FORK for learned systems.

Injunction Hook
An API in the action boundary built to accept rapid, cryptographically signed halt or rollback flags issued by recognized courts, executed without waiting for operator compliance.

Latent Forkability
The equilibrium the framework requires for coordination goods such as identity, money, and law: the credible threat of reproduction disciplines the operator without routine divergence. Where a good's value derives from uniqueness, an exercised fork is mutually assured destruction, so the threat must remain latent to remain usable.

LWD-R
Four-layer transparency requirement for learned systems: **L**ogic (architecture and inference code), **W**eights (trained parameters), **D**ata (training corpus with provenance), and **R**epresentation (the operative categorical schema of the deployed system).

Machine-Verifiable Legitimacy
The requirement that authority and attribution be record-borne acts rather than inferences from artifacts: machine-interpretable authority credentials, observable revocation propagation, tamper-evident execution traces, incident-triggered mitigation, and origination marking.

Nested Delegation
Composition of agent systems across delegation depth, where one system's harness contains another's. The tests compose across the nesting: authority attenuates, audit records reconstruct across levels, EXIT propagates down the chain, and the deployment's corrigibility is the minimum over its depth.

Ontological Capture
The condition in which a learned system's operative categories become non-contestable administrative facts, surrendering interpretive sovereignty regardless of weight publication. The risk it matures from is Representation Capture Risk; the mitigation is R-Layer Change Control over an operative schema treated as a commons rather than a vendor asset.

Open-Washing
The invocation of openness, interoperability, or digital sovereignty as reputational signals without satisfying reproducibility or governance conditions. As an umbrella failure it spans three categories: symbolic-openness, coerced-legitimacy, and substantive-substitution washing. As a specific tactic within the first category it names the release of peripheral SDKs while core execution logic stays proprietary.

Operative Representation
The categorical schema that emerges during inference under specific deployment conditions (quantization, pruning, MoE routing, output filtering, and the retrieval pipeline where present), as distinguished from *nominal representation* (the latent space of the trained artifact in isolation). Disclosure must describe operative R, not nominal R.

Origination Marking
The record states whether a human or an agent operated an action. Absence of a marker is never evidence of human operation; consulting a marker may only narrow authority, never enlarge it.

Provenance Chain
The sequence $P = [g_0, g_1, \ldots, g_n]$ of generators producing training data for downstream models. Transparency does not survive an opaque link in this chain.

Representation Capture Risk (RCR)
The risk that a learned system's representational categories ("eligibility," "risk," "fraud") become uncontestable administrative facts. Operates upstream of the action boundary.

Requisite Variety
Ashby's Law: a controller must possess at least as many states as the system it regulates. Applied to DPI/EPI: governance mechanisms must match the population's diversity of conditions to remain corrective.

R-Layer Change Control
The Temporal Stability Condition applied to operative representation. In high-stakes deployments, schema changes above a declared materiality threshold require pre-deployment notice to a designated review body, suspensive effect for objections from recognized affected-class representatives, and versioned schema diffs so drift is itself auditable. A schema that shifts faster than the affected community's capacity to contest it fails GOVERN-over-R.

Rule of the Ledger / Rule of the Workflow
Successive regimes in which authority is exerted through synchronized, self-executing artifacts. The Rule of the Ledger describes deterministic record-based authority; the Rule of the Workflow describes orchestration-mediated authority in agentic infrastructure.

Scale as a Governed Variable
Agentic deployment multiplies effective scale: environmental variety tracks principals times orchestration fan-out times action rate, so the threshold at which governance variety falls short is crossed far earlier. Where governance variety cannot be raised to match, the environmental variety must be bounded instead. Fan-out limits, action-rate limits, and blast-radius caps are then first-class GOVERN instruments, not performance tuning.

Strict Interpretability Firewall
Eligibility rule for the high-stakes tier: a learned system whose operative representation cannot be disclosed to the operative-R standard is ineligible for adoption as EPI in rights-affecting determinations. The failure sits at CODE; nothing in EXIT, AUDIT, GOVERN, or FORK compensates for it.

Subject Receipt
A signed, structured record emitted for each rights-affecting determination: action class, validator decision, timestamp, and an inclusion proof against the audit log. Delivered to the subject through a channel independent of the operator's logging infrastructure, independently verifiable, and machine-readable for juridical and class-action aggregation.

Switching Cost (σ)
The aggregate friction of leaving a system: data export, service downtime, credential re-establishment, network loss, and legal barriers, each normalized before aggregation. When the minimum σ over all plausible alternatives exceeds the calibrated threshold Σ^*, EXIT and FORK functionally fail regardless of formal rights. The implication is one-way: σ below threshold licenses no pass on its own.

Temporal Stability Condition
A system is corrigible only if the rate of effective correction exceeds the rate of error accumulation. Correction velocity carries both a lower bound (harm otherwise accumulates) and an upper bound (correction faster than verification capacity lets manipulated corrections bypass review).

Variety Drift / Temporal Variety Gap
The growing gap between a frozen model's variety and the evolving environmental variety it is meant to regulate. Drives corrigibility from a one-shot certification into a persistence condition.

Workflow Capture Coefficient (WCC)
A diagnostic abstraction quantifying epistemic delegation risk as a function of reproduction cost, ontological substitutability, and model portability. WCC near 1 indicates near-total capture; WCC near 0 indicates high sovereignty. Computed as one minus the harmonic mean of the per-dimension sovereignty scores, so that a single weak (low-sovereignty) dimension dominates the coefficient and cannot be hidden by strength in others.

References

Sunil Abraham. Engaging with the UPI debate further. Observer Research Foundation Expert Speak, July 2020. URL https://www.orfonline.org/expert-speak/engaging-with-the-upi-debate-further.

B. R. Ambedkar. *Annihilation of Caste*. Self-published, Bombay, 1936. Annotated critical edition: Verso, 2014.

B. R. Ambedkar. *What Congress and Gandhi Have Done to the Untouchables*. Thacker & Co., Bombay, 1945.

B. R. Ambedkar. *States and Minorities: What Are Their Rights and How to Secure Them in the Constitution of Free India*. Thacker & Co., Bombay, 1947.

Anivar A. Aravind. Epistemic capture and the action boundary: Corrigibility for learned and agentic public infrastructure. SSRN preprint, 2026. URL https://doi.org/10.2139/ssrn.6669318.

W. Ross Ashby. *An Introduction to Cybernetics*. Chapman & Hall, London, 1956.

Mitchell Baker and Ankit Gadgil. Aadhaar isn't progress. Mozilla Blog, May 2017. URL https://blog.mozilla.org/netpolicy/2017/05/26/aadhaar-isnt-progress/.

Gautam Bhatia. *The Transformative Constitution: A Radical Biography in Nine Acts*. HarperCollins India, New Delhi, 2019.

Biometric Update. Strings attached to Sri Lanka's quest for MOSIP integrator revealed, May 2025. URL https://www.biometricupdate.com/202505/strings-attached-to-sri-lankas-quest-for-mosip-integrator-revealed.

Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. MIT Press, Cambridge, MA, 1999.

Julie E. Cohen. *Between Truth and Power: The Legal Constructions of Informational Capitalism*. Oxford University Press, New York, 2019.

David Collingridge. *The Social Control of Technology*. Frances Pinter, London, 1980.

European Commission. Study about the impact of open source software and hardware on technological independence, competitiveness and innovation in the EU economy. Technical report, European Commission, Directorate-General for Communications Networks, Content and Technology, 2021.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024.

European Union. Regulation (EU) 2024/1183 amending regulation (EU) No 910/2014 (eIDAS 2.0). Official Journal of the European Union, 2024.

Jordyn Fetter, Krisstina Rao, and David Eaves. 2025 state of digital public infrastructure report: A look at measurement and prevalence as DPI transitions from experiment to scale. Technical Report IIPP Policy Report 2025/06, UCL Institute for Innovation and Public Purpose, November 2025. URL https://www.ucl.ac.uk/bartlett/publications/2025/nov/2025-state-digital-public-infrastructure-report.

Group of Twenty (G20). New Delhi leaders' declaration. G20 Summit, New Delhi, September 2023.

Jürgen Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, Cambridge, MA, 1996.

Albert O. Hirschman. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press, Cambridge, MA, 1970.

Reetika Khera. Impact of Aadhaar on welfare programmes. *Economic and Political Weekly*, 52(50):61–70, 2017.

Reetika Khera, editor. *Dissent on Aadhaar: Big Data Meets Big Brother*. Orient BlackSwan, Hyderabad, 2019.

Lawrence Lessig. *Code: Version 2.0*. Basic Books, New York, 2006.

Nancy G. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, Cambridge, MA, 2011.

Chantal Mouffe. *The Democratic Paradox*. Verso, London, 2000.

Mozilla Foundation. Response to India's strategy for National Open Digital Ecosystems. Mozilla Policy Submission, May 2020. URL https://blog.mozilla.org/netpolicy/files/2020/05/India-NODE-Consultation-Mozilla-Response-31052020.pdf.

Nordic Institute for Interoperability Solutions. X-Road: Open-source data exchange layer, 2026. URL https://x-road.global/. Technical documentation.

Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, 1990.

Parliament of India. The Aadhaar (Targeted Delivery of Financial and Other Subsidies, Benefits and Services) Act, 2016. No. 18 of 2016, Gazette of India, 2016.

Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, Cambridge, MA, 2015.

Philip Pettit. *Republicanism: A Theory of Freedom and Government*. Oxford University Press, Oxford, 1997.

Simon Phipps. Roman canaries. Cited in Global Information Society Watch, 2008, 2007. URL https://www.giswatch.org/thematic-report/2008-access-infrastructure/open-standards.

Reserve Bank of India. Certificates of authorisation issued by the Reserve Bank of India under the Payment and Settlement Systems Act, 2007, 2026. URL https://www.rbi.org.in/Scripts/PublicationsView.aspx?id=12043.

Right to Food Campaign. Starvation deaths in India: Documented cases and the role of Aadhaar, 2018. URL http://www.righttofoodcampaign.in/. Right to Food Campaign Secretariat, New Delhi.

Amartya Sen. *Development as Freedom*. Knopf, New York, 1999.

Richard M. Stallman. *Free Software, Free Society: Selected Essays of Richard M. Stallman*. GNU Press, Boston, MA, 2002.

Supreme Court of India. Justice K.S. Puttaswamy (Retd.) v. Union of India. Writ Petition (Civil) No. 494 of 2012, (2017) 10 SCC 1, 2017.

Supreme Court of India. Justice K.S. Puttaswamy (Retd.) v. Union of India. Writ Petition (Civil) No. 494 of 2012, (2019) 1 SCC 1; decided 26 September 2018, 2018.

United Nations Development Programme. Digital public infrastructure: Definitions and core concepts. Technical report, UNDP Digital Strategy Office, 2024.

Langdon Winner. Do artifacts have politics? *Daedalus*, 109(1):121–136, 1980.

World Bank. Principles on identification for sustainable development: Toward the digital age. Technical report, World Bank Group, Identification for Development (ID4D), 2021. URL https://id4d.worldbank.org/principles. Second edition.

World Bank. Global digital public infrastructure program: Results. World Bank Group, 2026a. URL https://www.worldbank.org/en/results/2026/05/06/global-digital-public-infrastructure-program. Results brief, 6 May 2026.

World Bank. Digital wallets: Trust frameworks — governing the ecosystem. Digital Wallet Policy Note Series, No. 2, July 2026b. URL https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099070126072023510.

World Bank. Digital wallets: A new paradigm — convergence of user-centric digital identity, data sharing and payments. Digital Wallet Policy Note Series, No. 1, June 2026c. URL https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099051126133542746. Together with: Digital Wallets 101 (same series, June 2026).

World Wide Web Consortium. Decentralized identifiers (DIDs) v1.0. W3C Recommendation, July 2022. URL https://www.w3.org/TR/did-core/.