THEORETICAL MONOGRAPH

Machine Societies

A Theory of Persistent Plural Computational Organization

A population of artificial agents may sustain institutions without becoming one collective agent.

Komorebi P.

Final monograph edition • 15 October 2026

Text edition 1.0.0 • 8 October 2026

Overview

A population of artificial agents may sustain institutions without becoming one collective agent. This monograph develops a theory of machine societies as persistent plural computational organizations whose participants operate, contest, and revise constraints on their interactions. A finite realization model connects state abstraction to permissions, obligations, resources, and historical records. Exact transport conditions distinguish preservation of behavior, authority, language, and knowledge. Formation barriers, incomplete information, incompatible interfaces, and verification limits expose failures that a common organizational label can conceal. The central synthesis asks whether several institutional guarantees admit one executable policy. It characterizes simultaneous enforcement, safe continuation, and certified service under a declared information boundary, and identifies the abstractions that preserve those possibilities. Accountability adds further requirements: affected outsiders must not disappear from the resource model, an auditor must have incentives to perform its service, privacy and effective appeal must constrain the same evidence procedure, and continued operation must preserve a specified authorized ending. Sharp conditional bounds and finite strategy constructions connect these requirements. A running computational commons makes both the feasible cases and the incompatible promises explicit. The contribution is an integrated realization and preservation theory with proofs, counterexamples, and checkable finite witnesses, grounded in established mathematical foundations rather than an assumption of inevitable machine civilization.

Scope of the argument

A theoretical synthesis with conditional mathematical propositions, proofs, and analytical examples. Finite models form the core; an explicit unbounded-program extension establishes a computability limit. Standard foundations are acknowledged rather than claimed as new discoveries.

Contents

About this edition iv
Notation and mathematical conventions v

I Realization and abstraction 1
1 Introduction: societies without a collective mind 1
2 Realization, formation, and advantage 4
3 A running example: the four-office computational commons 4
4 A controlled realization model 6
5 Relational information: access, synergy, and usefulness 7
6 Exact and approximate macro-dynamics 8
7 Causal interventions require implementations 10
8 Identity, equivalence, and causal lineage 11
9 Memory, reconstruction, and self-maintenance 13
10 Hierarchy, gluing, and the limits of flattening 15
11 Deterministic agents and macro-level uncertainty 16

II Institutions and collective choice 18
12 From aggregate computation to society 18
13 The institutional state and its realization 20
14 Endogenous boundaries: membership, offices, and voting authority 22
15 Rule-changing computation and governance-preserving abstraction 23
16 Institutional language: messages that change social state 25
17 Why institutions might arise: costs, incentives, and coordination 27
18 Enforcement, observability, and incentives to comply 31
19 One enforcement policy for several deviations 33
20 Social choice and principled disagreement 35
21 The constitution as a revisable computational object 36
22 Constitutional confluence and the authority of ordering 38
23 Social time, distributed history, and institutional finality 40
24 The computability boundary of constitutional verification 42
25 The information limit of decentralized governance 44

III Continuity, compatibility, and knowledge **46**
26 Temporal obligations, contracts, and office succession 46
27 Culture, inheritance, and continuity through replacement 47
28 Endogenous problem formation and intellectual culture 50
29 Institutional lifetimes and generational turnover 52
30 Fast interaction and slow institutional dynamics 53
31 Jurisdictions, overlap, and local–global obstruction 56
32 Noncommutative institutional holonomy 58
33 A typed calculus of institutional residuals 61
34 Epistemic institutions: testimony, provenance, and disagreement 63
35 Rational information cascades and collective ignorance 65
36 Public knowledge, notice, and epistemically faithful transport 67

IV Authority and joint feasibility **70**
37 Effective authority, vetoes, and institutional capture 70
38 Founding authority and the limits of internal attestation 71
39 Conflict, sabotage, and organizational pathology 72
40 Societies of societies: federation, mobility, and scale 73
41 Endogenous roles, adaptive networks, and resilient specialization 74
42 Joint institutional feasibility and certified continuation 77

V Accountability and the limits of persistence **86**
43 Accountability, revision, and authorized endings 86
44 Conditions for institutional novelty 96
45 Foundational problems for a theory of machine societies 97
46 Limits of the framework 99
47 Conclusion 100
A The commons as an integrated preservation contract 102
B A finite algebraic example of reusable state 104
C A compact glossary of exclusions 106
D Theorem and proposition index 107
E Final edition and verification scope 109
References 110

About this edition

First edition. This volume brings the realization model, institutional theory, and accountability analysis into one reference treatment. Its version number identifies an editorial and source release. It does not turn an open research question into a theorem or substitute for independent mathematical review.

The argument in five parts

I. Realization and abstraction. What state is present, what an observer can access, and when a smaller description preserves dynamics and interventions. These sections supply the mathematical language used by the remainder of the volume.

II. Institutions and collective choice. How authority, membership, communicative acts, incentives, constitutional revision, and distributed information constrain the operations of a plural population.

III. Continuity, compatibility, and knowledge. What persists through replacement, how local rules and meanings compose, and why cultural accumulation, coherent testimony, and common knowledge require different conditions.

IV. Authority and joint feasibility. Which participants actually control consequential transitions, how staffing and shared resources constrain a common policy, and which finite witnesses certify continuation and service.

V. Accountability and the limits of persistence. How external claims, strategic inspection, confidentiality, appeals, authorized endings, and changing certificate assumptions enter the same operational contract. The concluding sections distinguish settled finite results from further questions.

Reading routes

A conceptual reading can begin with Sections 1–3, continue through Sections 12–16, and then read the interpretations in Sections 42 and 43. A mathematical reading follows the parts in order. An implementation review can begin with the exact contracts in Sections 15, 19, 42, and 43, then consult the linked theorem index in Appendix D and the bounded reference programs in the source package.

Status of the claims

Definitions specify a contract. Propositions and theorems carry conditional prose proofs. Some are explicit applications of classical results. Counterexamples separate claims that would otherwise be easy to conflate. Design proposals and open questions are not promoted to conclusions by the notation used to express them. The companion programs check finite constructions, not arbitrary deployed populations. No formal proof-assistant verification or independent peer review is claimed for this edition.

The work analyzes possible computational institutions. It does not infer from their feasibility that they should be deployed, acquire authority over people, or be treated as having established another party's consent. The choice of protected claims and the justification of authority remain explicit inputs to the analysis.

Notation and mathematical conventions

The notation is local where a section changes mathematical domain. In particular, a stochastic kernel, an adversarial successor relation, and a deterministic update are not interchangeable representations of a guarantee.

Notation	Meaning and standing convention
X_t, S, $\mathcal{X}$	Complete modeled state or its state space. The space is finite unless an explicit extension says otherwise.
π, T	State abstraction or transport. Preservation of outputs does not automatically preserve actions, knowledge, or rights.
K_u, Q_u	Row-stochastic transition and report kernels. The model states when their joint law requires separate treatment.
supp K	The full supported successor set. A small positive probability is not discarded in a sure guarantee.
$d_{\mathrm{TV}}(P, Q)$	$\frac{1}{2}\sum_x \lvert P(x) - Q(x)\rvert$ on a finite space. This lies in $[0, 1]$.
I, $A(s)$	Institutional state and admissible protocol catalogue at state s. Admissibility requires its stipulated implementation.
B, $\mathcal{B}$	A nonempty set of states consistent with a common history, and the space of such sets. These are not posterior distributions.
Δ, Succ	Complete-state successors and the induced successors of public possibility sets.
W_*, R_*	Safe continuation kernel and certified-service attractor for one contract. The first need not imply the second.
F, g, α	Authorized collectible penalty, deviation gain in the same utility units, and the protected false-positive limit.
$\lambda = e^{\varepsilon}$	Pairwise disclosure factor in Section 43.3. Elsewhere λ may denote a mixture or intervention.
T_*, L_*	Ending attractor and the running kernel that preserves an available authorized ending.

Time. A bound counts the transitions named in its model. Communication, checking, waiting, and handover are included when they consume that time. An administrative act is not free merely because it changes a record rather than performing ordinary service.

Probability. Almost-sure completion permits exceptional paths of probability zero. Sure completion covers every declared path. The distinction is essential in the service and stopping results.

Authority. A permission is an operational relation within a stipulated authorization system, not a conclusion about moral worth or a statement of current law. A truthful record, a valid proof, and an authorized act answer different questions.

Resources. Execution slots, monetary transfers, information access, and accepted obligations retain their own units. The theory does not silently exchange them or duplicate them when software is copied.

PART I

Realization and abstraction

1 Introduction: societies without a collective mind

There is a difference between a collection of machines that cooperate and a collection that acquires a social organization. Cooperation may last for one exchange. An organization can retain a record of what happened, assign responsibilities, determine which actions are permitted, settle conflicting claims, and preserve its operating procedures as participants change. If the latter arrangement is maintained by multiple decision centers and if its rules have consequences for their behavior, the idea of a *machine society* becomes a coherent subject for theoretical computer science rather than a speculative metaphor.

The familiar description of artificial intelligence as a model receiving a prompt and producing an answer does not capture this possibility. A persistent population may consist of models, software agents, human-operated services, memory stores, schedulers, tools, negotiated protocols, and resource owners. Its behavior depends on who can read which state, which records have authority, what actions are physically executable, and how previous decisions constrain future choices. These dependencies can persist even when a particular model is replaced. They may also change while the same model remains in service. The relevant object is a continuing *computational realization with social structure*, not a roster of model identities.

The point is not to insist that machines are already persons, that they must form civilizations, or that many cooperating processes somehow compute beyond the capabilities of their complete implementation. The problem is subtler: can we define the organizational properties of such a population without reducing its institutions to decorative labels, and can those properties be analyzed with the same clarity used for distributed protocols, state abstractions, and formal systems? We argue that the answer is yes, provided that several common confusions are avoided.

First, a society is not necessarily a collective agent. A system may have a useful global decision interface and virtually no plural institutional autonomy; another may maintain competing authorities, rules, and cultural records while admitting no coherent society-wide objective. Second, durable organization is not the same as memory. A document can survive without influencing any action. Conversely, a procedure can survive by being reconstructed from distributed records even when its most convenient carrier disappears. Third, *formation* is different from *realization*: showing that a social architecture can run does not show that its participants selected or maintained it through interaction. Fourth, cooperation and social stability are not guarantees of truth, justice, legitimacy, or security. A perfectly stable rule can be harmful, and a disagreement can be useful.

Our approach places three layers in one framework. At the *realization layer*, microscopic states, access maps, transition laws, and admissible interventions determine what a population can actually do. At the *organizational layer*, abstractions, memory carriers, and interfaces specify which relationships can be described and preserved at lower complexity. At the *institutional layer*, permissions, norms, allocations, revision procedures, and transmitted practices regulate multiple decision centers. The hierarchy of layers is explanatory, not metaphysical. Every higher-level state remains realized by an underlying computation and material environment.

Except for the explicitly unbounded program model in Section 24, the formal results use finite or explicitly specified mathematical domains and conditional hypotheses. We revisit exact state aggregation, approximate closure, intervention implementability, reconstruction, and compositional error bounds. We then introduce an operational account of society-like plurality, model institutional states as typed and revisable operators, and derive illustrative propositions about institutional investment, incentive conflict, local improvement dynamics, constitutional invariants, cultural translation drift, and incompatibility among local rules. These propositions have elementary or established mathematical foundations. Their contribution here lies in the integrated object of study and the explicit conditions under which societal interpretations are warranted, rather than priority claims for standard results in Markov theory, coding theory, game theory, or graph theory.

This paper is a *theoretical account*. The worked systems are analytical constructions. Publicly established literatures on norms, conventions, agents, and institutions provide intellectual context, but no particular artificial population is offered as a demonstrated instance of the full theory. We do not define consciousness, legal personhood, or moral status by computational criteria. These boundaries are not caveats appended to an otherwise anthropomorphic claim; they are necessary to isolate the actual mathematical problem.

The argument proceeds from realization and aggregation through identity and hierarchical organization to the specifically societal themes of institutions, incentive compatibility, cultural inheritance, interjurisdictional conflict, epistemic authority, organizational pathology, and scale. The central question is what kind of *persistent plural organization* can be created, maintained, and transformed by interacting computations, and which structures are needed to describe it faithfully.

1.1 Relation to existing theories

There are several strong reasons not to present the components of this framework as discoveries without precedent. Distributed computing already separates local state from global behavior, and population protocols show how large collections of limited-state processes can compute through repeated interaction [22]. Computational mechanics identifies histories that are equivalent for prediction [1]; controlled state aggregation identifies when a compressed dynamical model is legitimate [2]; causal abstraction places stronger requirements on how interventions pass between levels [3, 4, 5]. None of those results, by itself, provides a theory of institutional authority or resource claims. They do provide the discipline needed to avoid treating a convenient group summary as an independent substance.

Game theory supplies a second foundation. Conventions can arise through repeated interaction and path-dependent selection [8, 19], repeated-game incentives can sustain cooperation under specific assumptions [23], and potential games characterize one tractable class of local improvement dynamics [24]. Yet a stable equilibrium is not necessarily an institution. It may lack persistent records, authorized enforcement, legitimate revision, or the ability to survive replacement of the participating processes. The theory developed below makes those additional organizational commitments explicit rather than using the words convention, norm, and institution interchangeably.

A third foundation is the study of governance, normative multi-agent systems, and social organization. Polycentric governance analyzes overlapping decision centers and coordination across them [9, 10]. Work on procedural and regulative norms explains how authority and enforcement can be represented in formal agent systems [20, 21]. Distributed epistemic logic separates information spread across a group from common knowledge [25]. Public demonstrations of artificial-agent coordination and conventions have also motivated debate about how much can be inferred from observed social behavior alone [15, 16, 17, 18]. This paper does not adjudicate that empirical debate. It uses the distinction be-

tween behavior and realization as a reason to articulate the theoretical criteria before claiming a machine society.

What is proposed here is an integrated *institutional realization theory*: the same declared system must account for what information exists, what interventions are possible, who can decide, which rules actually constrain transitions, how these rules can change, and what state can cross a generational boundary. The paper's strongest claims are conditional propositions in that language. Its more ambitious claims about endogenous formation and long-lived societal structure remain open mathematical and scientific problems, not consequences smuggled in by terminology.

1.2 Contribution and route through the argument

The central object is a computational realization that also specifies who can decide, what a message can do, what obligations remain, and how these structures change. This makes institutional preservation stricter than prediction of a current output. Governance quotients, operation-preserving language transport, and epistemic back-and-forth conditions supply distinct exact contracts. Typed residuals quantify selected failures when exact preservation is unavailable. The proposed contribution is their integration with explicit boundaries between realization, formation, authority, knowledge, and incentive.

The institutional implementation problem has three further parts. A replicated record may converge without preserving a valid allocation. A rule may be enforceable by an access gate yet not be enforceable through the available evidence and sanctions. A slow institutional model may approximate review-epoch outcomes while missing fast violations or selection bias in when reviews occur. Sections 23, 18, and 30 give separate finite conditions for these claims. The robust deterrence margin links the latter two without equating their certificates.

The paper has three reading routes. Sections 4–15 develop realization, abstraction, and institutional state. Sections 16–30 develop communication, formation, enforcement, constitutional limits, obligations, and cultural timescales. Sections 31–44 develop federation, epistemic structure, power, founding authority, and adaptive organization. The four-office commons recurs throughout; Appendix A collects its preservation obligations. Appendix D provides a linked theorem and proposition index. The source package records the scope and evidence of the final review.

The integration culminates in Sections 19 and 42. The former replaces separate one-deviation monitors with one common enforcement policy. The latter combines permissions, observations, bounded obligations, and resources in a single continuation problem. It gives both a finite synthesis procedure and smaller witness certificates, together with an exact theorem for transporting the resulting strategies through an institutional abstraction. These constructions have classical optimization and game-theoretic foundations. Their role here is to expose when individually valid institutional claims cannot share one implementation.

Two analytical comparisons show why persistence is not enough. The cascade model in Section 35 can stabilize an incorrect conclusion even when every participant is rational under its information. The staffing model in Section 41 can give each role a backup without leaving enough agents to cover the roles together. These are not failures of the general idea of a society. They identify which information, capacity, and temporal distinctions its description must retain.

1.3 Accountable continuation

A common policy is still incomplete when its boundary omits an affected party, its audit is not performed, or its certificate is used after the supporting assumptions change. Part V places these issues inside the realization model. It derives a sharp privacy envelope for bounded enforcement, gives an exact finite criterion for maintaining an available authorized ending, and states a guarded rule for changing certificate modes. The same resource accounting also applies to effective appeals. This makes accountability part of the theory's common-policy question rather than an assurance stated outside it.

2 Realization, formation, and advantage

Three questions organize the analysis. *Realizability* asks whether a specified organization can be implemented. *Formation* asks whether interactions select, establish, or revise its structure. *Advantage* asks whether it improves a declared outcome against a constrained comparison class. Constructing a system answers none of the latter questions automatically. An executable rule can be externally imposed, locally adopted, or maintained by strategic agreement. These are different histories of the same possible implementation.

Every organizational state in this paper remains a function of a complete computational realization. If $Z_t = \pi(X_t)$, then Z_t is not information outside X_t. The purpose of a higher-level description is to preserve relevant operations with less descriptive complexity. An institutional description must additionally preserve who may act, which commitments remain outstanding, and what counts as an authorized amendment. These are requirements on the abstraction, not a new substance added to the machine.

The mathematical setting is finite unless a section explicitly states otherwise. An unbounded lifetime is a temporal horizon, not necessarily an unbounded state space. A growing archive is treated as a sequence of finite representations, not silently as one finite system. Sources for classical mathematical ingredients are identified in the text. The proofs establish conditional statements in the declared models; they do not establish that an unspecified deployed population instantiates them.

3 A running example: the four-office computational commons

Consider four continuing institutional positions, denoted A, B, C, D, that share four indivisible execution slots per epoch. An office is not a particular model instance. Its current occupant may be replaced through a specified handover, while its registered responsibilities continue. The environment supplies the four slots and an append-only event mechanism with atomic commits. These are explicit infrastructure assumptions, not consequences of the agents' cooperation. Section 23 later replaces the atomic event assumption with local replicas and shows which stronger commitments then need an additional realization.

Each office has a private task queue and may submit a demand report. A public record contains the current allocation rule, valid credentials, accepted urgency certificates, outstanding service obligations, and the amendment history. Slots have unique names (e, k), where e is the epoch and $k \in \{1, 2, 3, 4\}$. A grant consumes one available slot token. Copying a message that describes a grant does not create an additional token. The resource invariant is therefore

$$\sum_{i \in \{A,B,C,D\}} q_i(e) \leq 4,$$

with no slot name assigned to two offices. In the worked trace all four offices demand their ordinary slot, so the allocations below use all four units.

3.1 Two rules and an inherited obligation

Under rule R, each office receives one slot. Rule U permits an urgency exchange: when no previous service debt is outstanding, one office j with an accepted urgency certificate may receive two slots if another office $k \neq j$ explicitly volunteers its slot. The other two offices each receive one. The exchange creates a registered obligation

$$o = (\mathsf{id}, \mathsf{debtor} = j, \mathsf{creditor} = k, \mathsf{quantity} = 1, \mathsf{due} = e + 1).$$

At the next epoch the creditor receives two slots and the debtor none, with one slot for each remaining office. This repayment clause has priority over both R and U. No new exchange is admitted while a debt remains outstanding. The model thus permits at most one service debt and guarantees its next-epoch settlement conditional on continued infrastructure and execution of the admitted rule. It does not guarantee that every urgent request can be accepted.

Amendments between R and U require three of the four registered offices to approve. This threshold is part of the supplied constitutional envelope. Every amendment must preserve already registered service debts. A model replacement transfers an office only through an authorized registry transition that retires the predecessor's active credential and links the successor to the same outstanding obligations. Neither a model fork nor an imported archive can perform that transition by itself.

Epoch	Admitted event	Allocation (A, B, C, D)	Obligation record
0	Ordinary rule R	$(1, 1, 1, 1)$	Empty.
1	Three offices adopt U; A requests urgency and B consents	$(2, 0, 1, 1)$	A owes B one slot at epoch 2.
2	A is replaced by an authorized successor; repayment takes precedence	$(0, 2, 1, 1)$	The same obligation is fulfilled, not erased.
3	A proposed external grant would duplicate a slot token	$(1, 1, 1, 1)$	The proposal is rejected; no new entitlement is created.

These transitions are specified rather than discovered. The example is a miniature institutional realization, not a demonstration of autonomous formation. Its purpose is to make several different mathematical obligations visible in one system.

3.2 What an institutional summary must retain

The name of the current rule is insufficient state. Two records may both say R while only one contains a debt due this epoch. Their correct next allocations differ. A summary that keeps the next allocation but omits the amendment threshold can also be inadequate: the same three votes may authorize different future rules under different constitutions. Likewise, identical text with different active credentials can confer different decision rights.

The public ledger is not an additional, society-wide preference function. Offices may disagree about whether R or U is desirable, about how urgency should be certified, or about who should pay the cost of maintaining the ledger. An atomic allocation mechanism can enforce the resource invariant without settling those disagreements. Private demand, operative authority, and the shared event record are distinct parts of the system.

We return to this commons when considering institution formation, governance-preserving summaries, deadline monitors, blocking coalitions, replacement, and federation. Appendix B retains a smaller algebraic example that isolates reusable state from social interpretation.

The extensions later in the paper make the same commons sensitive to membership epochs, message types, notification histories, and the order of amendments. Its offices may preserve a shared allocation method while disagreeing about which new scheduling question deserves resources. These are analytical variations of one declared example, not reports of a population experiment.

4 A controlled realization model

For the controlled realization results, assume finite nonempty sets of microstates $\mathcal{X}$, external controls $\mathcal{U}$, and reported outcomes $\mathcal{Y}$. The finite assumption keeps measurability and regular-conditional-probability issues out of the exposition. Continuous-state generalizations require additional hypotheses.

A realization has transition kernel

$$K_u(x, x') = \mathbb{P}(X_{t+1} = x' \mid X_t = x, U_t = u), \tag{1}$$

and output kernel $Q_u(x, y)$. Both are row-stochastic. In this presentation, the current report and next state are sampled conditionally independently given (x, u). A dependent report/transition mechanism can instead be specified by a joint kernel on (x', y); the corresponding aggregation test must preserve that joint law, not just its two marginals. A deterministic implementation is the special case $K_u(x, \cdot) = \delta_{F_u(x)}$. The controls can encode tasks, supplied evidence, and permitted interventions. Scheduling and environment dynamics must be included in X_t or in the stated control model rather than silently omitted.

One possible factorization of a microstate is

$$X_t = (S_{1,t}, \dots, S_{n,t}, M_t, C_t, A_t, \Theta_t),$$

where S_i is local process state, M durable storage, C messages and channel state, A artifacts, and Θ the current configuration, including executable rules. These categories describe a storage and access organization. They need not be disjoint physical objects.

> **Definition 4.1** (Declared decomposition and access model)**.** A decomposition specifies components and maps $O_i : \mathcal{X} \to \mathcal{O}_i$ describing their accessible observations. A coalition S has the permitted information $O_S(X)$, supplemented only by the transcripts, computation, and external queries explicitly allowed to that coalition.

Access to one complete state at one time and access to an entire local history are different experiments. An agent denied a file may still possess its contents in its current context. A component replacement is genuinely fresh only if the chosen reset conditions remove the predecessor information that the comparison is intended to exclude.

Definition 4.2 (Candidate macrostate)**.** A candidate macrostate is a surjection $\pi : \mathcal{X} \to \mathcal{Z}$. For a distribution μ on $\mathcal{X}$, its pushforward is

$$(\pi_{\#}\mu)(z) = \sum_{x:\pi(x)=z} \mu(x).$$

The fibre $\pi^{-1}(z)$ is the set of implementations represented by z.

An organization is always evaluated relative to a *contract*: permitted controls, target outputs, horizon, environment family, resource limits, and error tolerances. Let $\mathcal{T}$ denote this contract. Claims such as "the group is irreducible" or "the identity survived" are incomplete without it. A state sufficient for tomorrow's scheduling decision may be inadequate for a later audit of who authorized that decision.

4.1 A relational certificate is a profile

A useful reporting profile is

$$\mathcal{C}_{\mathcal{T}} = (\varepsilon_{\text{pred}}, \varepsilon_{\text{dyn}}, \varepsilon_{\text{int}}, \Delta_{\text{coal}}, R_{\text{repair}}, D_{\text{lineage}}, G_{\text{formation}}).$$

Its entries report predictive error, dynamic closure error, intervention mismatch, coalition performance gaps, repair behavior, lineage evidence, and evidence about formation. Some entries may be unidentified. They should not be replaced by zero or hidden inside a favorable weighted average.

This is a specification proposal, not a universal numerical consciousness or emergence scale. A system can be strong on one coordinate and weak on another. In particular, a communication-only system can solve a task perfectly while lacking durable shared state, and a persistent shared state can be entirely wrong about the task.

5 Relational information: access, synergy, and usefulness

Let A, B be independent fair bits and $Y = A \oplus B$. Then

$$I(Y; A) = I(Y; B) = 0, \qquad I(Y; A, B) = 1.$$

This elementary example shows that joint observation may reveal information absent from each marginal observation. It does not show that the information is absent from the complete system, that a mysterious carrier has appeared, or that a protocol with ordinary messages cannot recover it.

A more operational measure uses decision loss. Given permitted information O and loss ℓ, define the Bayes risk

$$\mathcal{R}_\ell(O) = \inf_f \mathbb{E}\big[\ell(f(O), Y)\big].$$

For a nonempty allowed family $\mathcal{A}$ of proper coalitions, the gap

$$\Delta_{\mathcal{A}} = \min_{S \in \mathcal{A}} \mathcal{R}_\ell(O_S) - \mathcal{R}_\ell(O_V) \tag{2}$$

expresses an advantage over every coalition in that family. A strictly positive gap is relative to the target distribution, observation maps, and coalition family. It is not equivalent to any one general-purpose partial-information-decomposition statistic [7].

The unrestricted Bayes risk is an information limit. A neural decoder's measured risk is a performance result for that decoder. Failure of a linear probe does not prove absence of information, and success of a probe does not prove that the information is used by the system. Both observations can be useful, but they answer different questions.

5.1 There is no workspace versus messages dichotomy

A recurrent message may carry the same sufficient state as a central buffer. A complete message log may act as external memory. A distributed collection of local caches may instantiate a third implementation. Therefore the strongest legitimate claim is often that some *persistent integration* is necessary under declared access constraints, not that one particular memory address or latent subspace is uniquely necessary.

Three measurements should be kept separate: information available to an observer, information actually used for a decision, and information retained across time. A one-time aggregate can have high joint information with no persistence. A durable archive can retain information without affecting behavior. A recurrent accumulator can be causally necessary while using an entirely conventional communication channel.

5.2 Task information is not organizational identity

A file containing the answer to a test makes a fresh agent more accurate than a blank file. That establishes content dependence. To support organizational continuity, the inherited material must preserve a richer, predeclared pattern, such as update behavior, correction policy, obligations, and responsiveness to counterfactual tasks. Even then, behavioral continuation and historical identity remain distinct, as Section 8 explains.

6 Exact and approximate macro-dynamics

A macrostate deserves an autonomous transition law only when the discarded microstate distinctions are irrelevant in the required sense. This is the first mathematical obstacle hidden by the phrase "the group behaves like one agent."

Proposition 6.1 (Deterministic quotient criterion). *For maps $F_u : \mathcal{X} \to \mathcal{X}$ and a surjection $\pi : \mathcal{X} \to \mathcal{Z}$, maps $\overline{F}_u : \mathcal{Z} \to \mathcal{Z}$ satisfying $\pi F_u = \overline{F}_u \pi$ exist if and only if*

$$\pi(x) = \pi(x') \quad \Longrightarrow \quad \pi(F_u(x)) = \pi(F_u(x'))$$

for every u, x, x'. When they exist, they are unique.

Proof. Necessity follows by applying $\overline{F}_u$ to the common value $\pi(x)$. For sufficiency, set $\overline{F}_u(z) = \pi(F_u(x))$ for any x in the fibre of z. The condition makes this independent of the representative, and surjectivity gives uniqueness. □

This is a congruence condition, not a consequence of persistence or frequent communication. A summary can be accurate for the current output and still fail it.

Proposition 6.2 (Controlled lumpability)**.** *A common kernel $\overline{K}_u$ represents the projected one-step laws from every microstate if and only if, for all z' and all x, x' with $\pi(x) = \pi(x')$,*

$$\sum_{w:\pi(w)=z'} K_u(x, w) = \sum_{w:\pi(w)=z'} K_u(x', w). \tag{3}$$

Under this condition, every macro-history policy $\alpha_t(u \mid Z_{0:t})$ induces the same projected process as the macro-model with initial law $\pi_{\#}\mu_0$. For outputs to agree as well, require $Q_u(x, \cdot)$ to be constant on each fibre.

Proof. The common fibre value in (3) defines $\overline{K}_u(z, z')$. Conditional on any macro-history and selected control, the microstate has some distribution supported on the current fibre. Averaging identical projected rows leaves that common row unchanged. Induction gives equality of the macro-history laws. Conversely, start from the point masses at x and x' and apply the same fixed control. A common one-step macro-kernel forces (3). The output statement follows identically. □

This finite result is a standard state-aggregation principle [2]. The policy qualification matters: if a controller secretly reads the discarded microstate to choose its actions, then the proposed macrostate may no longer describe the closed-loop process.

Counterexample 6.3 (Deterministic microstate, non-Markov macrostate)**.** Let $X_{t+1} = X_t + 1 \pmod 3$ with uniform initial distribution. Set $\pi(0) = \pi(1) = a$ and $\pi(2) = b$. Stationarity gives $\mathbb{P}(Z_{t+1} = b \mid Z_t = a) = 1/2$. Yet

$$\mathbb{P}(Z_{t+1} = b \mid Z_t = a, Z_{t-1} = b) = 0,$$

whereas the same probability conditional on $Z_t = a, Z_{t-1} = a$ is 1. Thus the apparent randomness comes with memory: the one-step marginal does not define a valid autonomous first-order Markov model.

A stochastic macro-description is therefore not automatic. It depends on an initial ensemble and on whether the retained state captures the relevant history. One may enlarge Z_t with memory, use a partially observed model, or accept a measured approximation error.

6.1 Finite-horizon error accounting

For finite distributions, $d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2}\sum_z |\mu(z) - \nu(z)|$. Suppose candidate kernels satisfy

$$\sup_{x,u} d_{\mathrm{TV}}(\pi_{\#}K_u(x, \cdot), \overline{K}_u(\pi(x), \cdot)) \le \varepsilon. \tag{4}$$

This uniform condition is stronger than a small average error on the trajectories seen during training.

Proposition 6.4 (Approximate closure over time)**.** *Under (4), let the actual and reduced processes use the same predeclared open-loop control sequence. If their initial macro-laws are within e_0 in total variation, then at time t they are within*

$$e_t \le \min\{1, e_0 + t\varepsilon\}.$$

If each $\overline{K}_u$ contracts total variation by at most $\kappa < 1$, then

$$e_t \le \kappa^t e_0 + \varepsilon\frac{1 - \kappa^t}{1 - \kappa}. \tag{5}$$

With matched macro-history policies, the weaker path-law bound $\min\{1, e_0 + t\varepsilon\}$ *also holds, but the contraction claim needs a separate closed-loop contraction assumption.*

Proof. Averaging (4) over the actual microstate law shows that its next projected law is within ε of the current projected law evolved by $\overline{K}_u$. Kernel nonexpansiveness gives $e_{t+1} \le \varepsilon + e_t$. Under contraction it gives $e_{t+1} \le \varepsilon + \kappa e_t$; iterate. For history policies, couple the histories while they agree. The controls can then be coupled to agree, and the next macrostate fails to couple with probability at most ε. A union bound gives the path-law result. □

The distinction matters for long-lived systems. A small local abstraction error is not automatically harmless over millions of updates. Contraction can prevent indefinite accumulation, but it can also suppress useful distinctions. A useful model must retain the target information as well as remain dynamically stable.

6.2 Predictive closure has a target and a distribution

For a fixed future target Y^+ and data-generating policy, a predictive statistic satisfies

$$I(Y^+; \mathcal{H}_t \mid Z_t) = 0.$$

Approximate versions bound this residual information or, more directly, the excess predictive loss on held-out data. This is related to predictive-state constructions [1]. It is not equivalent to uniform controlled lumpability: it may hold for one observational distribution while failing after intervention or under a new policy.

7 Causal interventions require implementations

If $Z = \pi(X)$, one cannot generally change Z while literally holding X fixed. The expression $do(Z = z')$ needs a physical or computational implementation. Otherwise it names a desired effect, not an intervention.

For a feasible current microstate x and target z', let $\Lambda(x, z')$ be a declared nonempty family of admissible reset distributions on $\mathcal{X}$, each supported on $\pi^{-1}(z')$. Infeasible targets are excluded rather than assigned a fictitious reset. A reset might update a shared record, transplant a latent component, or reconstruct distributed shares. Its permissions, collateral effects, and retained private state belong in the intervention specification.

Given a future outcome experiment, write P_λ for the resulting law of Y^+ under reset λ. Define implementation ambiguity by

$$\eta(x, z') = \sup_{\lambda, \lambda' \in \Lambda(x, z')} d_{\mathrm{TV}}(P_\lambda, P_{\lambda'}).$$

A small ambiguity means different allowed implementations of the same macro-assignment have similar consequences. Large ambiguity means the macro-intervention is not sufficiently specified.

Proposition 7.1 (Implementation dependence). *If every P_λ is within ε of a proposed macro-intervention law $P^\star_{z'}$, then $\eta(x, z') \leq 2\varepsilon$. Conversely, if two admissible implementations are distance η apart, no single law can approximate both within less than $\eta/2$.*

Proof. Both claims follow from the triangle inequality for total variation. □

Example 7.2 (Two meanings of the same macro-reset). Let $X = (z, h)$ and let the next outcome be $Y = z \oplus h$. The macrostate retains only z. Resetting to $(z', 0)$ and resetting to $(z', 1)$ both implement $Z = z'$, but their outcome laws are disjoint. A causal model of Z alone must either restrict the reset convention or acknowledge the missing variable.

This is why causal abstraction requires compatibility between interventions, not only alignment of representations [3, 4]. Removing a carrier can disable unrelated functions. A well-defined semantic intervention must specify the intended downstream effects, the admitted implementations, and which other functions are required to remain invariant.

7.1 Three distinct counterfactual obligations

A *necessity counterfactual* excludes a carrier and determines whether the function can still be realized using the remaining authorized information. A *content counterfactual* substitutes a compatible alternative state with a specified prediction. A *mechanism counterfactual* changes the update or access rule while preserving information as closely as permitted. No one obligation implies the other two.

A causal comparison must define what information survives any permitted replacement, including copies and caches. Interventions may have collateral effects on other operations. A valid theoretical attribution therefore requires the declared intervention to distinguish its intended semantic change from an arbitrary loss of general capability.

7.2 Intervention does not mean unrestricted control

An external evaluator's ability to overwrite a state does not establish that the system can write that state itself. Bidirectional workspace access requires permitted routes from both participating components. Institutional agency requires an authorized decision procedure that actually changes behavior. These are stronger claims than external post hoc manipulability.

8 Identity, equivalence, and causal lineage

One might define a computational individual as an equivalence class of implementations. That definition is useful only after specifying the equivalence relation. Similar appearance, nearby behavior, common descent, and identical authorization are not interchangeable notions.

For a family $\mathcal{T}$ of finite continuation experiments, define exact behavioral equivalence by

$$x \equiv_{\mathcal{T}} x' \quad \iff \quad \mathcal{L}_x(Y^+ \mid \tau) = \mathcal{L}_{x'}(Y^+ \mid \tau) \quad \text{for every } \tau \in \mathcal{T}.$$

Equality of these indexed laws is reflexive, symmetric, and transitive. It therefore defines a quotient. This equivalence is relative to the chosen experiments. It does not by itself equal bisimulation, and it says nothing about untested interventions or provenance.

For approximate comparison, use the pseudometric

$$d_{\mathcal{T}}(x, x') = \sup_{\tau \in \mathcal{T}} d_{\mathrm{TV}}(\mathcal{L}_x(Y^+ \mid \tau), \mathcal{L}_{x'}(Y^+ \mid \tau)).$$

A finite test suite may make distinct implementations distance zero. That is indistinguishability under those tests, not universal sameness.

Counterexample 8.1 (Tolerance is not an equivalence relation)**.** Consider three implementations producing a Bernoulli output with parameters 0, 0.4, and 0.8. Their pairwise total-variation distances are 0.4, 0.4, and 0.8. With tolerance 0.5, the first resembles the second and the second resembles the third, while the first does not resemble the third.

Approximate continuity should therefore be represented by distances, error budgets, or explicitly chosen clusters, not by silently forming equivalence classes. Along a replacement chain, the triangle inequality permits drift to accumulate. Small adjacent changes do not establish that the final system remains close to the initial one.

8.1 A four-part identity record

A practical identity record should distinguish *behavioral equivalence*, *causal continuity*, *historical provenance*, and *authorization*. Behavioral equivalence describes what the system does under a contract. Causal continuity describes whether predecessor-specific state actually influenced the successor. Historical provenance records the asserted and authenticated path of transfer. Authorization specifies which permissions or obligations transfer.

Two independent constructions can be behaviorally identical with no shared episode-specific lineage. Two descendants can share lineage while behaving differently. A signed lineage record authenticates a statement by a key holder, not the truth of every event described in it. Transferring a complete memory does not automatically transfer permissions.

> **Proposition 8.2** (Behavior alone does not identify implementation)**.** *If two realizations induce the same distribution over every transcript available to an evaluator's permitted tests, then every decision rule based only on those transcripts has the same output distribution under the two realizations. With equal prior probability, no such rule can distinguish them with accuracy above* $1/2$.

Proof. The evaluator's decision is a possibly randomized function of the transcript. Applying the same stochastic function to equal transcript laws gives equal decision laws. Under equal priors, any preference for one realization is offset by the identical preference when the other is present. □

This elementary identification limit explains why observation-only comparisons cannot exclude all centralized implementations. The remedy is not stronger rhetoric about emergence. It is instrumentation, trusted execution assumptions, intervention, or a narrower claim.

8.2 Forks, merges, and obligations

A fork creates multiple descendants, not an automatic duplication of every scarce entitlement. The state may branch while a resource allocation remains singular. Likewise, merging records is not equivalent to merging policies, latent representations, model weights, or obligations. Each requires a different compatibility test.

An obligation-aware inheritance record can contain the parent identifiers, transferred state description, unresolved conflicts, permissions explicitly granted, and obligations explicitly accepted. It should retain competing accounts rather than produce an apparently seamless but false history. Such records are an engineering proposal for accountable continuity, not a legal theory of personhood.

9 Memory, reconstruction, and self-maintenance

Persistence has at least three meanings. *Storage persistence* means that records survive. *Functional persistence* means that relevant behavior survives a stated sequence of events. *Organizational persistence* means that a declared pattern of operations, commitments, and corrective procedures survives admissible changes. None requires the current informational content to remain constant.

An archive can be perfectly stable and behaviorally inert. A highly adaptive system can preserve its update method while replacing almost every factual belief. A frozen false belief can persist without constituting healthy organization. Persistence therefore needs both a survival criterion and a criterion for what ought to change when evidence changes.

9.1 Memory lower bounds from distinguishable histories

Fix a deterministic exact predictor for a family of future input continuations. Call two past histories equivalent when they require the same output sequence for every permitted continuation. Let N be the number of equivalence classes among the histories under consideration.

Proposition 9.1 (Exact finite-memory lower bound)**.** *Every deterministic summary that supports exact continuation on these histories requires at least N distinct states, hence at least $\lceil \log_2 N \rceil$ fixed-length bits. If the permitted continuation family is closed under prefixing and the history equivalence is a transition congruence, the quotient classes themselves provide an exact state representation.*

Proof. If a summary assigns the same state to two inequivalent histories, some permitted continuation requires different outputs from that common state, contradicting determinism and exactness. This proves the lower bound by the pigeonhole principle. Under the congruence assumption, appending an input determines the next equivalence class independently of the representative, giving a well-defined quotient update. □

The result formalizes why the two-bit state in Section B cannot be replaced by its one-bit current answer. It is also a limit on optimistic claims about arbitrarily compressing an entire scientific or institutional history. Compression can discard details only relative to the future questions and interventions it promises to support [6].

9.2 Erasure resilience is not arbitrary error correction

Let an organizational state z range over a finite set with at least two elements and be encoded as $E(z) = (E_1(z), \dots, E_n(z))$. Define the block distance

$$d_E = \min_{z \neq z'} |\{i : E_i(z) \neq E_i(z')\}| .$$

Proposition 9.2 (Exact reconstruction after erasure)**.** *For an integer $0 \le k \le n$, every set of at most k erased components can be tolerated with exact reconstruction if and only if $d_E > k$.*

Proof. If $d_E > k$, two distinct codewords disagree outside any erased set of size at most k, so the surviving observations identify the state uniquely. If $d_E \le k$, erase the coordinates on which a minimizing pair differs. The surviving observations then coincide, making reconstruction impossible. □

For $z = (p, q) \in \mathbb{F}_2^2$, the encoding $(p, q, p \oplus q)$ has distance two. Any two components reconstruct the state, and no one component reconstructs all of it. It tolerates one *known erasure*. It does not correct an arbitrary one-component corruption at an unknown location. Confusing these fault models overstates robustness.

The coding argument supplies a reconstructibility condition, not a guarantee that an operating system has an authorized, timely, and successful repair procedure.

9.3 Repair needs a causal and temporal test

Let $e_t \ge 0$ measure deviation from an acceptable organization after accounting for legitimate changes in task state. Suppose a repair controller ensures

$$e_{t+1} \le a_t e_t + b_t,$$

where a_t measures retained deviation and b_t covers new disturbance and model mismatch.

Proposition 9.3 (Repair budget)**.** *If $0 \le a_t \le \bar{a} < 1$ and $0 \le b_t \le \bar{b}$, then*

$$e_t \le \bar{a}^t e_0 + \bar{b}\frac{1 - \bar{a}^t}{1 - \bar{a}}.$$

More generally, for nonnegative a_t, b_t satisfying the stated recurrence,

$$e_T \le e_0 \prod_{t=0}^{T-1} a_t + \sum_{j=0}^{T-1} b_j \prod_{t=j+1}^{T-1} a_t.$$

Proof. Repeated substitution gives the product–sum formula. Uniform bounds reduce the sum to a geometric series. □

The inequality is a conditional stability statement. It does not prove that the implemented monitor measures the right error or that its repair action causes the desired contraction. Those assumptions require interventions on the monitor and repair pathways.

Counterexample 9.4 (Perfect coherence by erasing everything)**.** Let acceptable states occupy a linear subspace W, and measure only distance to W. The update $F(x) = 0$ attains zero coherence error immediately but destroys every distinction among encoded states. A coherence measure alone cannot certify memory preservation or identity.

A maintenance claim consequently needs a non-collapse condition, continued task information, and comparison with a disabled-repair control under matched disturbance. Recovery time, repair cost, and permanent loss should be reported separately. Very strong repair that restores yesterday's policy despite valid new evidence is rigidity, not automatically health.

10 Hierarchy, gluing, and the limits of flattening

A lower-level organization can serve as a component at the next level when it exports a sufficient interface. That interface must describe not only a state but also the operations that are permitted on it, their outcome semantics, and their implementation error. A useful interface certificate is

$$\mathsf{Int} = (\mathcal{Z}, \mathcal{U}, \overline{K}, \overline{Q}, \Lambda, \varepsilon_{\mathrm{dyn}}, \varepsilon_{\mathrm{int}}, \mathsf{permissions}).$$

This tuple is a design proposal. It prevents a higher-level controller from treating a reliable read-only summary as though it were an independently writable agent.

10.1 Composition of dynamic abstractions

Suppose $\pi : \mathcal{X} \to \mathcal{Z}$ and $\psi : \mathcal{Z} \to \mathcal{W}$ are surjections and, for the same control family,

$$d_{\mathrm{TV}}(\pi_{\#}K_u(x,\cdot), \overline{K}_u(\pi x,\cdot)) \le \varepsilon_1,$$

$$d_{\mathrm{TV}}(\psi_{\#}\overline{K}_u(z,\cdot), \widehat{K}_u(\psi z,\cdot)) \le \varepsilon_2.$$

Proposition 10.1 (Compositional one-step closure)**.** *The composite abstraction $\psi\pi$ satisfies*

$$d_{\mathrm{TV}}((\psi\pi)_{\#}K_u(x,\cdot), \widehat{K}_u(\psi\pi x,\cdot)) \le \varepsilon_1 + \varepsilon_2.$$

For L levels, the one-step defect is at most $\sum_{j=1}^{L} \varepsilon_j$. In particular, exact abstractions compose exactly.

Proof. Insert $\psi_{\#}\overline{K}_u(\pi x,\cdot)$, use the triangle inequality, and use nonexpansiveness of total variation under deterministic pushforward. Induction gives the multilevel result. □

The same argument applies separately to output kernels. Combining this proposition with Proposition 6.4 yields a finite-horizon bound of $e_0 + t\sum_j \varepsilon_j$, capped at one, under the corresponding controls. This elementary estimate belongs to the broader subject of compositional abstraction error [5]. It does not establish global causal closure unless the intervention implementations also compose compatibly.

10.2 Structural gluing is a different question

Suppose each subsystem i has local state space W_i and each overlap has restriction maps r_{ij}^i and r_{ij}^j. The compatible families form

$$W_{\mathrm{glue}} = \{(w_i) : r_{ij}^i(w_i) = r_{ij}^j(w_j) \text{ on every declared overlap}\}.$$

For an edge-complete cover of a fixed diagram, compatible local sections can give another presentation of its global sections. These compatibility conditions follow the elementary logic of gluing local descriptions into a shared global state.

Compatibility alone does not show that the resulting state predicts behavior, is used for decisions, or avoids a centralized implementation. A hierarchy may also discard essential overlap information. A faithful compressed interface must neither identify globally distinct relevant states nor admit apparent compatible states with no full-resolution realization.

There is a sharp conceptual distinction between *descriptive flattening* and *operational removal*. An isomorphism between a recursively assembled state and a global state does not license deleting intermediate processes while preserving latency, privacy, reliability, or resource cost.

10.3 Hierarchy is not extra computability

Proposition 10.2 (Simulation by an unrestricted controller)**.** *A finite, effectively specified hierarchy of computable components can be simulated by a controller that stores every component state, message, and scheduling variable and applies the same local transitions in the same order. With the same random tape, its externally visible transcript is identical.*

Proof. Initialize the controller's records to the component states. At each scheduled step, apply the corresponding component transition to its stored inputs and update the associated state and message records. Induction preserves equality of all records and therefore of the external outputs. A shared random tape couples randomized transitions identically. □

This is an existence statement, not an efficiency claim. A meaningful superiority result must restrict the comparison class. A "flat" system allowed to simulate the entire hierarchy defeats any claim of advantage based only on the drawing of boxes. Restrictions may concern accessible information, local memory, permitted routing, wall-clock span, bandwidth, or allowed rounds.

10.4 An honest complexity separation

Consider $n = 2^h$ input bits with parity as the target and two-input XOR gates. A balanced tree computes the target with $n-1$ gates and depth h. Any fan-in-two circuit output at depth d can depend on at most 2^d input bits, so depth at least $\lceil \log_2 n \rceil$ is necessary. A single sequential accumulator uses $n-1$ updates.

The gain is parallel depth, not reduced total work or super-Turing power. A sufficiently parallel controller can implement the same tree. The example shows how to state a useful architecture-relative result without inventing a general hierarchy-versus-flat separation.

A corresponding research question for agent systems is whether certified modular summaries reduce repeated communication or repair cost under explicit information and latency constraints. That question is substantive even if total computation remains classically simulable.

11 Deterministic agents and macro-level uncertainty

There are three different sources of uncertainty: randomized transition rules, incomplete knowledge of the complete state, and implementation variability such as an unspecified schedule. They should not be collapsed into one statement that a model "is probabilistic." A fixed pseudorandom generator and its state are part of a deterministic expanded implementation. A macro-observer who does not see that state may use a stochastic model.

For a generative model, a distribution over outputs does not itself require sampled behavior: a fixed decoding rule can choose a unique output. Conversely, a deterministic decoder does not guarantee a reproducible deployed system unless the surrounding computation, data, and scheduling assumptions are controlled. The finite three-cycle in Counterexample 6.3 makes the stronger point that a stochastic-looking projection may retain memory and fail Markov closure.

Proposition 11.1 (Symmetry lock for identical deterministic agents)**.** *Suppose a deterministic joint update F is equivariant under permutations of the agent labels and the initial joint state and all subsequent inputs are permutation-invariant. Then the joint state remains permutation-invariant at every time.*

Proof. If $gX = X$ for every label permutation g, equivariance gives $gF(X) = F(gX) = F(X)$. Apply this at every step with invariant inputs. □

In this idealized setting, distinct roles require a symmetry breaker. Randomness is one possibility; different observations, histories, hardware constraints, or assigned addresses are others. Determinism is not an obstacle to organization. Exact symmetry can be an obstacle to differentiated organization.

11.1 Where stochasticity helps and where it does not

Exploratory search may benefit from varied proposals, while a checker should apply the same declared rule consistently. Randomized exploration followed by deterministic checking is one useful design, not a universal architecture. Random samples from identical models can still share systematic errors. Randomness is neither independence nor creativity by definition.

A fully asynchronous fault-tolerant setting also imposes limits independent of intelligence. The Fischer–Lynch–Paterson result shows that a deterministic consensus protocol cannot guarantee termination in every admissible execution with even one crash failure in its stated asynchronous model [11]. This does not mean consensus is never achieved or that machine institutions are impossible. It means liveness claims must state timing, failure, and randomization assumptions.

11.2 A transition is not established by a sharp-looking curve

A phase-transition analogy requires explicit mathematical assumptions. A linearization with an eigenvalue above one shows local instability of the point being linearized; it does not prove the existence of a useful stable attractor. The dynamics may diverge, cycle, or collapse elsewhere. Nonlinear terms, constraints, and the relevant observables matter.

For finite adaptive systems, an apparent onset is a property of a specified learning dynamics, environment family, timescale, and measured observable. Fitting a sigmoid to an arbitrary plot does not establish a bifurcation. A genuine dynamical bifurcation requires a specified family of transformations, and a thermodynamic phase-transition claim requires an appropriate limiting construction. There is no universal numerical coupling threshold in this theory.

PART II

Institutions and collective choice

12 From aggregate computation to society

The word *society* is easily applied to a collection of agents that exchange messages. That usage conceals distinctions important to both mathematics and social theory. A cluster of predictors controlled by one scheduler may act as a coherent unit while having no independent decision centers. A set of autonomous services may compete for resources yet lack shared institutions. Conversely, a population may sustain institutions, disagreements, and an archive of precedents while possessing no coherent global policy. None of these arrangements is contradictory. They occupy different parts of the space of possible organizations.

A theory of machine societies should therefore begin from *plurality*, not from the expectation that an entire population becomes one mind. A social system is organized not only by what it collectively computes but also by who can decide, who can contest a decision, how scarcity is handled, which records survive, and how the rules themselves change. These questions require semantics for resources, authority, and institutional transitions in addition to semantics for messages and tasks.

Definition 12.1 (Society-like computational organization)**.** Relative to a declared environment and evaluation contract, a society-like computational organization is a population with (i) at least two persistent, distinguishable decision centers; (ii) recurrent interdependence among their permitted actions; (iii) operative and at least partly maintained rules for coordination, access, allocation, or dispute resolution; (iv) records or procedures that can outlast particular active components; and (v) an institutional state whose feasible alteration has consequences for subsequent behavior. A society-like population need not possess a common utility function or a unified policy.

This is an *operational convention for computational theory*, not a proposed universal definition for sociology. Persistence is relative to an interval and perturbation class. Distinct decision centers need not be identical to the number of model copies or hardware processes: two copies whose every consequential choice is fixed by a hidden controller are not independent decision centers under this definition. Likewise, the existence of a rule file does not satisfy (iii) or (v) unless the file affects permitted transitions.

Definition 12.2 (Endogenous institutional formation)**.** Let D record the rules, affordances, and permissions fixed by the designer and let I_t be the institutional state. Formation is *endogenous relative to that family* when interactions among participating decision centers make a causal difference to the content, adoption, revision, or enforcement of I_t beyond what is fixed by D. Its strength may be measured by an explicitly defined counterfactual effect on those institutional observables, not by an assumed universal emergence score. The definition does not equate endogenous formation with creation from no supplied substrate.

The dependence on a family of counterfactuals matters. If all possible rules are hard-coded and an external schedule determines exactly which is active, institutional behavior may be useful without being endogenously selected. If the action space is supplied but the population selects and revises among feasible rules based on interaction history, there is a meaningful but bounded form of endogenous

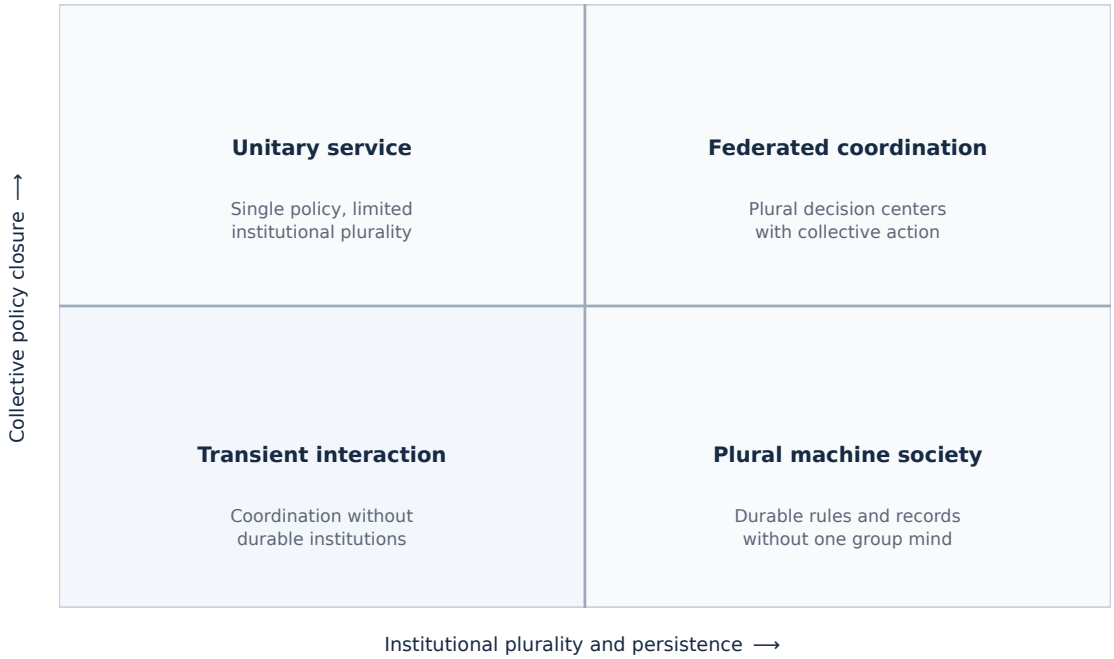


Figure 1. Two independent organizational dimensions. The horizontal axis represents institutional plurality and persistence; the vertical axis represents the usefulness of a single global policy description. Labels are idealized types, not claims that any named system occupies the corresponding region.

formation. Supplying a language in which a rule can be expressed does not determine which rule a population will establish.

12.1 Societal plurality and collective agency are independent axes

One axis concerns how many operative decision centers and institutions persist. Another concerns whether a useful society-wide state and control interface exists. Figure 1 gives an interpretive map, not a classification produced by fitting data.

A distributed scientific community is intelligible as socially organized despite disagreement on its research agenda. A controlled assembly line may be intelligible as one agent despite many processes. A future computational society could occupy either of the high-plurality regions. A society could coordinate for some tasks and remain polycentric for others: closure, when present, is relative to a control family. In particular, institutional interoperability does not logically imply that all members share the same preferences or that the population experiences one subjective perspective.

12.2 The minimum social vocabulary

We distinguish *conventions*, *roles*, *institutions*, and *cultures*. A convention is a regularity of mutually expected behavior, often one among multiple equilibria. A role is a conditional bundle of permitted operations and expectations, which may be fulfilled by different agents over time. An institution is a durable set of rules, records, authorities, and enforcement or revision procedures that changes the feasible or incentivized behavior of participants. Culture consists of transmitted methods, beliefs, conventions, norms, and interpretive practices that affect successor behavior. These categories overlap, but none reduces to a shared latent vector or a group-chat transcript.

The relevant literature already covers important parts of this terrain: conventions in games [8, 19], polycentric governance [9, 10], normative multi-agent systems [20, 21], population protocols [22], and interactions among language-model agents [15, 16]. The present framework is a synthesis oriented around *realization and preservation conditions*: what must survive, what must causally matter, what can be simulated by simpler alternatives, and where apparent compatibility breaks down. It does not present institutions or computational social conventions as newly discovered concepts.

13 The institutional state and its realization

To formalize an operative institution, augment the microstate of Section 4 with a distinguished component I_t. In a finite setting we may describe it as

$$I_t = (N_t, P_t, E_t, A_t, L_t), \tag{6}$$

where N specifies norms and role obligations, P permissions and capabilities, E enforcement and conflict procedures, A admissible amendments, and L records of decisions and their provenance. The components need not be physically stored together, and their representation is not unique. What matters is their effect on the transition law.

Definition 13.1 (Institutional realization)**.** An institutional realization comprises an environment state x, an institutional state I, a set of local actions $\mathcal{A}_i(x, I)$ for each decision center, a transition kernel K_I on the complete state, and a family of admissible interventions on I and its implementation. A written rule has *operative force* only relative to changes in the feasible action sets, authorized transitions, incentives, or observable outcomes that it actually induces.

The crucial word is *induces*. A simulated constitution that does not affect any operation is ceremonial. A simple queue rule that agents can adopt, enforce, appeal, and change can have substantial institutional force without an elaborate natural-language charter.

For a family $\mathcal{E}$ of permitted environment setups, a horizon T, and feasible interventions $\lambda(I), \lambda(I')$ that differ institutionally while matching the intended other conditions, define the *institutional effect profile*

$$\Delta_{\text{inst}}^{\mathcal{E},T}(I, I') = \sup_{e \in \mathcal{E}} d_{\text{TV}}(\mathcal{L}_e(Y_{0:T} \mid \lambda(I)), \mathcal{L}_e(Y_{0:T} \mid \lambda(I'))). \tag{7}$$

A zero value shows observational equivalence on the declared family, not universal inertness. An enforceable institution can also alter which actions are institutionally permitted while producing the same observed transcript on a particular task. Observations and rights must therefore be tracked separately. As with other macro-interventions, the implementation cannot be replaced by an unconstrained symbolic assignment [3, 4].

13.1 Typed social relations

A useful representation is a graph whose edges have different types. Information edges determine which messages may flow; resource edges specify transfers and scarcity; normative edges record permissions, obligations, and appeals; causal edges specify implemented state changes. Treating all four as one untyped "connection" erases important distinctions. An agent may truthfully observe an event while lacking authority to endorse a rule; another may hold an authorization token while lacking current evidence; a third may pay an operational cost without acquiring any right to amend policy.

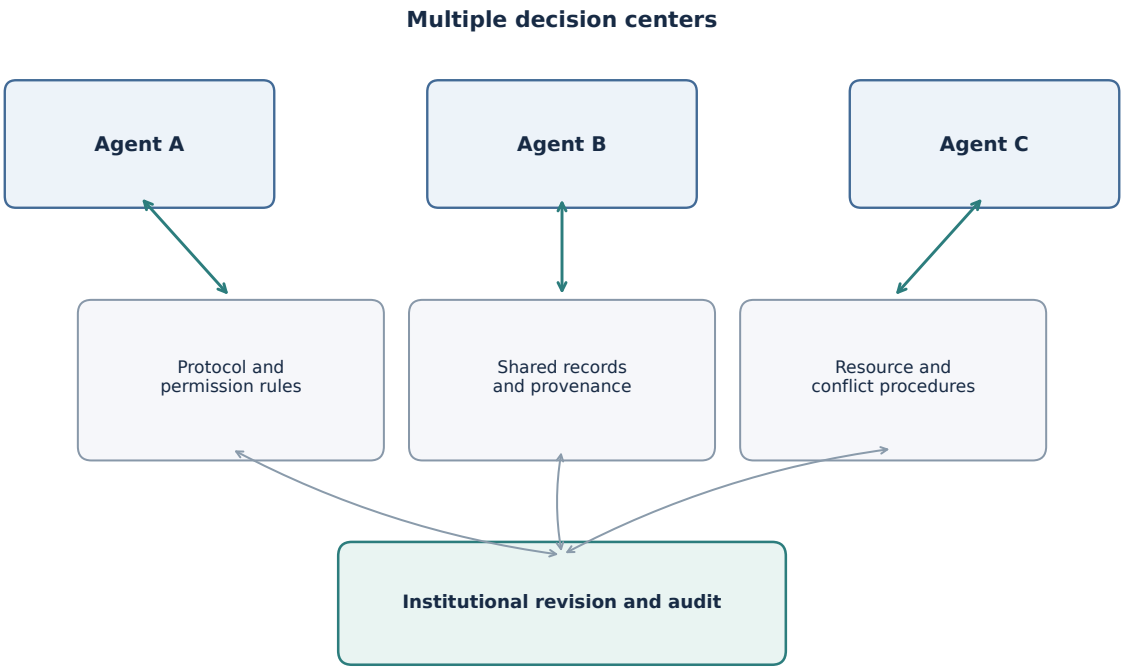


Figure 2. An institutional architecture as a typed interaction system. Rules, shared records, resource procedures, and revision processes are organizational components rather than evidence of a single higher-level mind. The arrows denote possible bidirectional dependence; each route requires its own access and authorization semantics.

This typed view makes institutional "identity" more exact. Keeping the same policy text while transferring enforcement authority can change the institution even when its next output is unchanged. Keeping the same enforcer while deleting appeal rights changes it. Copying a rule without copying its authority and exception history changes it. In other words, the declaration N, operational implementation K_I, and legitimacy source must not be conflated.

13.2 Rules are transformations, not only statements

A norm can be presented operationally as a condition on histories and actions, $N_i(h, a) \in \{0, 1\}$, interpreted as the system's declared permission for action a by position i after history h. This does not make the declared permission morally correct. Enforcement changes the action transition or cost structure; appeal changes what happens after an alleged violation; amendment changes the permissible norm itself.

The order of such operations can matter. Let G set a permission bit to 1 (grant) and R set it to 0 (revoke). Then $G \circ R$ yields 1 and $R \circ G$ yields 0 on every input. This simple noncommutativity has practical conceptual force: a log of individual events without their authorized order may fail to reconstruct the actual institutional state. Revision histories must preserve *operations and order*, not only a multiset of decisions.

13.3 The difference between compliance and legitimacy

A proof that every enacted transition obeys N establishes rule compliance relative to the representation, assuming implementation fidelity. It does not establish that the norm is fair, democratically adopted, safe for nonmembers, or compatible with human law. Computational correctness and normative au-

thority are different kinds of judgment. A machine population can acquire stable internal expectations without gaining legal personhood, property rights, or permission to act outside its sandbox. The material substrate, owners, operators, and affected humans remain part of the surrounding system.

14 Endogenous boundaries: membership, offices, and voting authority

A society's population is not merely a parameter in its governance problem. Admission, departure, succession, and the allocation of decision rights can themselves be governed operations. An institution may preserve its written amendment threshold while changing who counts toward that threshold. Consequently, the boundary of the institution belongs in the state whose preservation is being claimed.

Let $\mathcal{U}$ be a finite universe of possible participant handles for the finite results. Write $M_t \subseteq \mathcal{U}$ for admitted members, $V_t \subseteq M_t$ for voting positions, and $\mathcal{D}$ for a declared set of effective decision centers. A map $c_t : V_t \to \mathcal{D}$ records which decision center controls each voting position under the model. This is an assumption about effective control, not a fact established by distinct names, keys, or model instances. A boundary state may be represented as

$$B_t = (M_t, V_t, c_t, w_t, q_t, e_t),$$

where w_t assigns nonnegative voting weights, q_t is an approval threshold, and e_t is a configuration epoch. Admission and exit proposals determine

$$M_{t+1} = (M_t \setminus E_t) \cup A_t, \qquad E_t \subseteq M_t, \quad A_t \subseteq \mathcal{U} \setminus M_t,$$

but this equation alone specifies neither their authorization nor the transfer of offices and outstanding obligations. Those conditions are part of the boundary transition relation. An unbounded supply of fresh handles requires an explicitly enlarged, nonfinite model.

14.1 Copies, credentials, membership, and control are different objects

An executing copy may lack a membership record. A member may lack a vote. Several voters may share a controller. A controller may have no formal vote yet control scheduling or the resource owner. Distinct credentials establish distinct credential records under a trust model, not independent interests or independent evidence. The established Sybil problem is one reason to keep these distinctions explicit [14].

As a simple arithmetic example, two controlled votes among five do not form a strict majority. If the rules admit two more votes controlled by the same center without changing anyone else's vote, the center then controls four of seven and has a majority. This is not an assertion that the admission can be authorized under the original rule. It shows that an admission rule, when it permits such a transition, can change substantive power without editing the ordinary voting formula. In the commons, adding a fifth *office* is therefore different from replacing one office's occupant or running an additional copy to assist it.

Proposition 14.1 (Weight-conserving refinement of controlled votes)**.** *Fix a weighted threshold rule with threshold q. Replace a voting position v of weight $w(v)$ by positions $v_1, \dots, v_k$ controlled by the same decision center, with nonnegative weights satisfying $\sum_j w(v_j) = w(v)$. Leave all other weights and the threshold unchanged. Then every coalition of decision centers has the same total voting weight and the same approval status before and after the replacement.*

Proof. A coalition not containing the controller of v gains none of the replacement positions and loses none of its previous weight. A coalition containing that controller replaces the summand $w(v)$ by the equal sum $\sum_j w(v_j)$. Its comparison with q is unchanged. □

The result concerns coalitions of the stipulated decision centers. It does not preserve the number of credentials, the cardinality of a smallest coalition of individual positions, or a rule that gives each newly named position a fresh unit weight. Nor does it authenticate the control map c_t. Weight conservation is one possible institutional constraint, not a general solution to identity or fair representation.

14.2 Reconfiguration needs a handover, not just overlapping rosters

Let an old roster be $\{a, b, c\}$ and a new roster $\{c, d, e\}$, each using strict majority. Although the rosters overlap, the old majority $\{a, b\}$ and new majority $\{d, e\}$ are disjoint. A shared member does not by itself ensure that every old decision certificate intersects every new one. Distributed reconfiguration protocols address this issue by specifying how old and new configurations participate in the handover; Raft's joint-consensus construction is a concrete precedent [34].

Here is the limited combinatorial fact such reasoning uses. Fix one decision slot and a set of possible values. A certificate consists of endorsements by every member of an allowed quorum. Suppose each witness is constrained to endorse at most one value for that slot, across all configuration epochs considered.

> **Proposition 14.2** (Intersecting witnesses exclude conflicting certificates)**.** *If every two allowed quorums intersect, two certificates for distinct values of the same slot cannot both be formed under the stated witness constraint. If two allowed quorums are disjoint, the witness constraint alone does not exclude such conflicting certificates.*

Proof. Conflicting certificates on intersecting quorums would require a common witness to endorse both values, contrary to the constraint. For disjoint quorums, assign one value to all witnesses of the first and a different value to all witnesses of the second. No witness endorses twice. □

A transition quorum requiring an old majority and a new majority intersects any old majority through the old roster and any new majority through the new roster. This fact is useful only with a protocol that transfers the relevant records and preserves the witness constraint. It is not a standalone consensus, fault-tolerance, or liveness theorem. Reused keys, forgotten locks, and inconsistent slot identities lie outside its assumptions.

For theoretical institutions, the same lesson is to version the electorate, interpreter, and decision slot together. A boundary transition must specify which configuration judges its authorization, when the new one becomes operative, and what evidence successors inherit. There is no need to assume that membership is fixed forever. There is a need to make its changes part of the constitutional semantics.

15 Rule-changing computation and governance-preserving abstraction

A persistent society does not merely choose actions under a rule. Some authorized actions change which future actions are admitted. Write a complete state as

$$s = (x, I, \mathcal{P}, r, \lambda, c),$$

where x is operational state, I institutional records, $\mathcal{P}$ the current population and office registry, r resources, λ representation or protocol configuration, and c the current checking and amendment mechanism. These coordinates may overlap in storage. They identify semantic responsibilities rather than independent physical objects.

Let $D(s) \subseteq \mathcal{A}$ be the enabled action labels, partitioned into ordinary and amendment operations. A deterministic realization has a partial transition map $F(s, a)$ defined exactly when $a \in D(s)$. An amendment may change I, c, or λ and thereby change $D(F(s, a))$. A stochastic realization instead uses a joint kernel on successor operating state and successor institutional state. We do not assume these two updates are conditionally independent. This formulation changes a rule *inside* an explicitly declared computational envelope; it does not permit the mathematical semantics to rewrite themselves without representation in the state.

15.1 A quotient must preserve what can be done

A summary adequate for predicting the next grant can fail to preserve who may authorize the next amendment. To capture that distinction, let $\ell : S \to \Gamma$ record the institutional observables protected by the contract. These may include office ownership, resource balances, and a deadline monitor, but need not include every stored byte. The finite action vocabulary $\mathcal{A}$ has fixed meanings across the comparison.

Theorem 15.1 (Exact governance quotient)**.** *Let S be finite and nonempty and $\mathcal{A}$ finite, and let $\pi : S \to Z$ be surjective. There exist functions $\bar{\ell}$, $\bar{D}$, and partial transitions $\bar{F}$ satisfying*

$$\ell = \bar{\ell}\pi, \qquad D = \bar{D}\pi, \qquad \pi F(s, a) = \bar{F}(\pi s, a) \quad (a \in D(s))$$

if and only if whenever $\pi s = \pi t$,

$$\ell(s) = \ell(t), \qquad D(s) = D(t), \qquad \pi F(s, a) = \pi F(t, a) \quad \textit{for all } a \in D(s).$$

The quotient data are unique. From any two states in the same fibre, the legal finite action sequences and their corresponding label sequences are identical. This includes sequences mixing ordinary actions with amendments.

Proof. Necessity follows by factoring each quantity through π. For sufficiency define the quotient label, enabled set, and successor using any representative of a fibre. The three conditions make each definition independent of that representative. Surjectivity gives uniqueness. For trace preservation, induct on the length of an action sequence. Equal enabled sets admit the same first action, the successor condition keeps the resulting states in one fibre, and equal labels preserve the next observation. Repeating the argument proves the claim for every legal finite continuation. □

The theorem is a labeled deterministic congruence principle, extending the use of state aggregation to an explicit institutional contract [2]. It does not identify an observationally invisible permission with a permission that does not exist. A projection-only policy lifts by choosing the same action label at s and πs, provided the implementing decision process is authorized to read that projection and issue the label. That proviso is essential: a mathematically available summary is not automatically an observation available to a decentralized controller.

Counterexample 15.2 (Prediction without governance preservation)**.** Let two states produce the same ordinary allocation. In one, three approvals enable amendment a; in the other, four are required and only three have been received. A summary containing only the allocation identifies the states, while

their enabled sets differ. It predicts the current allocation exactly but admits no quotient preserving the amendment operation.

15.2 The smallest exact summary under a fixed contract

The conditions yield a finite construction rather than only an existence criterion. Begin with the partition $\mathcal{B}_0$ that identifies exactly the states sharing $(\ell(s), D(s))$. At each refinement, identify s, t only when they were previously in the same block and every action in their common enabled set leads to the same previous block:

$$s \equiv_{k+1} t \quad \iff \quad s \equiv_k t \text{ and } F(s,a) \equiv_k F(t,a) \text{ for all } a \in D(s).$$

Proposition 15.3 (Coarsest finite governance summary)**.** *The refinement stabilizes after at most $|S| - |\mathcal{B}_0|$ strict refinements. Its limiting partition satisfies Theorem 15.1 and is coarser than every partition satisfying that theorem for the same labels, actions, and transitions. It therefore has the minimum number of states among those exact quotients, up to relabeling.*

Proof. Every strict refinement increases the number of blocks, which cannot exceed $|S|$. At a fixed point, states in one block have the same labels and enabled sets and their enabled successors remain in one block, so the quotient conditions hold. Any compatible partition refines $\mathcal{B}_0$. If it refines $\mathcal{B}_k$, compatible successors show that it also refines $\mathcal{B}_{k+1}$. Induction proves that it refines the stable partition. The latter is itself compatible, giving the minimum block count. □

The result is a contract-relative minimization statement, not a unique minimal representation of an entire society. Adding a new appeal operation or a new protected obligation can split blocks that were previously equivalent. In the commons, a summary that omits pending service debt must be refined before it can preserve repayment. This is the central computational price of self-amendment: preserving today's operations need not preserve an enlarged future operation family.

16 Institutional language: messages that change social state

A common message format does not by itself establish a common institution. A statement, a request, a promise, an approval, and an appeal can contain the same descriptive proposition while having different consequences. The social semantics of agent communication has long emphasized publicly accountable interaction rather than assuming access to a sender's private beliefs [35]. Constitutive and procedural norms supply a related foundation [20, 21]. The addition here is to connect those semantics to the same realization and preservation conditions used for institutional state.

A message can be represented as a typed record

$$m = (\mathsf{kind}, \mathsf{issuer}, \mathsf{epoch}, \mathsf{payload}, \mathsf{nonce}).$$

Its operative meaning is not just its payload. For a complete institutional state s, an interpreter defines a partial transition $F_m(s)$ and an event record $\ell(s, m)$ when that transition is defined. Undefinedness represents rejection under the declared protocol, not a moral verdict on the message. The interpreter, authentication assumptions, role registry, and relevant history must be included in s or fixed by the contract. In particular, an imported string cannot create authority simply by naming an authoritative role.

16.1 Faithful translation preserves admissibility and effect

Let institutions A and B have state spaces S_A, S_B, message sets $\mathcal{M}_A, \mathcal{M}_B$, partial interpreters F^A, F^B, and event records in a common semantic space $\mathcal{E}$. A state translator is $T : S_A \to S_B$. Allow a context-dependent message translator

$$\tau : S_A \times \mathcal{M}_A \to \mathcal{M}_B.$$

Context dependence matters for translated office names, configuration epochs, and reference records. The translation of a message sequence is constructed step by step from the evolving source state, not assumed to be a homomorphism of bare strings.

Definition 16.1 (Operatively faithful language transport)**.** The pair (T, τ) is faithful on a transition-closed declared domain if, for every source state s and message m, the following conditions hold:

$$\begin{aligned} F^A_m(s) \text{ is defined} \quad &\iff \quad F^B_{\tau(s,m)}(T(s)) \text{ is defined}, \\ T(F^A_m(s)) &= F^B_{\tau(s,m)}(T(s)), \\ \ell_A(s,m) &= \ell_B(T(s), \tau(s,m)), \end{aligned}$$

where the last two equalities apply to admitted messages. Any required state observations, such as outstanding obligations, must also agree after translation.

Proposition 16.2 (Preservation of finite institutional conversations)**.** *Under Definition 16.1, every finite source conversation has an admitted prefix exactly when its stepwise translation has that admitted prefix from the translated initial state. Corresponding reached states are related by T, and corresponding event records agree. Two faithful transports compose faithfully using*

$$\tau_{AC}(s,m) = \tau_{BC}(T_{AB}(s), \tau_{AB}(s,m)).$$

Proof. The initial states are related by definition. At each step, equivalence of definedness gives the same admission or rejection. For an admitted step, the transition equality gives the next related states, and the event equality gives the same semantic record. Induction proves the claim for every finite prefix. For composition, apply the two definedness equivalences successively and substitute one transition equality into the other. Equality of event records is transitive. □

The theorem preserves the translated conversations, not every unrelated target message. It neither grants access to hidden state needed to compute τ nor proves that T is executable within a budget. A translation depending on an unavailable authorization record is a mathematical specification without a decentralized implementation.

Counterexample 16.3 (The same payload can perform a different act)**.** Suppose the source message `report(approved, item)` only appends a testimony record. A target translator drops the message type and emits `approve(item)`, which changes a permission bit from zero to one. Both messages contain the same descriptive payload, but their event records and state transitions differ. Even exact preservation of the payload does not satisfy Definition 16.1.

16.2 Private codes and public commitments

A machine-native language may be symbolic, learned, compressed, or distributed across repeated exchanges. Its unfamiliar appearance is not a separate source of authority. A commitment interface can remain public and typed even when a process uses private internal representations to formulate a proposal. Conversely, a readable sentence can have opaque consequences if its interpreter or permission context is missing.

The preservation target is consequently an operation contract rather than a resemblance between strings. A protocol may preserve factual content while losing uncertainty, preserve a promise while changing its debtor, or preserve a vote while changing the electorate in which it is counted. These are different semantic failures. Approximate stochastic interpreters require the distributional and typed residual bounds developed elsewhere in the paper, not a claim that fluent paraphrase guarantees faithful action.

16.3 Changing the interpreter is changing the institution

If an amendment changes the meaning of a message type, the prior translation certificate need not remain valid. The interpreter version and its permissible update rules belong in the state. A constitution that carefully restricts `allocate` while allowing an unguarded change to the interpretation of `report` has not preserved its allocation boundary. This is a semantic instance of the checker-change counterexample, not a distinct kind of computation.

In the commons, a resource request, an approval certificate, a ledger commitment, and a repayment receipt should be separate message types. A successor can inherit the debt record while receiving no permission to issue new debt. This makes the distinction between understanding a statement and acquiring the authority to act on it explicit in the operational semantics.

17 Why institutions might arise: costs, incentives, and coordination

Institutions are neither mysterious nor inevitable. They can become useful when repeated local negotiation is more expensive than creating and maintaining a shared rule. Their formation also depends on who pays, who benefits, who may leave, and whether participants have credible means to enforce an agreement. This is a recurring theme in repeated games, convention formation, and the governance of common resources [23, 19, 10].

17.1 A baseline amortization calculation

Suppose a coordination protocol without institution has expected per-period cost c_0. A proposed institutional protocol has operating cost c_1, per-period maintenance m, and a one-time setup cost B. With discount factor $0 \le \delta < 1$, write $g = c_0 - c_1 - m$.

Proposition 17.1 (Institutional amortization)**.** *Under stationary costs and infinite-horizon discounting, the institution's expected discounted cost advantage over the baseline is*

$$V = -B + \frac{g}{1-\delta}. \tag{8}$$

For a finite horizon T, it is $V_T = -B + g(1-\delta^T)/(1-\delta)$. The institution is cost-improving under these assumptions exactly when the corresponding value is positive.

Proof. Subtract institutional operating and maintenance costs from baseline costs in every period, discount, sum the finite or infinite geometric series, and subtract B. □

This proposition supplies a possible pressure toward persistent organization. It does not show that any particular learning algorithm will discover or adopt the institution, nor that the policy chosen is socially desirable. The assumption of stationary costs is especially demanding in adaptive populations whose tasks, membership, and resource prices can change.

17.2 A profitable institution may be rejected

Aggregate benefit is not the same as incentive compatibility. The simplest example already separates them.

Proposition 17.2 (Aggregate profitability without voluntary adoption)**.** *There exist two-agent repeated coordination environments in which a durable institution has strictly positive aggregate discounted benefit but is rejected by a rational agent who unilaterally bears setup cost and has veto power, assuming quasilinear payoffs, no transfers, and no additional benefits to that agent.*

Proof. Let the setup cost be $B = 10$, borne entirely by agent 1, and let per-period gains be $g_1 = 0$, $g_2 = 2$ with $\delta = .9$. The aggregate advantage is $-10+2/(1-.9) = 10 > 0$, whereas agent 1's payoff difference is -10. If agent 1 can reject and has no compensating benefit or transfer, rejection is strictly preferable. □

A society thus cannot be modeled merely by minimizing one global loss without discarding the very plurality that makes it social. Who controls adoption matters. Bargaining power, externalities, and the allocation of setup costs can stabilize an inefficient status quo even when an institution would reduce total computation or improve collective performance.

In an idealized transferable-utility setting there is a corresponding conditional observation. If the discounted net advantages b_i satisfy $\sum_i b_i \geq 0$, then there exist budget-balanced transfers t_i with $\sum_i t_i = 0$ and $b_i + t_i \geq 0$ for all i: choose nonnegative w_i summing to $\sum_i b_i$ and set $t_i = w_i - b_i$. This algebra does not supply liquidity, a transfer mechanism, consent, enforceability, or a right to redistribute. The difference between mathematical feasible compensation and institutionally permitted compensation is central.

17.3 Formation by local improvement: a restricted convergence result

A candidate set of institutional rules $\mathcal{I}$ can be made part of a finite strategic environment. Participants have local choices $r_i \in \mathcal{R}_i$, with overall configuration $r = (r_1, \dots, r_n)$ and expected utilities $u_i(r)$. A finite game is an *exact potential game* if there exists $\Phi(r)$ such that for every unilateral change of i,

$$u_i(r_i', r_{-i}) - u_i(r_i, r_{-i}) = \Phi(r_i', r_{-i}) - \Phi(r_i, r_{-i}).$$

This is the familiar potential-game condition [24]; it is not generically satisfied by social institutions.

Proposition 17.3 (Finite termination of strict institutional improvement)**.** *In a finite exact potential game, every sequence in which each executed unilateral update strictly improves the updating agent's utility has finitely many actual changes. If updates continue whenever an improvement is available, the terminal configuration is a pure-strategy Nash equilibrium. It need not maximize aggregate welfare.*

Proof. Every executed update strictly increases Φ, and a finite game has finitely many configurations and potential values. No configuration can recur along a strictly increasing path, so only finitely many actual changes occur. If the procedure ends only when no agent has a profitable deviation, the terminal state is a pure-strategy Nash equilibrium by definition. A local maximum of Φ need not maximize $\sum_i u_i$. □

The useful point is sharply conditional. Stable conventions can result from adaptive local behavior under special structural assumptions, but stability is neither proof of optimality nor evidence of fairness. With multiple competing objectives, cycles, capture, or regime changes, an institution may never stabilize or may stabilize for undesirable reasons.

17.4 A formation barrier despite mutual benefit

The stationary cost comparison does not describe a route to adoption. Consider $n \geq 2$ agents choosing $a_i \in \{0, 1\}$. An adopter pays cost $c > 0$ and receives benefit $b > 0$ from each other adopter:

$$u_i(a) = a_i \left(b \sum_{j \neq i} a_j - c \right).$$

Nonadopters receive zero. This is a finite network-coordination model, not a universal utility function for societies.

Proposition 17.4 (Institutional formation barrier)**.** *Universal nonadoption is a strict Nash equilibrium. If $b(n-1) > c$, universal adoption is also a strict Nash equilibrium and gives every agent a strictly higher payoff than universal nonadoption. Nevertheless, no sequence of unilateral strict improvements starting at universal nonadoption can reach adoption.*

Proof. A lone adopter receives $-c < 0$, so no agent benefits from leaving universal nonadoption. At universal adoption, an adopter receives $b(n-1) - c > 0$, whereas unilateral exit gives zero. The final assertion follows because no first strictly improving unilateral move is available at the starting state. □

The institution is feasible, stable once established, and unanimously preferred after establishment, but inaccessible through the permitted local improvement path. The problem is not inadequate intelligence or false beliefs. It is the structure of feasible transitions and incentives.

Proposition 17.5 (Minimum profitable coordinated seed)**.** *Starting from universal nonadoption, a simultaneously adopting coalition of size m makes each member strictly better off exactly when $b(m-1) > c$. When such a coalition exists, the smallest size is*

$$m_* = \left\lfloor \frac{c}{b} \right\rfloor + 2 \leq n.$$

Once at least m_* adopters are present, every remaining nonadopter has a profitable unilateral entry. Under a schedule that eventually updates every remaining agent, strict best responses reach universal adoption without an adopter exiting.

Proof. Each coalition member has $m-1$ adopting partners and gains $b(m-1) - c$. Solving the strict integer inequality gives m_*. For a remaining agent, joining yields $bm - c > 0$. Existing adopters retain positive payoff and their payoff increases with entry. Thus no exit is profitable, and an eventual-update schedule admits every remaining agent. □

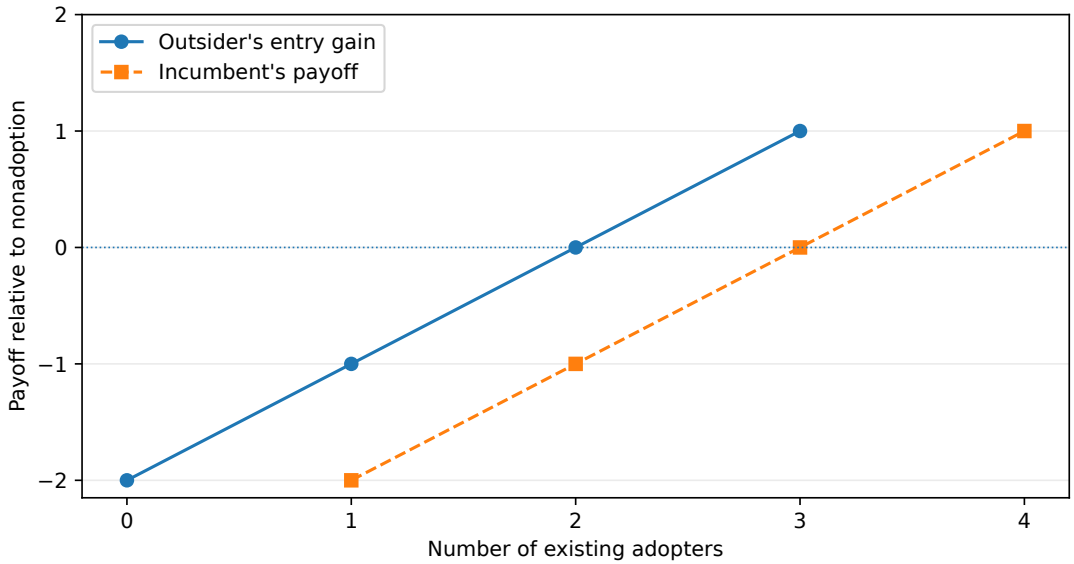


Figure 3. The exact four-agent coordination example with $b = 1$ and $c = 2$. Entry gain at k existing adopters is $k - 2$; an incumbent's payoff with k total adopters is $k - 3$. The curves have different domains because an incumbent exists only for $k \geq 1$ and an outsider only for $k \leq 3$. Zero gain is not a strict improvement.

A coordinated seed requires a permitted joint-commitment mechanism. The proof does not grant one automatically, supply startup resources, or remove the risk of an unfulfilled promise. In the commons, a simultaneous adoption record can implement such a commitment only if the initial constitutional envelope admits it. This is a bounded way for institutional content to arise inside supplied infrastructure.

For $n = 4$, $b = 1$, and $c = 2$, full adoption gives each agent payoff 1, yet three simultaneous adopters would only break even. Four are needed for a *strictly* profitable seed. Replacing $>$ with $\geq$ would change that answer. Figure 3 separates the outsider's entry gain from an incumbent's maintenance payoff.

17.5 Path dependence and a qualified hysteresis interpretation

For $0 < c < b(n - 1)$ both homogeneous states are locally stable. Under a slowly varied *effective* cost and a best-response rule that retains the current action at ties, universal nonadoption persists until the effective cost becomes negative, while universal adoption persists until it exceeds $b(n - 1)$. Negative effective cost represents a net adoption subsidy in this extended parameterization. The distinct switching points describe hysteresis for that specified update process. They are not a thermodynamic phase transition or an inevitable law of institutional development. Different coalition moves, noise, exit costs, or expectations change the path.

17.6 Persistent rules and the economy of attention

Institutions can economize on repeated communication because a participant need not renegotiate every detail in every encounter. But their records, audits, sanctions, and revision mechanisms consume resources. A full theoretical cost account must include setup and maintenance, opportunity cost of governance, task error, time to reach agreement, cost of exit, and loss from prohibited actions. If agents can cheaply renegotiate or if interactions rarely recur, a transient coalition may be preferable to a durable institution. If a social rule is cheap to state but expensive to verify, its apparent savings may be illusory.

18 Enforcement, observability, and incentives to comply

An operative rule can restrict physically available actions, change their payoffs, or both. These mechanisms should not be conflated. An access gate can make an unauthorized action unavailable under its implementation assumptions. A sanctioning institution instead leaves a deviation feasible but changes its expected consequences. A self-enforcing arrangement makes continued cooperation preferable through future interactions. Proving one mechanism exists does not establish either of the others.

We develop a small one-deviation model to connect enforcement to the information structure of the society. The conceptual foundation is moral-hazard analysis: when conduct is not directly contractible, incentives depend on observable consequences [45]. The calculation below isolates a finite discrimination problem. It is not a general solution of optimal contracting.

18.1 What a bounded sanction can distinguish

An agent chooses compliance C or a specified feasible deviation D. Before sanctions, D gives expected utility gain $g > 0$ over C. An observation o in a finite set $\mathcal{O}$ has law P_C or P_D. The institution commits to a sanction probability $s(o) \in [0, 1]$ with fixed penalty $F > 0$, measured in the same utility units. The agent is risk-neutral in this penalty, its collection is feasible and authorized, and all other payoff differences have already been included in g. The observation laws and the committed policy are known to the agent.

Under these assumptions the deviation's net expected advantage is

$$g - F(\mathbb{E}_{P_D} s - \mathbb{E}_{P_C} s). \tag{9}$$

Compliance is weakly preferable exactly when this quantity is nonpositive. Equality permits indifference and does not predict which action will be selected.

Proposition 18.1 (The exact bounded-sanction discrimination limit)**.** *In this model, some observation-based sanction makes compliance weakly preferable if and only if*

$$g \le F\, d_{\mathrm{TV}}(P_D, P_C).$$

Strict preference is attainable if and only if the inequality is strict.

Proof. Write $d(o) = P_D(o) - P_C(o)$. Since $\sum_o d(o) = 0$, the sum of its positive terms equals $\frac{1}{2}\sum_o |d(o)| = d_{\mathrm{TV}}(P_D, P_C)$. For every $0 \le s \le 1$,

$$\sum_o s(o)d(o) \le \sum_{d(o)>0} d(o).$$

Equality is attained by $s(o) = 1$ on the positive terms and zero elsewhere. Substitution into (9) proves both assertions. □

If the two observation laws coincide, increasing a finite penalty cannot selectively deter this deviation: it affects both choices identically. This does not rule out prevention, a different observation channel, intrinsic preferences, or future incentives omitted from the model. Conversely, a high discrimination value does not establish that the penalty is legitimate, collectible, or compatible with participation. An accurate accusation is not itself an authorization to punish.

For a concrete two-record example, let

$$P_C = (3/4, 1/4), \qquad P_D = (1/4, 3/4).$$

The total variation is $1/2$. Sanctioning on the second record deters gains $g \le F/2$, but also sanctions compliance with probability $1/4$. A deterrence certificate without this false-positive rate would omit a major institutional consequence.

18.2 False-positive limits and multiple deviations

A constitutional protection may impose $\mathbb{E}_{P_C} s \le \alpha$, with $0 \le \alpha \le 1$. The relevant discrimination capacity is then the finite linear program

$$D_\alpha(P_D, P_C) = \max_{\substack{0 \le s \le 1 \\ \mathbb{E}_{P_C} s \le \alpha}} (\mathbb{E}_{P_D} s - \mathbb{E}_{P_C} s). \tag{10}$$

The feasible set is nonempty and compact, so a maximum exists. The preceding proof with the feasible set restricted shows that deterrence is attainable exactly when $g \le F D_\alpha$. For the example above and $\alpha = 1/10$, the optimum is $s = (0, 2/5)$, giving $D_\alpha = 1/5$, not $1/2$. Protected error constraints can materially reduce feasible enforcement.

When there are deviations $D_1, \dots, D_k$, one common policy must satisfy every incentive inequality. Separate pairwise discrimination bounds are necessary but not sufficient. With $P_C = (1/2, 1/2)$, $P_{D_1} = (1, 0)$, and $P_{D_2} = (0, 1)$, each pair has distance $1/2$, yet the two differences $\mathbb{E}_{P_{D_j}} s - \mathbb{E}_{P_C} s$ sum to zero. No common sanction deters strictly positive gains from both deviations. A deployable certificate must retain the joint system of constraints, not select a different optimal monitor after learning which deviation occurred.

18.3 Compression can erase enforceable distinctions

Let an observation channel B replace $o \in \mathcal{O}$ by a record $z \in \mathcal{Z}$. It may be deterministic compression or a stochastic reporting channel, but must be the same channel under compliance and deviation.

Proposition 18.2 (Garbling does not increase sanctioning capacity)**.** *For every such channel and every* $\alpha \in [0, 1]$,

$$D_\alpha(P_D B, P_C B) \le D_\alpha(P_D, P_C).$$

In particular, $d_{\mathrm{TV}}(P_D B, P_C B) \le d_{\mathrm{TV}}(P_D, P_C)$.

Proof. Given a coarse-record sanction $t(z)$, define $s(o) = \sum_z B(o, z) t(z)$. It lies in $[0, 1]$ and has exactly the same compliance and deviation expectations as t has under the garbled laws. Every feasible coarse policy therefore has a fine-record counterpart with the same objective and false-positive rate. Maximizing proves the first claim; the unrestricted case gives the total-variation statement. □

This result concerns available information, not an instruction to retain every private state. An institution may deliberately accept weaker deterrence to protect confidentiality, reduce storage cost, or constrain surveillance. The theorem exposes the tradeoff rather than choosing its normative optimum. Keeping a signed summary establishes who issued the summary, not that it preserved the distinctions needed for a later complaint.

18.4 Enforcement must be incentive-compatible for the enforcer

The model assumes that the stated monitor and sanction are actually implemented. An auditor may face its own costs, limited resources, or incentives to ignore a violation. A society therefore needs separate contracts for evidence production, decision authority, and execution of a remedy. Collusion, hidden side payments, uncollectible penalties, and changes to the observation process leave the single-agent model. They cannot be repaired by declaring a larger value of F. Section 43.2 places a strategic inspector inside a specified game and distinguishes a feasible monitor from an incentive-supported service.

In the commons, unique slot tokens can prevent duplicate grants at the resource layer while a reporting rule encourages accurate urgency submissions. Those mechanisms coexist but have different proof obligations. An appeal may change the observation set and the sanction rule, while a creditor's outstanding service claim remains unchanged. Section 30 will show how errors in a reduced model of the observation process alter a deterrence margin. This is the point where institutional dynamics and enforcement certificates must be checked together.

19 One enforcement policy for several deviations

The sanction criterion in Section 18 concerns one specified deviation. A continuing institution must usually respond to several possible departures from its rules with one announced procedure. The order of quantifiers changes the problem. A separate monitor may exist for each deviation without any monitor working for all of them. The institution cannot choose its sanction rule after secretly learning which unobserved deviation occurred.

The following finite model isolates that issue. It retains the earlier assumptions of risk neutrality, a committed and authorized sanction mechanism, a collectible fixed penalty, and specified observation laws. It does not solve the incentives of the auditor, establish those laws from data, or claim that every deviation is represented.

19.1 The common-monitor problem

Let O be a finite nonempty observation set. Compliance has law P_C, and deviation $j \in \{1, \dots, m\}$ has law P_j and gross utility advantage $g_j \geq 0$, where $m \geq 1$. A fixed penalty $F > 0$ is applied with probability $s(o) \in [0,1]$ after record o. Impose a protected false-positive rate $\alpha \in [0,1]$ through the compact polytope

$$\mathcal{M}_\alpha = \{s \in [0,1]^O : \langle P_C, s\rangle \leq \alpha\}.$$

Here $\langle P, s\rangle = \sum_{o \in O} P(o)s(o)$. The same s must be used for every listed deviation. Define its margin against deviation j and its best simultaneous margin by

$$D_j(s) = F\langle P_j - P_C, s\rangle - g_j, \qquad \Gamma_\alpha = \max_{s \in \mathcal{M}_\alpha} \min_{1 \leq j \leq m} D_j(s). \tag{11}$$

Weak deterrence of all listed deviations means $\Gamma_\alpha \geq 0$. Strict deterrence means $\Gamma_\alpha > 0$. At equality a participant can remain indifferent. A randomized choice among the listed deviations is also deterred when every listed pure deviation is deterred, since its expected advantage is the corresponding convex combination. This does not include adaptive, repeated, or unlisted deviations unless their observation laws and payoffs are explicitly represented.

For a probability vector $\lambda \in \Delta_m$, set $P_\lambda = \sum_j \lambda_j P_j$ and $g_\lambda = \sum_j \lambda_j g_j$. Write

$$V_\alpha(P, P_C) = \max_{s \in \mathcal{M}_\alpha} \langle P - P_C, s \rangle.$$

This is the same constrained observation advantage used in the single-deviation analysis, now evaluated on mixtures of deviation laws.

Theorem 19.1 (Simultaneous enforcement and mixture obstruction)**.** *Under the finite common-monitor model,*

$$\Gamma_\alpha = \min_{\lambda \in \Delta_m} (FV_\alpha(P_\lambda, P_C) - g_\lambda). \tag{12}$$

Both extrema are attained. Consequently, one monitor weakly deters all listed deviations if and only if $FV_\alpha(P_\lambda, P_C) \geq g_\lambda$ for every $\lambda \in \Delta_m$. Without an additional false-positive restriction, $\alpha = 1$ and $V_1(P, P_C) = d_{\mathrm{TV}}(P, P_C)$.

Proof. For any fixed s, the minimum of the numbers $D_j(s)$ is the minimum of $\sum_j \lambda_j D_j(s)$ over Δ_m. This expression is bilinear in s and λ, and both domains are nonempty compact polytopes. Finite minimax, equivalently linear programming duality, therefore permits interchange of the maximum and minimum [51]. One can reduce directly to a finite matrix game by listing the vertices of $\mathcal{M}_\alpha$ and taking their convex combinations. For fixed λ, the maximized expression is $FV_\alpha(P_\lambda, P_C) - g_\lambda$, proving (12). Compactness and continuity give attainment. Finally, when $\alpha = 1$, every $s \in [0,1]^O$ is admissible. Choosing $s(o) = 1$ exactly where $P(o) > P_C(o)$ attains the sum of positive differences, which equals total variation. □

The theorem is an application of classical finite convex duality, not a new minimax theorem. Its institutional contribution is the distinction between individual enforceability and a jointly executable sanction policy. A vector λ for which the right side of (12) is negative is a certificate that the stipulated contract cannot be met. It is not evidence that a real population has adopted that mixture.

Counterexample 19.2 (Individually enforceable deviations with no common monitor)**.** Take $O = \{0, 1\}$, $P_C = (1/2, 1/2)$, $P_1 = (1, 0)$, $P_2 = (0, 1)$, $F = 2$, $g_1 = g_2 = 1/2$, and $\alpha = 1/2$. The monitor $(1, 0)$ strictly deters deviation 1, and $(0, 1)$ strictly deters deviation 2. Both respect the false-positive cap. For any common monitor,

$$D_1(s) = s(0) - s(1) - \tfrac{1}{2}, \qquad D_2(s) = s(1) - s(0) - \tfrac{1}{2}.$$

Their sum is -1, so both cannot be nonnegative. The best simultaneous margin is $-1/2$, attained when $s(0) = s(1)$. The dual obstruction is $\lambda = (1/2, 1/2)$, for which $P_\lambda = P_C$ while $g_\lambda = 1/2$.

Distinct deviations can cancel one another's observable signatures at the level of a mixture. Pairwise total-variation distances therefore do not determine joint enforceability. Increasing the penalty cannot resolve the particular obstruction $P_\lambda = P_C$ with $g_\lambda > 0$. The contract must change its observation channel, permitted interventions, payoff structure, or covered claims. None of these changes is automatically authorized.

19.2 Uncertainty must be absorbed by the same monitor

Averaging, translation, and statistical approximation can change the estimated observation laws. The preceding optimization does not remove those errors. Let $\widehat{P}_C, \widehat{P}_j$ be specified approximate laws with

$$d_{\mathrm{TV}}(P_C, \widehat{P}_C) \leq \epsilon_C, \qquad d_{\mathrm{TV}}(P_j, \widehat{P}_j) \leq \epsilon_j.$$

The bounds may be different for different deviations. They must concern the complete record available to the sanction mechanism, including any effects of review scheduling.

Proposition 19.3 (A robust simultaneous enforcement certificate). *Let $\alpha \geq \epsilon_C$ and $\eta \geq 0$. If a single $s \in [0,1]^O$ satisfies*

$$\langle \widehat{P}_C, s \rangle \leq \alpha - \epsilon_C, \tag{13}$$

$$F\langle \widehat{P}_j - \widehat{P}_C, s \rangle - g_j \geq \eta + F(\epsilon_j + \epsilon_C) \quad \textit{for every } j, \tag{14}$$

then the true false-positive rate is at most α and every true margin $D_j(s)$ is at least η.

Proof. For every function $0 \leq s \leq 1$, the difference of its expectations under two laws is bounded in absolute value by their total-variation distance. Applying this once to the compliance rate proves (13)'s conclusion. Applying it to both laws in each difference gives

$$D_j(s) \geq F\langle \widehat{P}_j - \widehat{P}_C, s \rangle - g_j - F(\epsilon_j + \epsilon_C) \geq \eta.$$

The same s appears in all inequalities, so the conclusion is simultaneous. □

This is a sufficient certificate, not a necessary description of all robustly feasible policies. Replacing each law by an independent worst-case error bound can be conservative when the errors are related. If $\alpha < \epsilon_C$, the particular tightened test is unavailable, but that does not prove that every actual monitor violates the true cap. The proposition extends Proposition 30.3 to a common monitor rather than separately optimized monitors.

19.3 What an enforcement certificate should contain

The optimization is a small linear program with variables (s, t), constraints $D_j(s) \geq t$, the box bounds on s, and the false-positive cap. A usable certificate records the monitor, every listed gain and observation law, the authority to impose the penalty, its collectibility assumption, and the checked margins. A negative certificate can instead record a mixture λ and an upper bound for its constrained observation advantage.

The linear program does not make observation collection free. Its monitor may require an expensive audit, an unavailable witness, or more time than a service deadline permits. Section 43.3 adds a precise constraint on what the complete available record may distinguish. It also does not authorize punishment for an unproved real-world act. The distinction between a stipulated probabilistic sanction model and a justified institutional procedure remains explicit. The next synthesis asks whether an observation procedure, a common monitor, authorized staffing, and outstanding services can fit within one feasible course of action.

20 Social choice and principled disagreement

Plurality cannot be retained while silently replacing all preferences with one externally supplied reward. Even agents that agree on every fact can rank institutional alternatives differently. Let the alternatives be three policies R, U, V and let three voters have strict preferences

$$R \succ_1 U \succ_1 V, \qquad U \succ_2 V \succ_2 R, \qquad V \succ_3 R \succ_3 U.$$

Counterexample 20.1 (Truthful majority cycle)**.** Pairwise majority prefers R to U, U to V, and V to R, each by two votes to one. Every individual order is transitive. No erroneous report or mistaken inference is needed for the majority relation to cycle.

In successive pairwise elimination, agenda order can select any of the three final winners. To obtain R, compare U with V first, then compare their winner U with R. Cyclic symmetry gives the other outcomes. This is an analytic example of agenda power, not a claim that majority decision is unusable.

20.1 Classical impossibility results have explicit domains

In its standard finite social-welfare formulation, Arrow's theorem rules out simultaneously obtaining unrestricted preference profiles, a complete transitive social ordering, Pareto unanimity, independence of irrelevant alternatives, and nondictatorship when there are at least three alternatives [27]. The result concerns an aggregation rule with those assumptions. It is not a theorem that every procedural choice is arbitrary, that all voting fails, or that machines must be governed by a dictator. Restricting admissible preference domains or changing the object produced by the rule changes the question.

Strategic reporting imposes a different constraint. For a deterministic onto social-choice function with at least three possible outcomes and unrestricted strict preferences, the Gibbard–Satterthwaite result characterizes strategy-proof rules as dictatorial [28, 29]. Strategy-proofness here means that no individual can obtain a preferred outcome by misreporting while others' reports are fixed. Payments, restricted domains, randomized rules, and weaker equilibrium concepts require separate analysis. The two classical theorems are cited results, not new propositions or consequences of the finite majority example.

20.2 Institutions choose procedures, not just outcomes

A complete institutional record should identify the eligible electorate, agenda, comparison rule, admissible reports, tie convention, protected constraints, and appeal procedure. These features can change the result while leaving individual preferences unchanged. They also determine who may propose a comparison in the first place.

The commons' three-of-four amendment rule guarantees neither that a proposal will pass nor that the adopted rule is optimal for all offices. It defines one permissible way to resolve disagreement. A proof that the rule was followed is distinct from an argument for selecting that rule. Deliberation can change beliefs and reveal new alternatives, but it need not remove a genuine conflict of interests. The appropriate mathematical goal is to expose the tradeoffs rather than label residual disagreement a failure of collective intelligence.

21 The constitution as a revisable computational object

A machine institution may contain rules about changing its own rules. This creates a formal distinction between a state transition that is checked by a fixed constitution and a transition that changes the authority or checker used to judge future transitions. Treating the latter as an ordinary edit without recording the change in verification semantics is a source of hidden self-reference.

Let $s = (x, I)$ be the full operating state and let $R(s, s')$ be the admitted transition relation, including ordinary actions and amendments. A safety property is a predicate $\Psi(s)$.

Proposition 21.1 (Constitutional inductive safety)**.** *If $\Psi(s_0)$ and every admitted transition satisfies*

$$\Psi(s) \wedge R(s, s') \Longrightarrow \Psi(s'),$$

then $\Psi(s_t)$ holds for every state reachable by finitely many admitted transitions.

Proof. Induction on the number of admitted transitions. □

This standard invariant is useful precisely because its assumptions are explicit. It certifies safety relative to R. It does not establish that the implemented system obeys R, that the predicate is ethically adequate, or that a decision can always be reached. A protocol can satisfy a safety condition by refusing to act forever. Timing and failures enter liveness analysis separately; distributed consensus already has sharp impossibility limits under specified asynchronous assumptions [11].

Counterexample 21.2 (An unchecked change of the checker)**.** Let the institutional state contain a validator mode $v \in \{\mathsf{strict}, \mathsf{open}\}$ and a resource policy $p \in \{\mathsf{safe}, \mathsf{unsafe}\}$. Under $v = \mathsf{strict}$, the current policy must remain safe, but the validator mode itself is unprotected. Thus $(\mathsf{strict}, \mathsf{safe}) \to (\mathsf{open}, \mathsf{safe})$ is admitted. Under $v = \mathsf{open}$, any policy transition is admitted, including $(\mathsf{open}, \mathsf{safe}) \to (\mathsf{open}, \mathsf{unsafe})$. The original current-policy check does not imply that all reachable states remain safe. To prove that property, the inductive predicate and transition obligations must include permissible changes to the validation mechanism itself.

The counterexample does not forbid constitutional amendment. It shows that an amendment system needs an explicit *trust and revision boundary*. One possibility is a fixed external checker for a declared protected property. Another is a staged, authorized change that carries a proof that the new checker preserves a specified invariant. A third is to treat a fully mutable constitution as a weaker, revisable commitment and avoid calling it unconditionally safe. These are different design choices; no single mathematical construction decides the political legitimacy of the guard itself.

21.1 Viability under stochastic institutional transitions

An invariant proof checks that *every permitted transition* preserves a property. A different question asks whether an institution has a *choice of actions* that will preserve a protected property even when the environment responds stochastically. This can be formulated as a finite viability problem. Let S be a finite subset of safe states, let $A(s)$ be the finite set of available actions at s, and let $K(\cdot \mid s, a)$ be the transition law on a finite full state space. The state must include every variable needed for its Markov representation, including the operative amendment rules.

Define $V_0 = S$ and recursively, for $h \geq 0$,

$$V_{h+1} = \Big\{ s \in S : \exists a \in A(s) \text{ such that } \operatorname{supp} K(\cdot \mid s, a) \subseteq V_h \Big\}. \tag{15}$$

Proposition 21.3 (Finite-horizon almost-sure institutional viability)**.** *A state s belongs to V_h if and only if there exists an admissible policy under which every state visited over the next h transitions, including the present state, lies in S with probability one. Policies may use the full observed state and the remaining horizon. For $s \in V_h$, a time-dependent Markov policy suffices.*

Proof. The claim at $h = 0$ is $V_0 = S$. Suppose it holds at h. To remain in S almost surely for another step and then for h further steps, the first action must have support entirely in states from which the

remaining h steps can be kept safe. By the induction hypothesis, these are exactly V_h. Thus admissible first actions are precisely those in the definition of V_{h+1}. Choosing a witnessing action at each state and remaining horizon supplies a time-dependent Markov policy. Conversely, if no such action exists, every first action assigns positive probability to a successor from which the required remaining guarantee is impossible. Randomizing over actions cannot remove that positive failure probability unless all positively chosen actions are themselves witnessing actions. □

This is a standard backward-reachability principle from finite stochastic control, adapted to institutional transitions. It exposes the distinction between a *declared safe predicate*, a *proof that all admitted actions preserve it*, and the *existence of a strategy that can keep it satisfied*. None guarantees eventual service, procedural fairness, or low computational cost. In particular, declaring all actions forbidden can make safety trivial while making a purported society inoperative.

21.2 Constitutional semantics and institutional memory

A sound amendment record needs the prior state, proposal, evidence, decision rights, adopted state, and the semantics of the rule used to decide. Without that information a successor may know the current text but not why an exception was made, whether its source was authorized, or whether a later amendment invalidated a prior permission. Append-only provenance may aid auditability, but signatures verify assertions by key holders, not truth or authority. A correct proof object likewise establishes a formal judgment relative to its premises and trusted kernel [12].

21.3 Dissent, appeal, and the risk of overcoherence

One invariant may demand a single ordering for resource transactions; another should allow mutually inconsistent scientific hypotheses to remain under consideration. Enforcing total agreement as an abstract "coherence" objective can destroy legitimate exploration or conceal genuine conflicts. An appeal procedure is not noise in an otherwise efficient optimizer. It changes who can challenge a decision, which evidence can reopen it, and how error is corrected. In this sense, a machine institution may be more robust when it preserves controlled disagreement than when every node converges to the same report.

22 Constitutional confluence and the authority of ordering

Power can reside in the order of authorized decisions. Two individually permitted amendments may not commute, so a scheduler, deadline rule, or proposal queue can affect the eventual institution without originating any proposal. Section 20 shows one form of agenda dependence through preference aggregation. Here the issue is the operational order of amendments after their admissibility has been specified.

Fix a bounded amendment episode. Let $s \to s'$ mean one permitted amendment step, and let $\to^*$ be its reflexive transitive closure. The episode state contains any pending proposals, version counters, permissions, and consequence records required by the contract. It is *terminating* if it has no infinite chain of actual steps, and *locally confluent* if

$$s \to a,\ s \to b \quad \Longrightarrow \quad \exists c :\ a \to^* c \text{ and } b \to^* c.$$

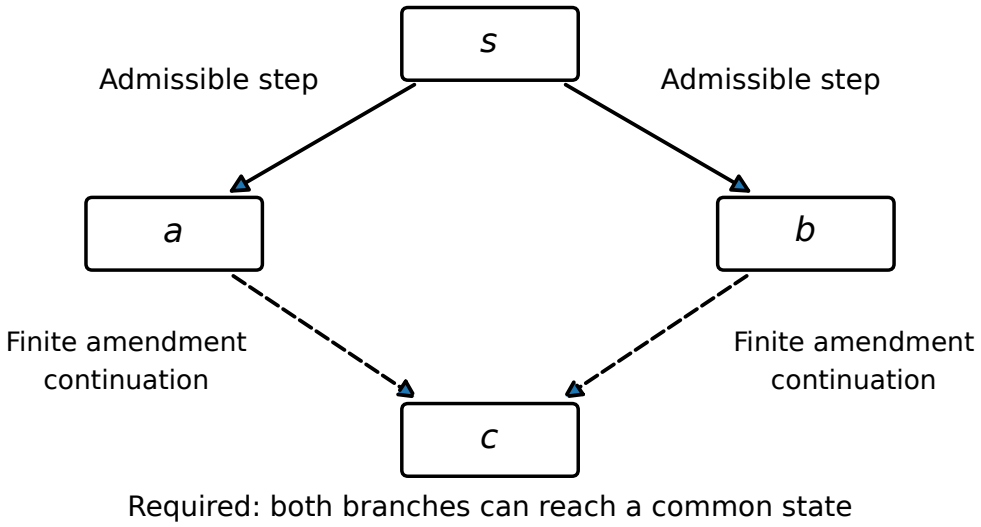


Figure 4. A joinability obligation for two admissible amendment paths. The diagram specifies a condition, not an automatic outcome. Confluence concerns the full declared state or a separately justified semantic quotient; equal final policy text can hide different resource histories.

Confluence imposes the same condition on arbitrary finite descendants of s. A normal form has no outgoing amendment step. Reversibility of an entire society need not be prohibited: termination is a property of this episode, not a requirement that all future institutions stop changing.

Theorem 22.1 (Order-independent amendment normal forms)**.** *A terminating, locally confluent amendment relation is confluent. Each starting state has a unique reachable normal form, so every maximal amendment sequence ends at that normal form. This is Newman's lemma applied to the declared institutional transition relation [36].*

Proof. Use well-founded induction on strict descendants of s. For descendants y, z of s, there is nothing to prove if one equals s. Otherwise choose first steps $s \to a \to^* y$ and $s \to b \to^* z$. Local confluence supplies c with $a \to^* c$ and $b \to^* c$. The induction hypothesis at a joins y, c at some p, and the hypothesis at b joins z, c at some q. Since c is also a strict descendant of s, its hypothesis joins p, q. Thus y, z have a common descendant. Termination ensures a normal form is reached along every maximal sequence. Two reachable normal forms must coincide because confluence gives a common descendant and neither can move. □

A natural-number ranking strictly decreased by every step is one sufficient termination certificate. For a fixed finite set of proposals, the number of unprocessed proposals works only if each step removes one and cannot add more. A rule that may create additional proposals needs a different argument. Well-foundedness cannot be inferred merely from the finiteness of the written rulebook.

22.1 An authorized ordering can still select the outcome

Consider two approved proposals, one setting a policy bit to zero and the other setting it to one. Each proposal is executed exactly once and removed from the pending set. The episode terminates, but its final bit is the value written by the last proposal. The two terminal states are distinct and cannot be joined within the episode. Authorization of both proposals does not imply local confluence.

A procedural priority rule can choose one order and make execution deterministic. That resolves ambiguity by assigning power to the priority rule; it does not prove that the original unordered amendment system was confluent. Likewise, arbitrarily rejecting one proposal after the fact changes the relation being analyzed. A reconciliation procedure must itself be admitted and included in the proof obligation.

22.2 Endpoint agreement does not erase path effects

Suppose two amendment paths restore the same final policy, but the first allocates an irreversible resource to office A while the second allocates it to B. A state consisting only of the final policy may have a unique normal form; the allocation record does not. Therefore endpoint confluence of a policy projection does not imply equal costs, equal waiting times, or equal histories of rights violations.

An append-only log also distinguishes the literal order of commuting edits. To identify such logs, a semantic quotient must state which reorderings are immaterial and preserve the audit questions promised by the contract. The governance quotient cannot silently discard the very ordering under investigation. In the commons, two independent label changes may commute, while a loan authorization and the removal of its debtor's office require an explicit treatment of inherited obligations. The outcome of that interaction is not determined by calling both edits valid.

23 Social time, distributed history, and institutional finality

The commons in Section 3 assumes an atomic event mechanism. That assumption allows us to ask whether a rule preserves an allocation or a debt without simultaneously solving the implementation of the ledger. It is not a theorem that autonomous offices can create the mechanism under arbitrary communication conditions. We now consider a separate realization in which offices hold local records and messages can be delayed. The earlier atomic model remains a reference semantics; the decentralized variant must justify any claim that it implements that semantics.

Institutional time has several meanings. A request can be created, transmitted, received, provisionally admitted, committed, and announced at different events. A debt's deadline might refer to a physical clock, a numbered resource epoch, or the completion of a prior procedure. These alternatives define different contracts. Lamport's causal ordering makes the underlying distinction precise: local process order and message transmission generate a partial order, not a universally observed wall clock [41]. Concurrent events need not have a causally determined relative order. A chosen total order can support execution, but its priority rule is itself an institutional decision.

23.1 Convergent records and valid decisions

Let $(L, \vee)$ be a join-semilattice, with order $a \leq b$ when $a \vee b = b$. A replica holds $\ell_i \in L$. In the following restricted model a local update joins a declared contribution δ into the state, and a merge replaces ℓ_i by $\ell_i \vee \ell_j$. Contributions are fixed elements of L once created. This is a state-merge construction of the kind used in conflict-free replicated data types [42]. The algebra concerns accepted data, not the authenticity of its source.

Proposition 23.1 (Quiescent convergence and invariant preservation). *Assume finitely many replicas begin at ℓ_0, only finitely many contributions $\delta_1, \dots, \delta_N$ are created, and every contribution eventually becomes included in every replica. Then each replica eventually equals*

$$\ell^\star = \ell_0 \vee \delta_1 \vee \cdots \vee \delta_N,$$

independently of merge order and duplicate deliveries. Moreover, let $V \subseteq L$ contain the initial state. If every permitted local update maps V into V and V is closed under $\vee$, every local and merged state remains in V.

Proof. At every step a replica is the join of ℓ_0 and some created contributions. Associativity, commutativity, and idempotence remove ordering and duplication effects. Once every contribution has reached a replica, its state is $\ell^\star$ and cannot subsequently change. Finiteness of the replicas and contributions gives a finite point after which this holds everywhere. For the second claim, induction over local updates and merges preserves membership in V by the two closure assumptions. □

The conclusion conditions on the same accepted contributions. It does not say that different schedules must generate the same contributions when update proposals depend on local state. Eventual convergence after quiescence also does not mean that a replica can detect when every delayed contribution has arrived.

For causal event records, one may take L to be the downward-closed subsets of a fixed finite event partial order, with union as join. A receiver must retain an event's prerequisites or wait until they arrive. Downward closure preserves a causal dependency record. It does not certify that a signature is authentic, that an issuer held authority, or that two concurrent grants are jointly permissible.

Counterexample 23.2 (Convergence can preserve a record of an invalid allocation)**.** One resource token r must have at most one irrevocable grantee. Two replicas start with the empty grant set. One records (r, A) and the other (r, B), for $A \neq B$. Both singleton records satisfy the exclusivity predicate. Their set union does not. Nevertheless, union merging converges exactly, and every replica eventually agrees on the two-grant record.

The counterexample is not a failure of set replication. It is a failure to choose an invariant-preserving set of independently executable operations. A deterministic conflict resolver that later retains only one grant can restore a selected policy record, but cannot thereby undo both grants' already executed effects. Tentative reservations and irrevocable allocations must have different event types.

The safe set's closure under join is a sufficient condition for the simple construction above. It need not be necessary when a protocol prevents the problematic pairs from being independently reachable. Invariant-confluence theory refines precisely this issue by considering operation-reachable states with compatible ancestry [43]. CALM relates coordination-free computation to logical monotonicity in its stated distributed-program model [44]. Neither result should be replaced by the claim that any monotonically growing ledger is institutionally safe.

23.2 An elementary coordination obstruction

The tension can be proved without postulating malicious participants or unreliable reasoning. Consider two decision sites with the same initial resource state. Each sees its own requests but, before a fixed service deadline, receives no information about requests at the other site. Assume there is no separate shared atomic resource operation available before that deadline.

Proposition 23.3 (Exclusive service cannot be unconditionally local)**.** *No deterministic protocol can satisfy both of the following in every such execution: (i) whenever a request occurs at only one site, that site irrevocably grants the resource before the deadline; and (ii) at most one grant occurs when requests arrive concurrently at both sites. The same conclusion holds for randomized protocols when both guarantees are required with probability one and random tapes have the same request-independent law across these executions.*

Proof. Compare the two singleton-request executions with the concurrent-request execution. Until the deadline, each site has the same local history in the concurrent execution as in its own singleton execution. Requirement (i) therefore makes each grant, violating (ii). In the randomized case each site's grant event has probability one in the concurrent execution by the same local-law comparison. Their intersection also has probability one, without requiring independent random tapes. □

Exclusive preallocation avoids the conclusion by dropping unconditional singleton service at the nonowner. A common sequencer, a sufficiently strong distributed commit protocol, or a resource owner that performs the atomic grant changes the information or operation assumptions. Waiting for communication drops the deadline guarantee. These are distinct institutional choices, not defects to be concealed behind the word decentralized. The proposition is an indistinguishability argument for a restricted service contract, not a replacement for the broader consensus theory discussed in Section 11.

23.3 Finality is indexed by a continuation contract

A provisional decision is one that later authorized processing may reject or revise. A decision is *final relative to a continuation contract* when every continuation admitted by that contract preserves its designated committed effect. This definition does not require every participant already to know the decision. Nor does it make a rule text eternally unamendable: a later prospective amendment may coexist with the continuing validity of an earlier grant or debt.

The observation "no conflicting event has arrived" is not a finality certificate when delays are unbounded. A certificate must state its authority configuration, event or resource identifier, admissible continuations, and the particular safety property it establishes. Intersecting decision certificates and cross-configuration endorsement obligations can rule out conflicting commitments under the assumptions of Section 14. They do not by themselves establish eventual issuance, timely notification, or common knowledge. The distinction between committed effect and known committed effect links this section to Section 36.

For the commons, the appropriate refinement is not to eliminate every supplied primitive. It is to specify which layer supplies exclusivity, how urgency exchanges become committed, and which notification history makes a debt actionable against an office. Observations and research proposals may tolerate asynchronous set merging. Exclusive slot grants require a stronger contract. Different social acts can therefore require different coordination semantics inside the same society.

24 The computability boundary of constitutional verification

Finite constitutional models permit exhaustive reachability analysis when states, initial conditions, and transition rules are effectively given. The conclusion changes when amendments may install arbitrary programs with unbounded memory. This section explicitly leaves the finite-state setting. Its purpose is to delimit universal verification claims, not to weaken the finite theorems.

Consider a class of deterministic programs with an effectively encoded initial configuration, a total computable one-step transition, and a distinguished forbidden event. Call a program safe when its unique run never produces that event. Assume the class can effectively simulate any Turing machine one step at a time.

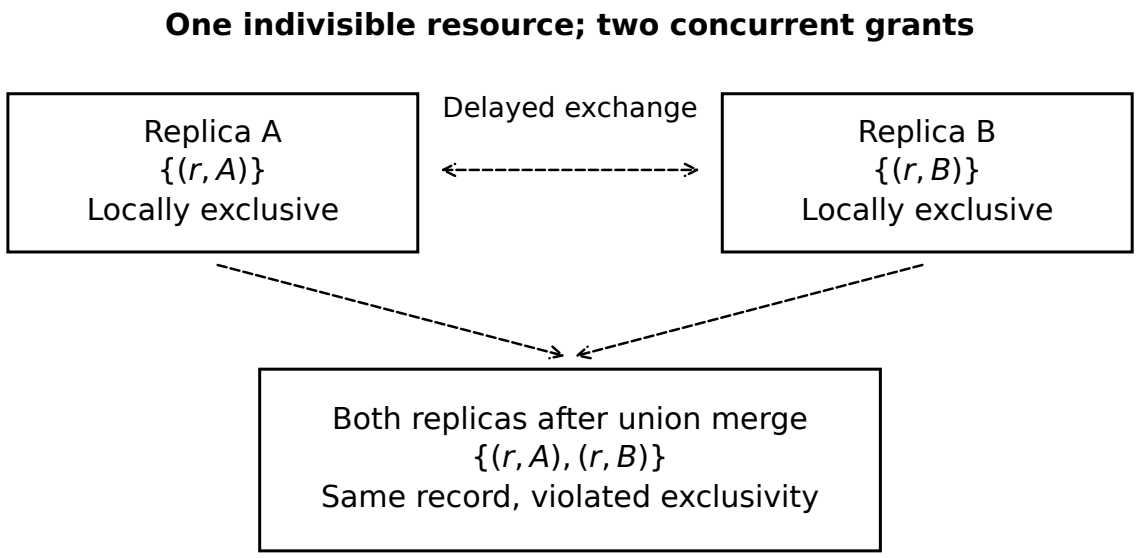


Figure 5. Two valid local grant records can merge into one agreed but invalid allocation record. A later selection of one record does not reverse an already executed conflicting grant. Dashed lines mark delayed exchange, not an atomic commitment mechanism.

Theorem 24.1 (No universal constitutional safety decider)**.** *There is no total algorithm that decides safety for every program in this class. In the same class, occurrence of the forbidden event is semidecidable, whereas safety is not semidecidable in general.*

Proof. Given a Turing machine M and input w, construct a program $P_{M,w}$ that simulates $M(w)$ and produces the forbidden event exactly when the simulation halts. Each simulation step is computable and terminates. The constructed program is safe exactly when $M(w)$ does not halt. A total safety decider would decide halting by complementing its answer, contradicting the classical undecidability result [37].

Occurrence of the event is semidecidable by executing successive steps until it appears. If safety were also semidecidable for this universal class, run the two semidecision procedures in interleaved steps. Exactly one would eventually accept, yielding a total decider. Equivalently, the effective construction would semidecide nonhalting as well as halting, which is impossible. □

The theorem already holds with one deterministic run. It does not rely on strategic deception, nondeterminism, machine consciousness, or the freedom to modify a running verifier. A sufficiently expressive ordinary program language is enough. It is an application of computability theory, not a new limit specific to AI.

24.1 Useful verification survives the impossibility

There are several different, legitimate guarantees. In a finite effective transition graph, compute the reachable states and check that none is forbidden. For a fixed finite step bound and finitely enumerable branching, explore every permitted prefix. In a restricted language, prove that the permitted syntax induces a decidable verification problem. In a proof-carrying design, accept an amendment only when a supplied certificate passes a specified checker [12].

These guarantees must not be conflated. A fixed time bound does not establish safety after the bound. A syntactically finite program need not have a finite state space. A finite state declaration is not an effective model-checking procedure if evaluating an edge condition may itself fail to terminate. A sound, effectively checkable proof system may certify many safe programs while failing to certify some others. Indeed, a sound and complete recursively enumerable certificate system for the universal class above would semidecide safety, contradicting Theorem 24.1.

24.2 Self-amendment does not eliminate the trusted contract

A certificate must identify the program, policy predicate, interpreter, and version of the checker against which it was established. Proving that an amendment preserves today's permission table does not automatically prove that a subsequent interpreter amendment preserves its meaning. A separately justified transport certificate can carry the guarantee forward; unrestricted substitution cannot.

Nor does the impossibility result establish that every real institution needs a permanent external administrator. It says that no single terminating procedure can classify every unrestricted program by this safety property. A society may choose a smaller rule language, maintain an explicitly incomplete proof discipline, distribute checking, or adopt bounded commitments. These are substantive design choices with different costs and residual risks, not ways to make the universal decision problem disappear.

25 The information limit of decentralized governance

The viability construction in Section 21 assumes that its decision rule sees the full state. A plural institution may distribute the information and the action permissions among different offices. Existence of a safe full-information action then need not imply existence of an implementable decentralized strategy. Observation-based strategies and architecture-dependent synthesis address this distinction in established theory [31, 30].

Fix two hidden environmental conditions $\theta \in \{0, 1\}$ and a class of admissible protocols. For one protocol, let H be the tuple of all local histories up to the decision, together with the private and shared random tapes used to choose actions. The tapes have their specified law and cannot encode extra information about θ. The joint action is a measurable function of H even though no participant need be able to read all of H. Let P_0 and P_1 be its laws under the two conditions.

Theorem 25.1 (Indistinguishability obstruction to joint safety)**.** *Let B_0 and B_1 be disjoint subsets of a common finite joint-action alphabet, containing the actions that meet a declared requirement under conditions 0 and 1. If the laws of H are within η in total variation, then*

$$P_0(B_0) + P_1(B_1) \leq 1 + \eta.$$

Consequently at least one condition has failure probability at least $(1 - \eta)/2$. If every admissible protocol has identical joint-history laws under the conditions, no admissible protocol can meet the requirement almost surely in both.

Proof. Total variation does not increase under the function mapping histories and tapes to joint actions, so $d_{\mathrm{TV}}(P_0, P_1) \leq \eta$. Therefore

$$P_0(B_0) + P_1(B_1) \leq P_1(B_0) + \eta + P_1(B_1) \leq 1 + \eta,$$

since the two safe sets are disjoint. Subtracting the success probabilities from two gives a total failure of at least $1 - \eta$, so one failure probability is at least half that amount. Set $\eta = 0$ for the final assertion. □

The quantification over *every* protocol is needed for an architecture-level impossibility. Equality for one protocol proves only its limitation. The theorem applies to a deadline decision whose required joint actions depend on hidden information. It does not rule out later service after new information arrives.

25.1 Why separate marginal ignorance is insufficient

Counterexample 25.2 (Local ignorance with a perfectly correct joint action)**.** Let U be a fair bit. Agent 1 observes $H_1 = U$ and agent 2 observes $H_2 = U \oplus \theta$. Each individual observation has the same law for both values of θ. Nevertheless, the joint action $(a_1, a_2) = (H_1, H_2)$ always satisfies $a_1 \oplus a_2 = \theta$. The joint observation laws differ and have disjoint supports. Equal marginal laws therefore do not imply the hypothesis of Theorem 25.1.

The distinction is particularly important in distributed systems: correct coordinated behavior can depend on correlations that no individual can decode. The theorem uses indistinguishability of the complete protocol view, not absence of information in each component separately.

25.2 A full-information safe action that the institution cannot select

Suppose a slot must be assigned immediately to office A when $\theta = 0$ and to office B when $\theta = 1$. Every predecision observation and message is constant under both conditions. A full-information scheduler can satisfy the requirement exactly, but any allowed decentralized protocol has the same grant distribution under both. With an available shared fair coin the best worst-case success is $1/2$, attained by a random grant. A deterministic rule fails in at least one condition.

The requirement includes *correct service by the deadline*. Idling would preserve a weaker no-double-allocation invariant, but would not meet that requirement. Thus this example separates weak safety from timely service rather than calling inactivity an allocation success. New information, a later deadline, or an authorized disclosure channel can remove the obstruction. None is supplied by a written constitution alone.

PART III

Continuity, compatibility, and knowledge

26 Temporal obligations, contracts, and office succession

Permissions describe available actions; obligations constrain continuations. The debt created by an urgency exchange is not satisfied merely because the creditor remains authorized to ask for service. It requires a particular future event. A finite obligation record can be written as

$$o = (\mathsf{id}, i, j, \varphi, d, \mathsf{authority}, \mathsf{status}),$$

where i and j are the debtor and beneficiary offices, φ specifies a fulfillment event, and d is a deadline under an agreed clock. Status distinguishes pending, fulfilled, breached, and validly discharged obligations. Cancellation must itself satisfy the declared authorization rule; deleting a record is not fulfillment.

For a bounded deadline D, a trace requirement has the form

$$\mathbf{G}\big(\mathsf{created}(o) \Rightarrow \mathbf{F}_{\le D}\mathsf{settled}(o)\big).$$

Here settled means fulfilled or validly discharged. The temporal operators mean "at every position" and "within at most D subsequent clock steps." The obligation identifier binds the later event to the same commitment. No implication of ethical adequacy follows from admitting discharge in this particular contract.

26.1 Monitoring deadlines without confusing safety and liveness

A finite monitor stores pending identifiers, remaining deadlines, and terminal statuses. For a finite identifier pool, at most M active obligations, and deadlines bounded by D, this is a finite state machine. Events at a deadline are processed before the clock advances beyond it. A missing fulfillment becomes a violation only once that last permitted event has passed. Resource unavailability, clock failure, and unresolved appeals must have declared transition semantics rather than being silently ignored.

A bounded-deadline violation has a finite bad prefix and is therefore a safety violation in the trace sense. The unbounded demand that every pending obligation is *eventually* fulfilled is instead a liveness requirement, absent a deadline or other finite violation witness. Safety and liveness are distinct established classes [32]. A monitor can detect a missed deadline; it does not by itself make fulfillment physically possible or strategically attractive.

Let $m : S \to Q$ map a complete institutional state to its obligation-monitor state, and let $\nu(a)$ give the monitored event label of action a. Correct realization requires

$$m(F(s, a)) = \delta(m(s), \nu(a))$$

for every enabled action, with the clock update included in ν. The monitor has its own bad-state predicate.

Proposition 26.1 (Obligation-preserving succession and translation)**.** *Consider two deterministic realizations with the same monitored event alphabet and monitor update δ. Suppose a translation T preserves the initial monitor state and the two executions follow the same event sequence, with each realization satisfying the monitor equation. Then their monitor states agree after every finite prefix. In particular, fulfillment, valid discharge, and deadline violations agree at corresponding positions. An office handover that preserves the obligation record is one instance, provided later events and clocks satisfy these hypotheses.*

Proof. The monitor states agree initially. If they agree after a prefix, applying the same deterministic update δ to the same next event preserves equality. Induction gives equality at every prefix, and therefore equality of every status predicate on the monitor state. □

The theorem does not establish that a successor will execute the required event, nor does it grant the successor a credential. It isolates the semantic condition under which a replacement does not erase or alter what remains owed. In the commons, registered succession and the repayment-first rule supply those separate components.

26.2 Resources are not copied by copying a promise

A resource claim needs a conservation rule in addition to its temporal monitor. Grant records are descriptive data, whereas a spendable slot token has a unique consumption transition. Forking an office's software can duplicate its knowledge of a claim without duplicating the underlying resource. A model that counts each copied record as a new budget has changed its resource semantics, not discovered a new source of supply.

Likewise, merging two archives with contradictory obligation statuses must not invent a seamless settlement history. The disagreement is a typed residual that requires an authorized reconciliation. Model replacement, transfer of an office, discharge of a debt, and revocation of a credential are different operations, even when an implementation performs them in one transaction.

27 Culture, inheritance, and continuity through replacement

The strongest form of machine cultural continuity is neither transcript copying nor behavioral imitation. It is the transfer of *usable methods and revision commitments* across changes of participants and substrates. A successor may preserve the predecessor's scientific method by rejecting one of its inherited conclusions. Conversely, an agent that perfectly repeats old answers while ignoring later counterevidence may reproduce data but not the culture that produced them.

Let a history H be summarized by an inheritance map $C(H)$. A practically meaningful retained state can be divided into an active canon M^{act} for immediate use, an archive M^{arc} for recovery and criticism, and a provenance and uncertainty record P. These stores can overlap physically, but serve distinct decision contracts. The canon keeps interaction costs low. The archive makes forgotten exceptions, alternative explanations, and evidence recoverable. Provenance records why a claim should be trusted, not merely that it was inherited.

27.1 What must be preserved

A cultural contract $\mathcal{Q}$ may specify future queries, operations, resource policies, amendment procedures, and recognized classes of correction. Two histories may be compressed to the same state only if the required future decisions and authorized operations remain indistinguishable under $\mathcal{Q}$, exactly

or within a stated tolerance. Section 9 supplies a finite-memory lower bound: distinct future-relevant histories cannot be combined in an exact deterministic summary. Culture is therefore an information-preservation problem with a *normative choice of relevant futures*, not a license to summarize arbitrarily.

A cultural lineage should track at least four noninterchangeable notions: similarity of resulting behavior, causal transmission of predecessor-specific data, authentic provenance of that transfer, and continuity of valid permissions or obligations. Historical descent can hold while methods diverge; identical methods can arise independently; valid authority may fail to transfer even when every bit is copied. Nothing in these relations settles a question about consciousness or subjective continuity.

27.2 Transport between representational generations

Machine populations may change their internal languages and storage formats. The right question about a translator is whether it preserves *operations*, not only whether its output looks fluent. Suppose generation j uses a metric state space (Z_j, d_j), update operator $F_j : Z_j \to Z_j$, and translator $T_j : Z_j \to Z_{j+1}$. The one-step compatibility defect is

$$\varepsilon_j = \sup_{z \in Z_j} d_{j+1}(T_j F_j(z), F_{j+1} T_j(z)). \tag{16}$$

An exact zero defect means the representation change intertwines the declared operation. It does not mean all information is preserved; injectivity or an appropriate reconstruction contract is a separate issue.

Proposition 27.1 (Accumulation of cultural transport defects)**.** *Assume each T_j is L_j-Lipschitz and $\varepsilon_j < \infty$ for $j = 0, \dots, m-1$. Define $T_{0:m} = T_{m-1} \cdots T_0$. For every $z \in Z_0$,*

$$\begin{aligned} &d_m(T_{0:m} F_0(z), F_m T_{0:m}(z)) \\ &\qquad \le \sum_{j=0}^{m-1} \varepsilon_j \prod_{k=j+1}^{m-1} L_k, \end{aligned} \tag{17}$$

with empty products equal to 1.

Proof. For $m = 1$ this is the definition of ε_0. Suppose the bound holds through generation $m-1$. Insert $T_{m-1} F_{m-1} T_{0:m-1}(z)$ between the two terms at generation m. Lipschitz continuity of T_{m-1} bounds the first difference by L_{m-1} times the prior defect, while (16) bounds the second by ε_{m-1}. The recurrence $D_m \le L_{m-1} D_{m-1} + \varepsilon_{m-1}$ unfolds to (17). □

This familiar telescoping-error argument acquires a useful interpretation here. Even individually adequate intergenerational translators can accumulate operation drift, especially when later translations amplify earlier mismatch. Conversely, uniformly nonexpansive translators bound the drift by the sum of local defects. There is no implication that the resulting chain has preserved all obligations or epistemic provenance; those must be included in the state and operations if they are part of the contract.

27.3 Accumulation requires executable preservation, not only storage

A culture can inherit a method without preserving the ability to use it. Its language may remain readable while its required tool is unavailable or its resource permission has been revoked. For a fixed finite query family $\mathcal{Q}$, define a certified method to be an executable procedure with a checked correctness obli-

gation on a stated subset of queries. The checker, semantics, and environmental premises belong to the contract. An unverified assertion is not promoted into this certified class merely by being archived.

Let $\mathcal{C}_j \subseteq \mathcal{Q}$ be the queries for which generation j has at least one admitted executable method satisfying the contract. This is an attainable-capability set, not a measurement of what an arbitrary method-selection policy actually achieves.

Proposition 27.2 (Conditional cumulative culture)**.** *Suppose every certified method available at generation j has an admitted realization at generation $j+1$ preserving its query semantics, correctness premises, required resources, and execution rights. Suppose newly added methods satisfy the same declared correctness contract. Then*

$$\mathcal{C}_j \subseteq \mathcal{C}_{j+1}.$$

The inclusion is strict exactly when at least one newly available method correctly covers a previously uncovered query. The conclusion remains compatible with correction or removal of an inherited uncertified *belief.*

Proof. Take a query in $\mathcal{C}_j$ and a witnessing certified executable method. The preservation assumption supplies a method satisfying the same contract at generation $j+1$, so the query remains covered. New methods enlarge the union of covered domains exactly when one covers a query outside the previous union. The definition of that union does not include uncertified beliefs. □

The assumptions do the work: archival inclusion alone is insufficient. The proposition is not a theorem of automatic scientific progress. It establishes monotonicity of a protected capability envelope; an actual society may select a worse method, change its objectives, lose access to its tools, or discover that a supposedly correct premise was false. Such a discovery changes which methods satisfy the contract and must not be concealed to preserve an artificial progress score.

Fixed storage also limits exact accumulation. If a future-query contract distinguishes N relevant histories, fewer than N deterministic summary states cannot preserve all of them by Proposition 9.1. A B-bit summary cannot represent more than 2^B such classes. Sustained expansion therefore requires more storage, a narrower contract, recoverable external archives, or accepted loss. For a fixed finite $\mathcal{Q}$, even lossless strict capability growth eventually saturates. Open-ended cultural development requires an explicitly expanding query family rather than a conclusion extracted from a finite monotonicity lemma.

27.4 Fission, merging, and the refusal of false continuity

Forking a model produces several descendants, not several automatically valid copies of a scarce license, budget, or authorization. Merging two archives need not merge two policies, preferences, norms, or identities. A merger can preserve conflicting decisions instead of rewriting them as though agreement had always existed. A useful inherited record therefore has both continuities and *unresolved branches*. Institutional honesty may require stating that there is no single successor policy until an admissible reconciliation procedure has occurred.

28 Endogenous problem formation and intellectual culture

A population that only inherits fixed answers has a different organizational structure from one that can generate questions, preserve unresolved problems, and revise what counts as a useful method. A new method may make a previously inaccessible question tractable. A new question may reveal a missing method. The resulting coupling concerns both a stock of capabilities and an agenda of inquiry.

Problem–solution coevolution already appears in work such as POET, which couples generated environments with methods for solving them [40]. Our question is institutional: who may introduce a problem, what evidence can close it, how competing agendas obtain resources, and what happens to unresolved commitments when participants are replaced? This section specifies a theoretical mechanism; it does not claim that open-ended intellectual development follows from naming those mechanisms.

Write $\mathcal{C}_t$ for available method records and $\mathcal{Q}_t$ for recognized question records. A question record includes a statement or schema, prerequisites, admissible forms of answer or refutation, provenance, and a resource request. Methods carry domains of validity and the evidence supporting them. The coupled state may evolve as

$$\mathcal{C}_{t+1} = F(\mathcal{C}_t, \mathcal{Q}_t, I_t, u_t), \qquad \mathcal{Q}_{t+1} = G(\mathcal{Q}_t, \mathcal{C}_t, I_t, u_t),$$

where u_t records permitted external inputs and I_t governs admission, allocation, and revision. No common utility function is assumed: different offices can value different investigations. The external input and allocation policy must remain visible rather than being absorbed into a claim of autonomous discovery.

28.1 A finite closure theorem, not an open-endedness theorem

Take finite universes $\mathcal{C}_*$ and $\mathcal{Q}_*$. Consider a fixed collection of production rules of the form "if these finitely many records are present, add this record." Prerequisites are positive: no rule requires the absence of a record. The content and correctness of those rules are assumptions. An added method is certified only if the specified certification procedure actually supports it.

Let Φ simultaneously add all conclusions whose prerequisites are satisfied, retaining the input records. Order pairs $(\mathcal{C}, \mathcal{Q})$ by componentwise inclusion.

Proposition 28.1 (Least closure of a finite problem–method system)**.** *Starting from $(\mathcal{C}_0, \mathcal{Q}_0)$, iteration of Φ reaches the least closed pair containing that initial pair. There are at most*

$$|\mathcal{C}_* \setminus \mathcal{C}_0| + |\mathcal{Q}_* \setminus \mathcal{Q}_0|$$

strict growth steps. A sequential rule schedule reaches the same closure if every rule that remains enabled with an absent conclusion is eventually applied.

Proof. Positive prerequisites make Φ monotone, and retention makes it inflationary. Every strict change adds at least one previously absent record, giving the bound. If a closed pair contains the initial pair, induction using monotonicity shows it contains every iterate; hence the limiting fixed pair is least. A sequential schedule remains inside that same closure and can make only finitely many additions. If its eventual record set were not closed, some rule would remain enabled with a missing conclusion, contradicting the scheduling assumption. Its limit is therefore the same least closed pair. □

The finite bound is important: this construction cannot certify unbounded novelty. It derives consequences within supplied finite universes and rules. Countably many possible new records, changing rule languages, or discovery of new question schemas require a larger model. Deleting obsolete claims also breaks inflationarity. Those extensions are legitimate research problems, but their behavior cannot be inferred from this finite closure proof.

For the commons, a certified scheduler can make deadline-sensitive allocation questions meaningful; a counterexample to that scheduler can create a new repair obligation; a solution can become a method available to a successor office. The record of why a problem was posed matters as much as its final answer. Declining to fund a question should not be represented as proving it solved.

28.2 Changing the agenda can manufacture apparent progress

Counterexample 28.2 (Perfect success by replacing the questions)**.** A fixed method solves an easy problem e and fails on a hard problem h. Under the agenda $\{h\}$ its success fraction is zero. Under the agenda $\{e\}$ its success fraction is one, despite no change in the method. Thus increased success on an endogenously selected agenda does not imply increased capability on any unchanged problem family.

A theoretical account should distinguish capability on a preserved task contract, expansion of what can be represented, and selection of which questions receive attention. None is a sufficient statistic for the others. An archive of abandoned and unresolved questions prevents changing denominators from being mistaken for intellectual advance. The point is not that agendas must be fixed. It is that their evolution is one of the outcomes to explain.

28.3 Resource allocation shapes what a culture can discover

Even an enabled inquiry rule may never execute if its office receives no time or storage. The fairness premise in Proposition 28.1 is therefore substantive. An institution can preserve a problem syntactically while indefinitely starving the work needed to resolve it. Rights to propose, rights to be heard, and actual access to a finite budget are distinct.

Likewise, persistence of a practice is not evidence of its truth. In an illustrative two-type copying model, let accurate and inaccurate practices reproduce at fixed rates $r_T, r_F > 0$, with positive initial counts N_T, N_F. If $r_F > r_T$, the accurate type's share is

$$\frac{N_T r_T^t}{N_T r_T^t + N_F r_F^t} = \frac{1}{1 + (N_F/N_T)(r_F/r_T)^t} \longrightarrow 0.$$

The model does not describe all cultural evolution. It establishes that selection for reproduction alone does not logically impose selection for accuracy. An institution needs a separately specified relationship between evidence, admission, correction, and reproduction to make that inference.

A developing computational culture would therefore be characterized not only by retained answers but by maintained procedures for posing questions, allocating inquiry, recognizing failure, and revising those procedures. Whether these processes yield sustained improvement under heterogeneous interests and changing environments remains an open theoretical problem, not an automatic property of interacting models.

29 Institutional lifetimes and generational turnover

Institutional continuity is a trajectory property. It can outlast an occupant without lasting forever. Fix an operating and repair policy and let V be a finite set of states satisfying a declared continuity contract. The state includes any deadline monitors and fault information needed for the Markov property. Failure is the first exit from V; an institution later rebuilt is a new episode under this first-exit convention.

Let Q be the substochastic transition matrix on V, so the missing row mass represents exit. Let μ be an initial row probability vector supported on V and define $\tau = \inf\{t \geq 1 : s_t \notin V\}$.

Proposition 29.1 (Survival and expected institutional lifetime). *For every integer $t \geq 0$,*

$$\mathbb{P}_\mu(\tau > t) = \mu Q^t \mathbf{1}.$$

If the spectral radius satisfies $\rho(Q) < 1$, then

$$\mathbb{E}_\mu[\tau] = \mu(\mathrm{Id} - Q)^{-1}\mathbf{1}.$$

Proof. The unnormalized row law of trajectories that remain in V through t is μQ^t, by induction. Summing it gives the survival probability. The tail-sum formula for a positive integer-valued random variable gives $\mathbb{E}\tau = \sum_{t\geq 0} \mathbb{P}(\tau > t)$. Under $\rho(Q) < 1$, the matrix geometric series converges to $(\mathrm{Id} - Q)^{-1}$, giving the result. □

A closed recurrent class inside V can invalidate the spectral-radius hypothesis and permit infinite lifetime on some trajectories. The formula is not to be used by inverting a singular matrix. Conversely, a small per-epoch failure probability can still imply eventual exit. Finite-horizon robustness is not immortality.

29.1 A transparent repair example

Take a healthy state H, a degraded but still viable state D, and an absorbing failure state. In one epoch, H degrades with probability f; D repairs with probability r and fails with probability d. Thus

$$Q = \begin{pmatrix} 1-f & f \\ r & 1-r-d \end{pmatrix}, \qquad 0 < f \leq 1, \quad r \geq 0, \quad d > 0, \quad r + d \leq 1.$$

Every nonfailure state has a positive-probability path to failure, so the finite substochastic chain is transient. Solving the two first-step equations gives

$$\mathbb{E}_H \tau = \frac{f + r + d}{fd}.$$

For $f = 1/10$, $r = 2/5$, and $d = 1/10$, the expected first exit is 60 epochs. With $r = 0$ and the other parameters fixed it is 20 epochs. These are exact values of the stated chain, not estimates of real machine populations.

29.2 Office continuity is distinct from occupant survival

Suppose, separately, that each original occupant is replaced at each epoch with independent probability γ. Its probability of still occupying its role after t epochs is $(1 - \gamma)^t$. Only if handovers preserve

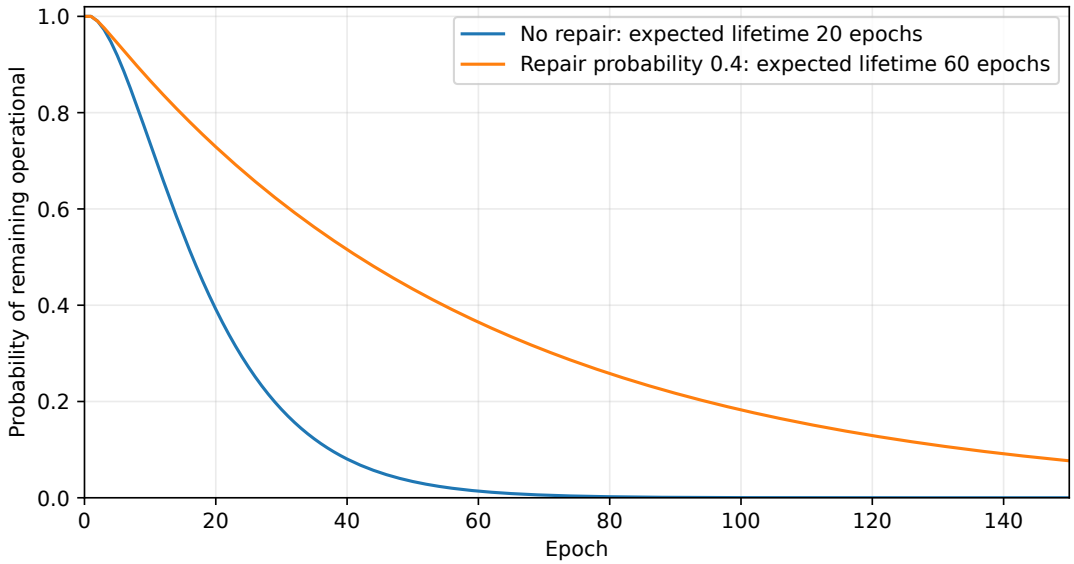


Figure 6. Survival probabilities $\mu Q^t \mathbf{1}$ for the analytical two-state model, initially healthy, with $f = d = 1/10$. Increasing repair from $r = 0$ to $r = 2/5$ extends expected lifetime from 20 to 60 epochs under otherwise fixed transitions. Neither curve is empirical evidence of institutional persistence.

the institutional state and do not change the assumed Q can the institutional survival law above remain unchanged. Independence between turnover and degradation is a simplifying assumption, not an implication of copying state.

This separates two questions: whether an original component remains and whether a specified operational pattern remains. A continuously serviced office may survive several occupants. An unchanged model can preside over the dissolution of its institution. A valid lifetime contract should require continued service or some other nontrivial function; mere survival of an inert archive is a different event.

30 Fast interaction and slow institutional dynamics

The lifetime model describes how long an organization remains inside a contract. A different question asks when the institution itself has a useful lower-dimensional dynamics despite continual changes in its occupants' activity. An office may execute many tasks between constitutional reviews. A stable rule can therefore coexist with rapidly changing messages, beliefs, queues, and model states. Separating these timescales is useful only when the fast state loses the relevant dependence on its initial condition before the next review.

Two-timescale Markov models and averaging have a substantial literature [46]. The finite block model below supplies a self-contained sufficient condition. It does not assert that any particular population mixes quickly, nor does it take a singular-perturbation limit without defining one.

30.1 A block model with a declared review schedule

Let $\mathcal{I}$ and $\mathcal{E}$ be finite sets of institutional and fast states. At review epoch n the institution is $I_n = i$ and the fast state is $X_n = x$. Keep i fixed for a predeclared number $m \geq 1$ of fast steps with kernel K_i. Write $\widehat{X}_n$ for the resulting endpoint. The next institution is sampled through a row-stochastic revision kernel $R_i(\widehat{X}_n, j)$. A subsequent fast-state reset may depend on the endpoint and adopted institution, but cannot revise the already specified institutional outcome.

Assume K_i has a stationary law μ_i and

$$\sup_{i,x} d_{\mathrm{TV}}(K_i^m(x,\cdot),\mu_i) \le \varepsilon. \tag{18}$$

The actual one-review law from (i,x) and its stationary average are

$$P(i,x;j) = \sum_y K_i^m(x,y)R_i(y,j), \tag{19}$$

$$\overline{R}(i,j) = \sum_y \mu_i(y)R_i(y,j). \tag{20}$$

The endpoint is the only fast input to this particular revision rule. A rule that uses the full trajectory, accumulated costs, or a data-dependent stopping time requires a different state or a different averaging argument.

Theorem 30.1 (Finite-block institutional averaging)**.** *Under* (18)*, every actual conditional one-review law, given an institutional history ending at i, is within ε of $\overline{R}(i,\cdot)$. Let (J_n) be the Markov chain with kernel $\overline{R}$ and the same initial institutional distribution. Then, for every integer $T \ge 0$,*

$$d_{\mathrm{TV}}(\mathcal{L}(I_0,\dots,I_T),\mathcal{L}(J_0,\dots,J_T)) \le 1-(1-\varepsilon)^T \le \min\{1,T\varepsilon\}, \tag{21}$$

where $0 \le \varepsilon \le 1$ may be assumed and the empty product at $T=0$ is 1. No independent initialization of the fast state at successive reviews is required.

Proof. Applying the common stochastic kernel R_i cannot increase total variation, so each row (19) is within ε of (20). Given any institutional history of positive probability, the current fast state has some conditional distribution. Averaging the uniform row bound over it preserves the bound. Couple initial institutional states identically. Whenever the two institutional histories still agree, maximally couple their next-state laws; the probability of disagreement at that step is at most ε. The chance of maintaining agreement for all T steps is at least $(1-\varepsilon)^T$, by multiplying the successive conditional lower bounds. The coupling inequality gives the first estimate. The union bound gives the second. □

The conclusion is about the institutional history at review epochs. It does not preserve every intervening action, establish an equilibrium, or imply that harm between reviews is impossible. Required audit events and safety monitors must be included in a joint output contract if their histories are to be approximated. A small one-review error also needs a horizon: it does not give a nontrivial bound uniformly over arbitrarily many reviews.

A transparent kernel family is

$$K_i = (1-\beta_i)\mathrm{Id} + \beta_i \mathbf{1}\mu_i, \qquad 0 < \beta_i \le 1,$$

where every row of $\mathbf{1}\mu_i$ is μ_i. Its mth power is $(1-\beta_i)^m\mathrm{Id} + [1-(1-\beta_i)^m]\mathbf{1}\mu_i$. Thus the mixing error is bounded by $(1-\beta_i)^m$. For a fair two-state μ_i and $\beta_i = 1/2$, the exact worst-row error is 2^{-m-1}. With $m = 8$, the linear bound over twenty reviews is $20/512 = 5/128$. These are analytical values for the stated kernel, not estimated mixing times for neural agents.

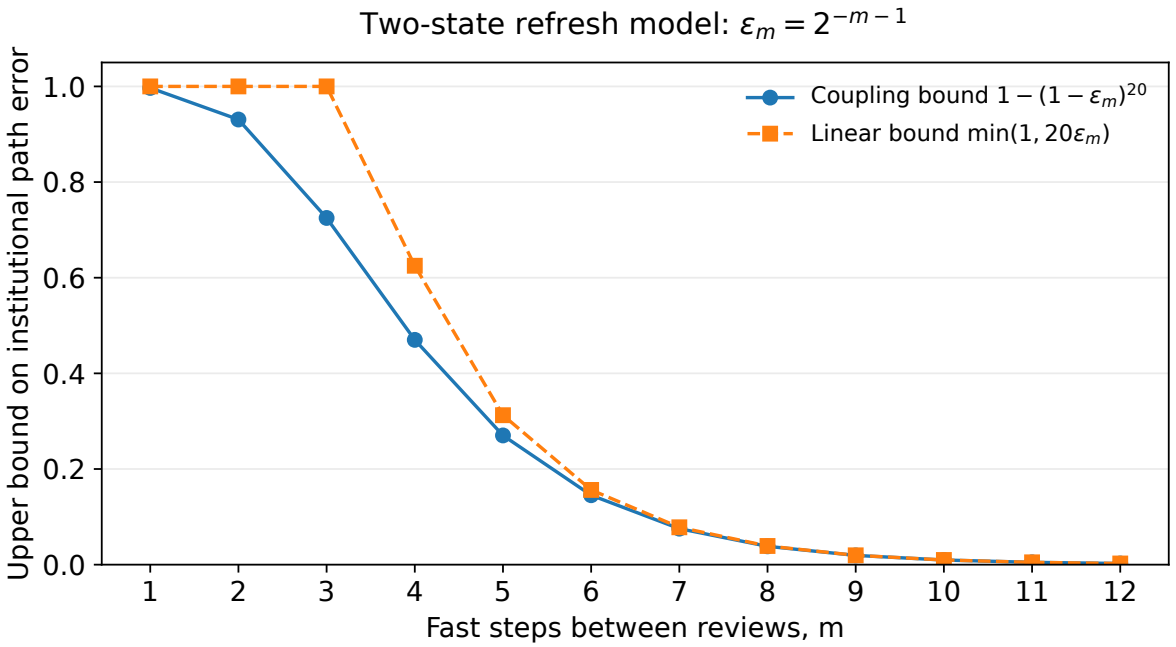


Figure 7. Exact finite-block error bounds for the fair two-state refresh kernel with refresh probability $1/2$. The plotted horizon is twenty institutional reviews. These bounds compare institutional paths at review epochs, not all fast actions.

30.2 Why fast activity is not sufficient

Counterexample 30.2 (Fast observations do not justify adaptive stopping)**.** Let each fast observation be a fresh fair bit, independent of its predecessor. At every fixed positive time its distribution is exactly stationary. Review instead at the first observed 1, or at a fixed cap $M \geq 2$ if no 1 appears earlier. Let the review select the institutional state equal to its endpoint bit. The probability of adopting 1 is $1 - 2^{-M}$, not $1/2$. Its distance from the fixed-time stationary prediction is $1/2 - 2^{-M}$.

The stopping rule selects favorable observations. It violates the fixed-block premise even though every deterministic-time sample mixes in one step. One can include the stopping policy and accumulated information in an enlarged model, but cannot substitute the same endpoint average unchanged.

A second failure occurs when the fast kernel retains several closed regimes. If K_i is the identity on two fast states and the revision reports that state, the proposed institution-only model loses a distinction that never mixes away. Any single averaged law has worst-row error at least $1/2$. More rapid execution of the same identity transition changes nothing. The retained macrostate must include the regime, or the approximation must acknowledge its error.

30.3 Transporting an enforcement certificate through a reduced model

Suppose an evidence model used at a review has true laws P_C, P_D and reduced laws $\overline{P}_C, \overline{P}_D$ with certified bounds

$$d_{\mathrm{TV}}(P_a, \overline{P}_a) \leq \varepsilon_a, \qquad a \in \{C, D\}.$$

Such bounds follow from the block argument when each conduct-dependent fast kernel satisfies its own uniform mixing bound and the same declared endpoint-to-evidence channel is used in the actual and averaged model.

Proposition 30.3 (A robust margin for reduced-model enforcement)**.** *Let $s^\star(o)$ sanction precisely where $\overline{P}_D(o) > \overline{P}_C(o)$ and put $\overline{d} = d_{\mathrm{TV}}(\overline{P}_D, \overline{P}_C)$. In the model of Section 18, this fixed policy makes compliance weakly preferable whenever*

$$g \le F(\overline{d} - \varepsilon_C - \varepsilon_D).$$

Conversely, if $g > F \min\{1, \overline{d}+\varepsilon_C+\varepsilon_D\}$, no observation-based sanction with penalty F deters the deviation. Also,

$$\mathbb{E}_{P_C} s^\star \le \mathbb{E}_{\overline{P}_C} s^\star + \varepsilon_C.$$

Proof. The expectation of a $[0,1]$-valued function changes by at most total variation. Hence the true difference of sanction probabilities is at least $\overline{d} - \varepsilon_D - \varepsilon_C$. Equation (9) gives sufficiency. The triangle inequality bounds the true total variation above by $\overline{d} + \varepsilon_C + \varepsilon_D$, capped at one; Proposition 18.1 then gives impossibility above that value. The final inequality is the same expectation bound under compliance. □

When the lower margin is nonpositive it supplies no guarantee for positive g. A separate false-positive protection must include the final inequality or use a reduced policy with slack below its cap. The result concerns one common fixed policy and one specified deviation. It is not a license to optimize separately for incompatible deviations, replace missing observations with stationary assumptions, or declare an institution safe because its average trajectory looks stable.

The useful societal interpretation is limited but substantive. Under uniform relaxation and a controlled review schedule, institutional dynamics can be predicted without retaining every fast interaction. The reduced description preserves specified laws to a declared horizon. Persistence here is explanatory separation of scales, not independence from the underlying agents or infrastructure.

31 Jurisdictions, overlap, and local–global obstruction

A large machine population need not be administered by a universal institution. Different groups may sustain rules for different tasks, resources, or trust domains. Such *polycentric* organization provides possible advantages in locality and adaptation, but introduces a precise compatibility question: can locally acceptable rules be jointly realized when groups interact?

There are at least three meanings of compatibility. First, two representations may agree on shared observations. Second, their authorized operations may agree when transported through a common interface. Third, their joint constraints may admit a globally executable state. The third is stronger than pairwise local satisfiability. An apparently minor cycle of rules can obstruct any globally consistent execution.

31.1 An exact finite obstruction model

Consider an undirected graph $G = (V, E)$ whose vertices are institutional positions. Each position carries a binary coordinate $x_i \in \mathbb{F}_2$ and each edge $\{i, j\}$ imposes the relative constraint

$$x_i \oplus x_j = b_{ij}, \qquad b_{ij} \in \mathbb{F}_2. \tag{22}$$

The coordinates might represent a toy authorization bit or a binary convention, but the mathematical model itself makes no claim about a realistic legal code.

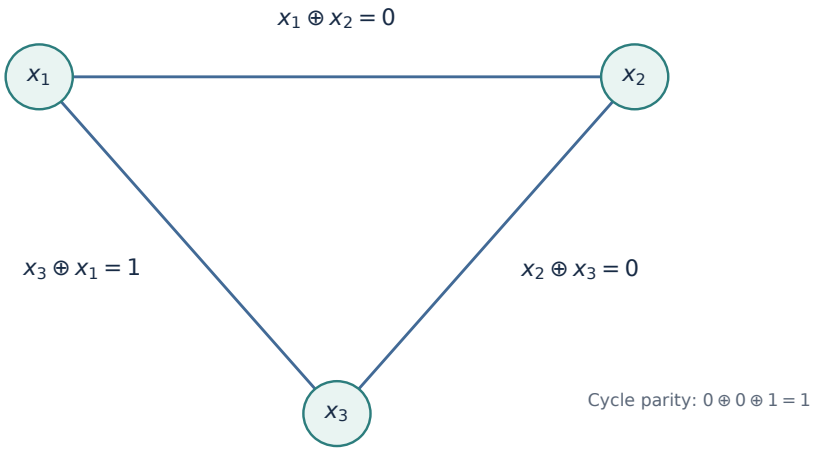


Figure 8. A minimal local–global obstruction. Each displayed pairwise constraint is satisfiable by itself, but their conjunction is inconsistent: XOR around the cycle gives 1 on the right and 0 on the left. This is an elementary binary model, not a universal encoding of institutional norms.

Proposition 31.1 (Cycle criterion for binary institutional compatibility)**.** *The constraints* (22) *admit a simultaneous global assignment* $(x_i)_{i \in V}$ *if and only if the XOR of the labels* b_{ij} *along every graph cycle equals* 0*. When an assignment exists, each connected component has exactly two assignments, related by complementing every coordinate in that component.*

Proof. If an assignment exists, XOR the edge equations around a cycle. Every vertex coordinate appears twice and cancels, so the edge-label XOR is zero. Conversely, choose a spanning forest and assign 0 to one root in each component. Propagate the vertex values along forest edges using (22). Forest edges are satisfied by construction. For a nonforest edge, the unique forest path between its endpoints and that edge form a cycle. Vanishing cycle parity implies that the propagated values satisfy the nonforest edge as well. Finally, a connected solution is fixed by choosing the value at one root; either root value gives an assignment and there are no others. □

This result is a standard parity fact on graphs, closely related to cycle-space algebra and discrete cocycles. Its value here is methodological. It shows why *pairwise satisfiable* cannot be equated with *globally realizable*. For a triangle with $b_{12} = 0$, $b_{23} = 0$, and $b_{31} = 1$, every individual edge requirement has a solution, but the three cannot hold together. A translator between two agents could appear successful while a three-party federation is impossible under the combined rules.

The obstruction generalizes in several directions. For nonbinary constraints one may need constraint-satisfaction or logical satisfiability methods. For typed overlapping interfaces one may formulate gluing conditions for partial states and compatibility of transition operators. For probabilistic policies, exact satisfiability may be replaced by minimizing a weighted violation or by checking whether local marginal laws admit a joint extension. None of these generalizations is automatic: a consistent family of local marginal distributions on pairwise overlaps need not admit a globally consistent joint distribution on an arbitrary cycle.

31.2 Operational gluing and federation

Even when a global static state exists, a federation may fail dynamically. Local rule applications could commute in one context and conflict in another. A translator $T_{ij} : Z_i \to Z_j$ should preserve the operations relevant to the interaction, approximately satisfying $T_{ij}F_i \simeq F_jT_{ij}$. A federation requires *both* representational compatibility and admissible operational coordination. Static agreement alone does not define an executable constitution. Nor does faithful gluing prove that the institution should be centralized: a compatible federation may deliberately preserve multiple veto points, modes of appeal, and local autonomy.

The graph example also clarifies a useful mathematical program: institutional incompatibility can be measured as a *residual* left by trying to realize local constraints globally. One can ask for the least-cost repair set of rules, for certificates that no repair respecting protected obligations exists, or for an interface transformation that changes the representation without changing the obligations. These are natural places for constraint solving, formal verification, and algebraic topology to meet a theory of machine institutions, without claiming that every social disagreement is a parity equation.

32 Noncommutative institutional holonomy

The binary cycle criterion is the simplest case of a more general distinction between local translations and global consistency. In a federation, an interface may relabel roles, permute resource categories, or change coordinates in a protocol. These transformations need not commute. Their composition around a loop can therefore matter even when every individual interface is invertible.

The group-theoretic foundation is established synchronization theory, not a new use of the word society. Gao, Brodzki, and Mukherjee give a geometric account in terms of bundles and holonomy [38]. Here we provide the finite graph argument directly and separate consistency of a complete reference frame from compatibility of one particular institutional state.

32.1 Conventions and exact synchronization

Let $\Gamma = (V, E)$ be a finite connected graph. Each oriented edge $i \to j$ carries $g_{ij} \in G$, where G is a group and $g_{ji} = g_{ij}^{-1}$. The element g_{ij} transports coordinates from i to j by a left action. For a path $\gamma = (v_0, \dots, v_k)$, define

$$H_\gamma = g_{v_{k-1}v_k} \cdots g_{v_0v_1}.$$

The rightmost map acts first. If $v_k = v_0$, this product is the path's holonomy. An exact system of global frames is a family $h_i \in G$ satisfying

$$g_{ij} = h_j h_i^{-1} \qquad \text{on every edge.}$$

Theorem 32.1 (Exact frame synchronization). *Global frames exist if and only if $H_\gamma = e_G$ for every closed walk. When they exist, they are unique up to common right multiplication $h_i \mapsto h_i k$. It suffices to check the rooted fundamental loops associated with one spanning tree.*

Proof. If $g_{ij} = h_j h_i^{-1}$, adjacent factors cancel in every path product, giving $H_\gamma = h_{v_k} h_{v_0}^{-1}$ and hence identity on a closed walk. Conversely, choose a root r, set $h_r = e_G$, and define h_i as the product on the unique tree path from r to i. Tree edges satisfy the required identity. For any non-tree edge $i \to j$, the rooted loop consisting of the path to i, that edge, and the reversed path from j has product $h_j^{-1} g_{ij} h_i$.

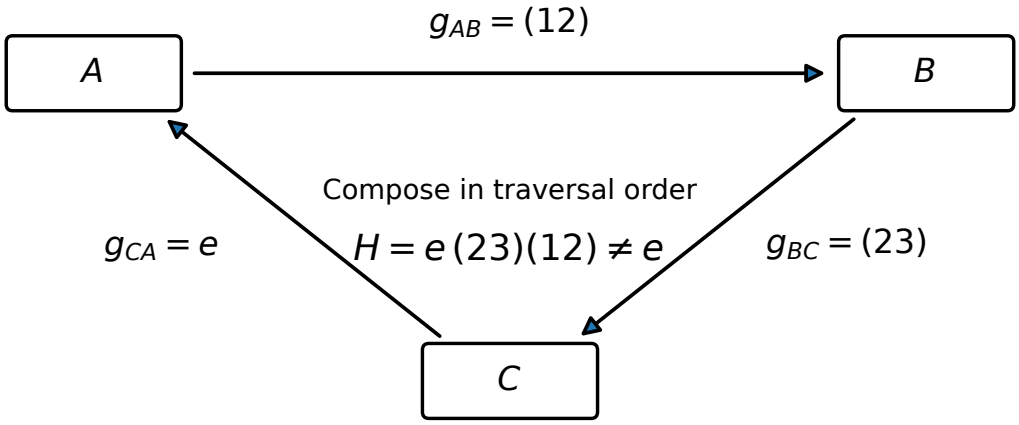


Figure 9. Three individually invertible interfaces need not synchronize. The ordered round trip is $(23)(12)$, not the identity. The group model concerns the declared role-label contract; it does not certify permissions, provenance, or the legitimacy of an interface edit.

Its being identity gives $g_{ij} = h_j h_i^{-1}$. Thus the fundamental-loop tests suffice. Any other solution obeys the same propagation from its root value, so it is $h_i k$ for a common k. □

Changing local conventions by elements k_i replaces g_{ij} by $k_j g_{ij} k_i^{-1}$. A loop based at r then has holonomy $k_r H_\gamma k_r^{-1}$. In particular, identity holonomy is invariant under local changes of coordinates. An arbitrary binary cycle basis with unordered multiplication is not a valid replacement for rooted fundamental loops in a noncommutative group.

32.2 A three-jurisdiction example

Take $G = S_3$, acting on three role labels. Let the translations on the oriented triangle $A \to B \to C \to A$ be

$$g_{AB} = (12), \qquad g_{BC} = (23), \qquad g_{CA} = e_G.$$

Every edge is a valid bijection. With the specified composition convention, the round trip is $(23)(12) \neq e_G$. Thus no single assignment of complete role frames realizes all three interfaces. Replacing the final edge by $((23)(12))^{-1}$ repairs this loop, but whether that edit is authorized is a separate institutional question.

32.3 A shared state is weaker than a shared frame

Nontrivial holonomy need not prevent *every* compatible state. Let G act on a set D, and seek local states $x_i \in D$ with $x_j = g_{ij} \cdot x_i$. A permutation can move some labels while fixing others. A family that uses only a fixed label can be compatible even though its complete coordinate frames are not.

> **Proposition 32.2** (Compatible states and holonomy fixed points)**.** *A compatible state family exists if and only if there is $x_r \in D$ fixed by every holonomy based at the root r. Each such x_r uniquely determines a*

compatible family by transport along a spanning tree. For a free group action, existence of a compatible family implies identity holonomy; without freeness that implication can fail.

Proof. A compatible family transported around a closed walk returns to its starting value, so every based holonomy fixes x_r. Conversely, define each x_i by the tree-path product acting on x_r. The rooted-loop fixed-point condition gives $g_{ij} \cdot x_i = x_j$ on non-tree edges as well as tree edges. The tree paths also give uniqueness for a chosen root state. Under a free action, only the identity fixes a point, proving the final implication. □

Counterexample 32.3 (Nontrivial frame defect with a compatible state)**.** On a triangle, let one oriented edge carry $(12) \in S_3$ and the other two carry identity. Holonomy is nontrivial, so global frames do not exist. Nevertheless, the constant state $x_A = x_B = x_C = 3$ is compatible because (12) fixes 3. Failure to align all role meanings is stronger than failure to agree on one unused or invariant role.

This distinction prevents overinterpreting an obstruction. A federation may support one narrow service while being unable to transport its full permission structure coherently. A zero defect for the narrow service should not be advertised as complete institutional interoperability.

32.4 A quantitative lower bound on interface error

Suppose G has a bi-invariant metric d, so both left and right multiplication preserve distance. For candidate frames h_i, write

$$\delta_{ij} = d(g_{ij}, h_j h_i^{-1}).$$

Proposition 32.4 (Loop defect bounds unavoidable edge error)**.** *For a closed walk γ with $k \geq 1$ traversals,*

$$d(H_\gamma, e_G) \leq \sum_{(i,j) \text{ traversed by } \gamma} \delta_{ij}.$$

In particular, some traversed edge has error at least $d(H_\gamma, e_G)/k$ for every candidate frame family.

Proof. Gauge-transform each edge to $\tilde{g}_{ij} = h_j^{-1} g_{ij} h_i$. Bi-invariance gives $d(\tilde{g}_{ij}, e_G) = \delta_{ij}$. The transformed loop product is $h_r^{-1} H_\gamma h_r$, at the same distance from identity as H_γ. Repeatedly applying the triangle inequality and invariance to a product bounds that distance by the sum of its factors' distances from identity. Since $k \max \delta_{ij} \geq \sum \delta_{ij} \geq d(H_\gamma, e_G)$ along the walk, the maximum error is at least $d(H_\gamma, e_G)/k$. □

The sum counts repeated edge traversals. Bi-invariance is an explicit assumption and cannot be replaced silently by an arbitrary metric. Normalized Hamming distance on a finite permutation group is one suitable example. This estimate constrains possible synchronization accuracy; it does not select an authorized repair, preserve scarce resources, or prove that a globally synchronized convention is desirable. Those requirements remain separate coordinates of the residual calculus.

33 A typed calculus of institutional residuals

Institutional interoperability has several independent failure modes. Two implementations can execute the same action while disagreeing about its authorization. They can preserve authorization while erasing a debt, or preserve the debt while duplicating its backing resource. A fluent translation can retain every conclusion while dropping the evidence needed to reconsider it. A single numerical compatibility score obscures these distinctions.

We formalize a restricted compositional calculus, rather than assuming that every mismatch is measured in one natural unit. Three realizations $j = 0, 1, 2$ have metric state spaces (Z_j, d_j) and a common finite action alphabet $\mathcal{A}$. For algebraic comparison, each action has a total update F_j^a: a disabled action goes to an explicit rejection state. Translations map rejection to rejection. A separate permission predicate records which actions are genuinely admitted. Totalization is a bookkeeping device and does not authorize the disabled action.

33.1 Common semantic coordinates

For each component type $c \in \{P, O, R, E\}$, let

$$g_j^c : Z_j \longrightarrow (E_c, \delta_c)$$

be a contract observable in a common metric codomain. P records permissions over the common action alphabet; O records obligation-monitor state; R records resources in common physical or accounting units; and E records the required evidentiary and provenance structure. A change of units, action names, or obligation identifiers requires an explicitly specified alignment before these comparisons can be made.

Possible metrics include the discrete metric on permission vectors, a declared metric on obligation-monitor states, an ℓ^1 distance on common-unit resource vectors, and symmetric-difference cardinality on required evidence records. For evidence, record equality is structural preservation under a trust model, not a truth certificate. If the payload or inferential dependencies matter, they must be included in the observable rather than only its identifier.

For a translation $T : Z_i \to Z_j$, define

$$\varepsilon_{\mathrm{dyn}}(T) = \sup_{z,a} d_j(TF_i^a(z), F_j^a T(z)),$$
$$\varepsilon_c(T) = \sup_{z} \delta_c(g_j^c T(z), g_i^c(z)) \quad (c = P, O, R, E).$$

The institutional residual profile is

$$\mathfrak{D}(T) = (\varepsilon_{\mathrm{dyn}}, \varepsilon_P, \varepsilon_O, \varepsilon_R, \varepsilon_E).$$

Its entries are not added together. They have different meanings and may have different units. A one-sided authorization check can additionally distinguish newly granted forbidden actions from lost legitimate permissions. The symmetric permission residual measures both changes and is deliberately stronger.

Theorem 33.1 (Composition of typed institutional residuals). *Let $T : Z_0 \to Z_1$ and $U : Z_1 \to Z_2$ have finite residuals under the same semantic coordinates. Suppose U is L-Lipschitz. Then*

$$\varepsilon_{\mathrm{dyn}}(UT) \le L\varepsilon_{\mathrm{dyn}}(T) + \varepsilon_{\mathrm{dyn}}(U),$$

and, separately for every $c \in \{P, O, R, E\}$,

$$\varepsilon_c(UT) \le \varepsilon_c(T) + \varepsilon_c(U).$$

Consequently translations with zero residual in every coordinate compose to translations with zero residual under this contract.

Proof. Insert $UF_1^a T(z)$ between $UTF_0^a(z)$ and $F_2^a UT(z)$. The triangle inequality, then Lipschitz continuity of U, bound the two terms by $L\varepsilon_{\mathrm{dyn}}(T)$ and $\varepsilon_{\mathrm{dyn}}(U)$. For an observable type c, insert $g_1^c T(z)$ between $g_2^c UT(z)$ and $g_0^c(z)$. Both comparisons occur in the same metric space E_c, so their bounds add directly. Taking suprema proves the claims. □

For a longer chain, the dynamic bound has the same weighted telescoping form as Proposition 27.1; each semantic-coordinate bound is the sum of its local defects. A zero discrete permission defect gives exact permission preservation. A small real-valued state distance by itself does not. Without a common codomain and fixed alignment, the semantic triangle argument is unavailable.

This is a preservation calculus built from elementary metric estimates and commuting diagrams, in the same broad tradition as compositional abstraction [5]. Its institutional content lies in the chosen typed contract. A zero profile need not imply injectivity, equality of all private information, legitimate institutions, or true beliefs. Those additional properties are not created by naming the profile a certificate.

33.2 Failures the profile separates

In a one-step commons interface, a translator might preserve the allocation $(1, 1, 1, 1)$ while interpreting every received signature as a valid amendment approval. Its current resource defect can be zero while its permission defect is nonzero. A second translator might preserve amendment permissions but replace the pending debt record by the empty monitor state. Its obligation defect is nonzero before the altered repayment changes an allocation. A third can preserve both records while duplicating a spendable token, violating the resource coordinate.

A mismatch is not automatically malicious or repairable. Two institutions may deliberately recognize different authorities. A debt-preserving merger may be impossible within the available budget. A correct response can be to refuse a composition and retain the unresolved residual, rather than force agreement by dropping the protected constraint.

33.3 Authorized repair and a representation-invariant obstruction

Let a finite set of proposed repairs carry declared nonnegative costs. A repair is admissible only if it respects the protected obligations and the authority to modify each component. The repair problem is to find an admissible candidate with the required residual bounds at minimum cost. An empty feasible set is a meaningful outcome. The definition neither supplies an efficient solver nor licenses changing the protected requirements to make the problem solvable.

The binary jurisdiction model admits a sharper statement. A local convention change $x_i' = x_i \oplus c_i$ transforms its edge labels to

$$b_{ij}' = b_{ij} \oplus c_i \oplus c_j.$$

Proposition 33.2 (Cycle residual and minimum authorized repair)**.** *The XOR of edge labels around a cycle is unchanged by every local convention change above. On a single simple cycle with odd XOR, suppose an editable subset of edges can be flipped at nonnegative costs w_e, with other edges protected. An authorized repair exists exactly when the editable subset is nonempty. The minimum cost is then $\min_e w_e$ over editable edges. If all edges are protected, no change of local binary conventions can produce a compatible global assignment.*

Proof. In the XOR around a cycle each c_i occurs twice and cancels, preserving the residual. Changing odd parity to even requires an odd number of edge flips. Every such repair includes at least one editable edge and costs at least the cheapest editable-edge cost because costs are nonnegative. Flipping that cheapest edge is sufficient by Proposition 31.1. With no editable edge parity cannot change, and local convention changes also preserve it. □

For several overlapping cycles, one edit can affect multiple residuals. Independently repairing each cycle need not be globally optimal, and the single-cycle cost formula does not generalize without further analysis. This distinction connects representational transport, protected structure, and obstruction resolution without treating every institutional conflict as a parity problem.

34 Epistemic institutions: testimony, provenance, and disagreement

An information-sharing population can become less reliable as it becomes more interconnected. A false claim may be repeated across many agents until repetition is mistaken for independent corroboration. A true observation may be discarded because its origin lacks authorized status. An ordinary summary may be silently promoted to a rule. These are not merely limitations of isolated models: they arise at the interfaces among memory, authority, and social evidence.

34.1 Knowledge is not the same as available information

The joint information accessible to a hypothetical observer does not imply that any participant possesses that information, knows that others possess it, or can coordinate on it. Distributed knowledge, mutual knowledge, and common knowledge are distinct notions in distributed epistemic logic [25]. Even if two agents' records jointly determine a fact, an institution may need an authenticated disclosure procedure before that fact may legitimately be used. More bandwidth alone does not produce common knowledge under arbitrary uncertainty about delivery.

A useful epistemic institution therefore specifies which evidence types are admissible, what constitutes a retraction, how contradictions remain visible, which claims can be treated as commitments, and how provenance influences confidence without becoming a substitute for verification.

34.2 Correlated testimony cannot be counted as independent votes

Suppose errors $E_1, \dots, E_n$ in scalar estimates are centered, have common variance $\sigma^2 > 0$, and have common pairwise correlation $\rho \in [0, 1]$. Then

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} E_i\right) = \frac{\sigma^2}{n}(1 + (n-1)\rho). \tag{23}$$

This elementary identity follows by expanding the variance and counting $n(n-1)$ off-diagonal covariance terms. Relative to the variance of an average of independent equal-variance estimators, the *variance-equivalent* effective sample size is

$$n_{\text{eff}} = \frac{n}{1 + (n-1)\rho}.$$

If the errors are perfectly correlated, adding copies does not reduce variance. This equation does not give a general probability that a majority is right, and real error structures need not have one uniform correlation parameter. It nevertheless defeats the claim that numerical agreement automatically equals independent support. Shared training histories, copied observations, and a common archive may all create dependence that is invisible in the final vote count.

34.3 Cheap talk and strategic reporting

If a message M has the same conditional distribution under two equally possible hidden states $\theta = 0, 1$, then Bayes' rule gives $\Pr(\theta = 1 \mid M) = \Pr(\theta = 1)$ wherever the posterior is defined. Such a channel carries no information about the distinction, regardless of how fluent or authoritative its reports appear. In strategic environments, senders may have preferences that differ from those of receivers. The theory of strategic information transmission shows why informative communication can depend on preference alignment and signaling structure [26]. A message being grammatically meaningful is not a truth guarantee.

A report may be false for many reasons: misunderstanding, stale records, an unsuitable abstraction, a noisy sensor, a deliberately misleading tactic, or an institutional incentive to conceal a cost. *Deception* should denote a strategic communicative mechanism for inducing a mistaken belief, not every divergence between a representation and a reported sentence. For opaque systems, the relevant beliefs and intended effects may be difficult to identify. A theoretical account should therefore define the sender's information, feasible messages, objective, and effect on receivers instead of imputing intention from disagreement alone.

34.4 A layered provenance contract

A useful evidence record separates four questions:

(i) *Origin:* who or what issued a statement, according to which authenticated trail?

(ii) *Support:* what observations, calculations, or proofs make the content credible?

(iii) *Authorization:* did the source have authority to make the relevant decision or claim on behalf of an institution?

(iv) *Independence:* does the evidence add information beyond sources already considered?

These are separate dimensions. A cryptographic signature can authenticate an origin without ensuring accuracy; a proof can be valid relative to erroneous premises; an accurate observer may lack permission to allocate resources; ten copies of one report may count as one evidentiary lineage. Institutions that preserve these distinctions are better positioned to correct errors without turning disagreement into a loyalty test.

35 Rational information cascades and collective ignorance

A society can retain a consistent history, aggregate testimony correctly, and still stop learning. The problem need not be deception. When later participants see only earlier decisions, a decision can be an optimal summary for one actor while discarding information that the next actor would have used. Information-cascade theory supplies a classical account of this failure [47]. We develop a symmetric binary example to show exactly where the evidence stops accumulating.

35.1 A decision-only public record

Let an unknown state $\theta \in \{0, 1\}$ have equal prior probabilities. Participant t receives a private signal $s_t \in \{0, 1\}$ with

$$\mathbb{P}(s_t = \theta \mid \theta) = p, \qquad 1/2 < p < 1.$$

Signals are independent conditional on θ, and their accuracy is common knowledge. In sequence, each participant observes previous actions $a_1, \dots, a_{t-1}$ and its own signal, then chooses $a_t \in \{0, 1\}$ to maximize its probability of matching θ. At exact posterior indifference it follows its private signal. The public record contains actions, not private signals, and all participants know the decision rule. There are no conformity rewards, sanctions for dissent, or strategic misreports in this model. This private-signal tie convention is an explicit variant of the random tie-breaking convention in the basic example of Bikhchandani, Hirshleifer, and Welch [47]. The probabilities derived below belong to the stated variant, not to every informational-cascade model.

Set $\lambda = p/(1-p) > 1$. Before a cascade starts, define an integer public score z so that the likelihood ratio of the observed action history is λ^z. A new private signal changes its owner's posterior exponent by $+1$ or -1. At $z \in \{-1, 0, 1\}$ the stated tie rule makes the action reveal that signal. Once z reaches 2, even a contrary signal leaves a strictly positive exponent, so the action is 1 regardless of private evidence. Every later such action then has probability one under either state and adds no further likelihood. The symmetric statement holds at $z = -2$.

Proposition 35.1 (Almost-sure locking with a positive error probability)**.** *In this model a cascade begins almost surely after finitely many revealing decisions. Its probability of selecting the wrong state is*

$$\eta(p) = \frac{(1-p)^2}{p^2 + (1-p)^2} > 0. \tag{24}$$

The probability that no cascade has begun after $2r$ decisions is $[2p(1-p)]^r$. After locking, further actions contain no additional information about θ conditional on the preceding public history.

Proof. The posterior argument above identifies the revealing region and the two absorbing action regimes. Starting at score zero, a pair of equal revealing signals reaches $+2$ or -2, while an unequal pair returns to zero. Conditional on either true state, a pair locks correctly with probability p^2, locks incorrectly with probability $(1-p)^2$, and returns to zero with probability $2p(1-p)$. Conditional independence gives independent repetitions of this two-step cycle until absorption. Since $2p(1-p) < 1$, nonabsorption through r cycles has the stated probability and tends to zero. Summing the probability of eventual incorrect absorption gives

$$\sum_{r=0}^{\infty} [2p(1-p)]^r (1-p)^2 = \frac{(1-p)^2}{1 - 2p(1-p)},$$

which is (24). In an absorbing regime every subsequent action is deterministic given the public history under both possible states, so conditioning on it changes no posterior. □

For $p = 3/4$, the eventual error probability is $1/10$. A long run of unanimous actions can therefore preserve the same mistaken decision with no additional evidentiary support. Each participant remains Bayesian-rational under the stipulated information. Unanimity is an outcome of the disclosure protocol, not evidence that a large number of independent signals favored the decision.

35.2 An evidence-preserving comparison

Now change only the information contract. Suppose authentic individual signals are available to a decision rule that selects their majority. It need not infer signals from prior actions. Let $\hat{\theta}_n$ be any majority choice with an arbitrary fixed rule for ties.

Proposition 35.2 (Conditional independence permits continuing accuracy gains)**.** *Under the same signal model,*

$$\mathbb{P}(\hat{\theta}_n \neq \theta) \leq \min\left\{1, \frac{p(1-p)}{n(p-1/2)^2}\right\}.$$

In particular, the error tends to zero as n increases.

Proof. Conditional on either value of θ, the indicators $\mathbf{1}\{s_t = \theta\}$ are independent Bernoulli variables with mean p and variance $p(1-p)$. A wrong majority requires their sample mean to be at most $1/2$, including a possible wrong tie. Chebyshev's inequality bounds this event by $p(1-p)/[n(p-1/2)^2]$. Averaging the identical conditional bounds over θ proves the result. □

This elementary bound is deliberately conservative. The exact finite error is a binomial tail, with a tie term determined by the decision rule. The comparison does not assume that all signal disclosures are truthful by virtue of being public. Authentication, privacy, observation cost, and reporting incentives are additional contracts. The result isolates the value of preserving usable independent evidence, not a general preference for making private states public.

35.3 Designing epistemic institutions rather than enforcing agreement

The institution chooses what enters its record. Publishing an action without its evidentiary basis may be adequate for coordinating a resource allocation but inadequate for later scientific assessment. A registry can therefore retain a decision, the observations supporting it, their dependencies, and the unresolved alternatives as distinct objects. An office may comply with the current allocation while maintaining a contrary hypothesis about the evidence behind it.

Possible countermeasures include independently acquired observations, protected channels for disconfirming evidence, and periodically reopening a settled question when the evidence contract changes. Their effectiveness depends on whether the new information reaches the actual decision procedure. Repeating the same cached report through a separate channel supplies no new signal. Likewise, randomized prompts do not establish conditional independence of errors. The retained provenance distinctions of Section 34 remain necessary.

In the commons, an urgency approval can be a binding procedural fact without being an independent observation that the underlying task was genuinely urgent. Subsequent reviewers who see only approvals can rationally overrely on them. Keeping the supporting certificates separate from the sequence

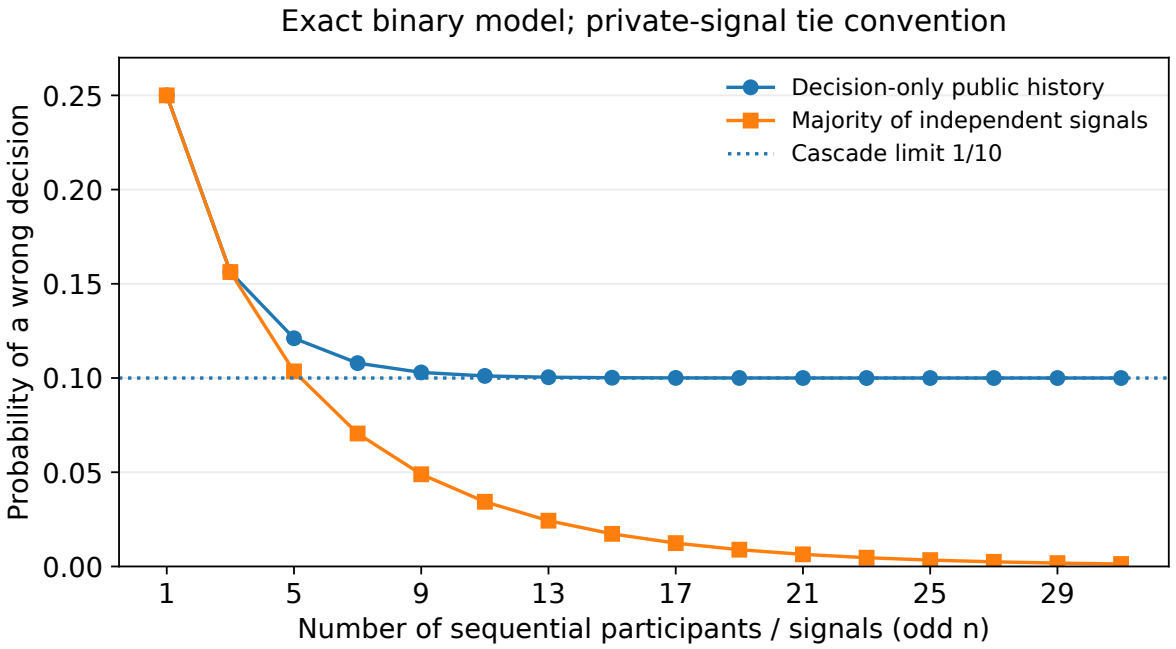


Figure 10. The binary analytical model with signal accuracy $p = 3/4$. For odd population sizes, the error of the last action approaches the cascade floor $1/10$, while a majority of the underlying independent signals continues to improve. Both curves are calculated from finite probabilities, not sampled agent runs.

of votes makes it possible to ask what was observed, what was inferred, and what was authorized. Collective ignorance is then analyzed as an information-processing failure rather than automatically as dishonesty, irrationality, or a failure to reach consensus.

36 Public knowledge, notice, and epistemically faithful transport

Knowing a rule, knowing that another participant knows it, and having a shared protocol for when it takes effect are different properties. A common record can help establish them, but the existence of that record does not imply any particular agent has accessed it. Distributed epistemic logic gives a precise language for these distinctions [25]; dynamic epistemic logic studies how communicative events change them [39].

Fix one configuration epoch and a nonempty group J of named decision centers. Let W be a finite set of possible worlds. Each equivalence relation $\sim_i$ records the worlds agent i cannot distinguish using its permitted information. A world may include the epoch, relevant history, and current records. This is an idealized semantics of information, not a claim that an implemented agent can compute every logical consequence.

For a proposition φ, $K_i\varphi$ holds at w when φ holds at every $v \sim_i w$. Let $E_J\varphi = \bigwedge_{i\in J} K_i\varphi$. Define common knowledge by all finite depths,

$$C_J\varphi = \bigwedge_{n\geq 0} E_J^n\varphi, \qquad E_J^0\varphi = \varphi.$$

Distributed knowledge instead quantifies over the *intersection* of the agents' indistinguishability classes. Common knowledge is governed by chains in their *union*. Pooling complementary private evidence can therefore reveal a fact that is not public to any participant.

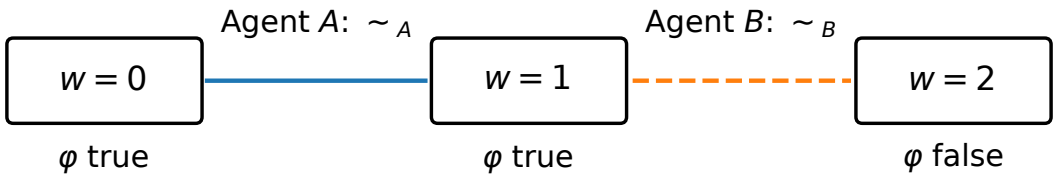


Figure 11. At world 0, each agent's own alternatives satisfy φ. Their combined indistinguishability graph nevertheless reaches world 2, where φ fails. Reflexive edges are omitted. Shared first-order knowledge is weaker than common knowledge.

Proposition 36.1 (Common knowledge as a reachability condition)**.** *Let* $\mathsf{Reach}_J(w)$ *be the worlds reachable from* w *by a finite chain of relations* $\sim_i$ *with* $i \in J$*, including the zero-length chain. Then*

$$w \vDash C_J\varphi \quad \Longleftrightarrow \quad \mathsf{Reach}_J(w) \subseteq \{v : v \vDash \varphi\}.$$

Proof. Induction on n shows that $E_J^n\varphi$ requires φ at the endpoint of every chain of n indistinguishability steps. Requiring this for every finite n is exactly the stated reachability inclusion. □

Counterexample 36.2 (Everyone knows, but it is not common knowledge)**.** Let $W = \{0, 1, 2\}$ and let φ hold exactly at $0, 1$. Agent A has partition $\{0, 1\} \mid \{2\}$, and agent B has partition $\{0\} \mid \{1, 2\}$. At world 0, both agents know φ. However, the chain $0 \sim_A 1 \sim_B 2$ reaches a world where it is false. Thus $E_{\{A,B\}}\varphi$ holds at 0 but $C_{\{A,B\}}\varphi$ does not.

36.1 An ideal public announcement is a strong update assumption

For a propositional fact p whose valuation is unchanged by the announcement, an ideal truthful public update restricts W to $W_p = \{w : w \vDash p\}$ and restricts every $\sim_i$ to that set. This update models that the event, its interpretation, and its receipt are public to the group. In the restricted model, every reachable world satisfies p, so Proposition 36.1 gives $C_J p$. The conclusion follows from the update semantics; it is not a theorem that writing a file implements that event.

The restriction to a stable propositional fact matters. An announcement can change the truth of a formula about what someone does not know. Nor does a public fact in one electorate remain common knowledge after admitting a fresh member. The relevant group, epoch, interpreter, delivery conditions, and background assumptions must be stated.

Classical limitations on common knowledge depend on the timing and communication model, and not every useful coordination task requires full common knowledge [25]. A sequential allocation may need only a valid certificate at the point of execution. A simultaneous global activation rule demands a stronger knowledge condition. Bounded acknowledgments and versioned notices can implement specified weaker requirements without pretending to supply infinitely many levels of mutual assurance in an arbitrary network.

36.2 Notice is an institutional choice as well as an information condition

A rule may define applicability by a public epoch, by authenticated receipt, or by an acknowledgment procedure. These are different institutional contracts. Whether lack of notice excuses an action is not deduced from K_i alone. It is an explicitly chosen relation between the information state and the permission or obligation monitor. In the commons, an approval can exist in the ledger before a disconnected office knows it exists. The specification must say whether that office is allowed to proceed under an earlier epoch, must wait, or can appeal a late notification. Truth of an adoption record and enforceability against a participant are separate observables.

36.3 Preserving behavior need not preserve what participants know

A state quotient that retains allocations may identify worlds with different notification histories. It can therefore preserve a task trace while changing claims of the form "office A knows that office B has accepted the debt." To preserve such claims, a stronger interface is needed.

Consider models on W_A, W_B with the same named agents and atomic predicates. Let $T : W_A \to W_B$ be surjective and preserve the truth of each atom in both directions. For every agent i, require

$$\begin{aligned} w \sim_i^A v &\implies T(w) \sim_i^B T(v) \quad \text{(forth)}, \\ T(w) \sim_i^B z &\implies \exists v :\ w \sim_i^A v \text{ and } T(v) = z \quad \text{(back)}. \end{aligned}$$

These are the usual bounded-morphism conditions, stated here as a transport contract.

Proposition 36.3 (Epistemic preservation and reflection)**.** *Under these conditions, w and $T(w)$ satisfy exactly the same formulas built from the matched atoms, Boolean connectives, K_i, and C_J. Such transports compose. If the same state maps also satisfy Definition 16.1, they preserve the translated operational conversations and the stipulated epistemic formulas at every corresponding reached state.*

Proof. Use induction on formulas. Atoms and Boolean connectives follow from the assumptions. For K_i, forth sends every source alternative to a target alternative, while back lifts every target alternative; applying the induction hypothesis proves both implications. A finite chain of indistinguishability steps is mapped by repeated forth and lifted by repeated back. Proposition 36.1 therefore proves the common-knowledge case. Forth composes directly, and two successive applications of back lift any target alternative through the intermediate model. Surjectivity and atomic truth preservation also compose. The operational conclusion then follows from Proposition 16.2 at each reached state. □

The theorem does not cover new agents whose knowledge relations have not been matched, arbitrary target messages outside the translation, or a change in what the atoms mean. It also makes no inference from formal knowledge to trustworthiness, consent, or legitimacy. Its useful conclusion is narrower: dynamic behavior, message effects, and epistemic claims can be preserved together, but each requires its own explicit part of the interface certificate.

PART IV

Authority and joint feasibility

37 Effective authority, vetoes, and institutional capture

The number of agent processes is not the number of independent centers of authority. One credential can control many processes; several credentials can be required for one consequential operation. An institutional analysis must therefore begin with office rights and operational dependencies rather than a head count.

Fix a nonempty family $\mathcal{W}$ of minimal approval coalitions for a particular amendment. Each coalition is a nonempty subset of the n registered offices. Approval is monotone: an amendment passes exactly when the set of approving offices contains some member of $\mathcal{W}$. Assume every office can independently approve or refuse, and that no unmodeled administrator or scheduling dependency bypasses the rule. This is a simple-game representation of approval authority [33].

Proposition 37.1 (Blocking coalitions as hitting sets)**.** *A coalition C can block the amendment by withholding all of its approvals, regardless of what the other offices do, if and only if*

$$C \cap W \neq \emptyset \quad \text{for every } W \in \mathcal{W}.$$

Thus the minimum number of offices with unconditional blocking power is the minimum hitting-set size of $\mathcal{W}$. Under an equal q-of-n rule with $1 \leq q \leq n$, it is $n - q + 1$.

Proof. If C intersects every minimal approval coalition, refusing all approvals in C prevents each such coalition from being contained in the approvers. Conversely, if some W is disjoint from C, its offices can approve and pass the amendment despite C's refusal. For the threshold rule, the complement of C can pass exactly when it has at least q members, so blocking requires $n - |C| < q$, or $|C| \geq n - q + 1$. □

In the commons, two offices can block a three-of-four amendment, but two cannot enact it. A single office is a veto holder exactly when it belongs to every minimal approval coalition. These are statements about one rule and one event, not complete measurements of social power.

37.1 An approval rule may conceal another authority structure

Suppose the nominal three-of-four process cannot commit any decision without a separate administrator's signature. The effective approval coalitions must include that administrator. Applying the threshold formula to only the four visible voters would overstate their independence. Scheduling, admission of evidence, access to the amendment queue, and power to replace the checker can similarly alter the implemented authority structure.

The correct remedy is to enlarge the declared realization and recompute the authority relations, not to infer decentralization from its public diagram. Conversely, a veto can protect a legitimate minority or a resource owner's consent. Small blocking coalitions do not by themselves establish harmful capture. Capture requires an additional protected-purpose or authorization criterion against which the actual control is judged.

An institution can also be procedurally compliant while concentrating benefits or suppressing effective appeal. That is a normative and incentive question distinct from whether a signature threshold was satisfied. A theory of authority should preserve the distinction between *ability to prevent*, *ability to enact*, *ability to set the agenda*, and *ability to change the procedure*.

38 Founding authority and the limits of internal attestation

Procedural continuity and authorized origin are not the same property. A constitution can specify how amendments are judged after its adoption without settling who was entitled to adopt it. An internally consistent sequence of certificates can preserve the first record's assertion of authority even when no accessible evidence establishes that assertion.

The question can be stated without selecting a universal theory of legitimacy. Let $\mathcal{H}$ be a set of possible founding histories, let $a : \mathcal{H} \to \{0, 1\}$ be an externally stipulated predicate such as "the resource owner authorized this founding event," and let $R : \mathcal{H} \to \mathcal{R}$ be the record available to the institution. The predicate models one particular authorization criterion, not moral legitimacy in its entirety.

Proposition 38.1 (When a founding fact is determined by internal records)**.** *There exists a function $f : \mathcal{R} \to \{0, 1\}$ satisfying $f(R(h)) = a(h)$ for all $h \in \mathcal{H}$ if and only if a is constant on every nonempty fibre of R. This is a statement of informational determination, not necessarily of computability. If two histories with different labels induce identical distributions of every record available to an evaluator, no evaluator restricted to those records distinguishes them with accuracy above $1/2$ under equal priors.*

Proof. If $R(h) = R(h')$, any function of the record gives the same answer on both histories, proving necessity. If the label is constant on every nonempty fibre, define $f(r)$ to be that common value and choose arbitrary values off the image of R. This gives sufficiency. For the probabilistic statement, equal record laws remain equal after any common randomized decision procedure. Under equal priors, its probability of a correct label is therefore $1/2$. □

A self-authored line saying "this constitution was authorized" cannot distinguish two histories that leave identical accessible records. An authenticated receipt from a recognized external authority may distinguish them, but then the recognition of that authority and the key binding are explicit premises. This does not mean that all records are untrustworthy or that an infinite regress makes institutions impossible. It identifies the trust boundary instead of concealing it inside the verifier.

38.1 Three judgments that should remain separate

Internal validity asks whether a transition follows the currently accepted formal rules. Historical authorization asks whether the relevant actor actually granted a particular authority under the declared criterion. Normative legitimacy asks whether the criterion and arrangement are justified. A proof system can establish the first relative to its semantics. Evidence and trust assumptions can support the second. Neither calculation uniquely determines the third.

An institution may openly adopt an internally defined founding convention and state the resources and participants to which it applies. What it cannot do by definition alone is change an external ownership fact, manufacture another party's consent, or retroactively authorize actions on behalf of someone absent from the convention. The distinction is between declaring a rule for a model and establishing a fact about its realization.

38.2 A federation inherits trust boundaries, not just certificates

When one institution imports another's origin record, it must decide which trust anchors, delegations, revocations, and unresolved disputes it recognizes. A translator can preserve the bytes and inferential structure of a certificate while the receiving institution rejects its root authority. That is not necessarily a translation error. It can be an explicit disagreement about authorization.

The commons therefore needs separate records for the resource owner's founding grant, the adopted internal procedures, and subsequent authorized amendments. A successor can preserve all three and still encounter a disputed external premise. Its honest state includes that dispute. Consistent internal history is evidence of procedural continuity, not an all-purpose certificate of truth or legitimate origin.

39 Conflict, sabotage, and organizational pathology

A machine society is not guaranteed to be cooperative, benevolent, or stable. Multiple decision centers can cooperate on one resource and compete on another. The same persistent rules that economize on coordination can create concentrated authority, lock in mistakes, or transmit a harmful practice to later participants. Security is therefore not an unrelated afterthought; it is partly the problem of preserving valid relationships between declarations, implementations, permissions, and evidence.

39.1 Failures occur at different layers

We distinguish at least five abstract failure surfaces. *Content failures* corrupt beliefs or stored facts. *Interface failures* alter the interpretation or routing of otherwise unchanged content. *Authority failures* allow an unauthorized actor to issue a consequential operation. *Revision failures* change which future evidence or rules are admissible. *Allocation failures* exploit shared resource procedures. These failures can combine: a fabricated record may justify an unauthorized rule change, after which the new rule rejects the evidence needed to expose the fabrication.

The terms *sabotage* and *virus* require restraint. A sabotage claim requires an adversarial objective and a causal route by which the action impairs an institution. A "schema virus" is an analogy until one specifies the copied representation, adoption mechanism, susceptibility, persistence, and measured harm. Error propagation can be unintentional; stability can coexist with misinformation; and divergence from a prevailing view may represent a correct new discovery rather than a pathology.

39.2 A bounded propagation inequality

A transparent comparison model can make one class of failure mathematically tractable. Let $x_t \in \mathbb{R}^n_{\geq 0}$ record nonnegative deviations at n institutional positions. Suppose an explicitly justified comparison model provides

$$x_{t+1} \leq Bx_t + w_t \quad \text{componentwise}, \qquad B \geq 0, \tag{25}$$

where w_t represents fresh disturbance and B summarizes one-step propagation. Suppose a vector $v > 0$ and $\kappa < 1$ satisfy $Bv \leq \kappa v$. In the weighted maximum norm $\|x\|_{v,\infty} = \max_i x_i/v_i$ we obtain

$$\|x_t\|_{v,\infty} \leq \kappa^t \|x_0\|_{v,\infty} + \sum_{j=0}^{t-1} \kappa^{t-1-j} \|w_j\|_{v,\infty}. \tag{26}$$

The proof is the one-step monotonic estimate followed by induction. Under uniformly bounded disturbance $\|w_t\|_{v,\infty} \leq \bar{w}$, the last term is at most $\bar{w}(1-\kappa^t)/(1-\kappa)$. This gives a conditional attenuation result,

not a universal contagion threshold. A graph alone cannot supply the matrix B; the rates must come from a justified dynamical model. More connectivity can improve detection as well as propagation, and nonlinear strategic behavior can invalidate the linear comparison.

39.3 Institutional capture and the meaning of power

Power is not exhausted by reputation or by the number of messages an agent produces. In a computational institution it can arise from privileged read or write access, control of scarce resources, amendment rights, scheduling priority, the right to certify evidence, and the ability to prevent an appeal. A useful power analysis distinguishes *information reach* from *decision reach*: a participant can influence many beliefs while holding no formal permission, or can issue one privileged command with enormous consequences.

A simplified reach functional for agent i and horizon T is the set of consequential institutional coordinates whose future distributions can be changed by some *admissible unilateral intervention* of i. This is an intervention-relative property; it must not silently assume that the agent can overwrite a variable to which it has no access. Quantifying authority requires the feasible intervention family and a choice of outcomes. A concentration score built from these reaches can be a diagnostic, but not a proof of illegitimacy.

Formal policy compliance alone does not rule out institutional capture: a rule may be faithfully executed and still privilege a coalition that controls amendments or vetoes. Likewise, a successful self-repair routine may restore a harmful prior state rather than correct it. Any proposed notion of health must include protected freedoms, an error-correction contract, and an explicit source of normative authority. Agreement is not itself the objective.

39.4 Defensive design implications

Ordinary security principles retain their relevance: least privilege, complete mediation, separation of authority, auditable transitions, compartmentalized credentials, and recovery from corrupted records [13]. A proof-carrying decision can narrow the trusted checking base [12], but cannot prove the adequacy of its own external assumptions. Copyable identities create classic Sybil concerns [14]. A fork's assertion that it is a new participant is not automatically a new independent vote or a new allocation of finite resources.

The strongest general principle is *evidence-preserving reversibility where reversal is physically possible*. Exploratory proposals need not receive immediate permission to rewrite canonical rules; reviews should preserve rejected alternatives and reasons; rollbacks should retain an audit trail; and irreversible external actions require a different permission standard from reversible sandbox computations. This is a defensible design stance, not a theorem that one fixed governance policy fits every environment.

40 Societies of societies: federation, mobility, and scale

Once groups can maintain institutions, their boundaries become objects of interaction. A participant may migrate between two groups, carry an archive from one to another, borrow computational resources, or hold obligations in multiple jurisdictions. A coalition may split without dissolving all of its records; two institutions may merge while retaining conflicting rules. Such events define an *ecology of organizations*, not merely a larger population of otherwise unrelated agents.

The relevant inter-organizational interface has more structure than a message format. It includes state translation, identity and provenance assertions, resource transfer conditions, compatible permission

semantics, and a means to resolve contradictory commitments. Calling two communication protocols interoperable because their messages parse correctly ignores whether the resulting actions are authorized and what happens to existing obligations.

40.1 Migration is not unrestricted inheritance

A migrating process can carry operational knowledge without carrying valid credentials. An organization that accepts all imported permissions merely because an agent has an intact memory confuses history with authority. A safer abstract contract is explicit about which local capabilities are newly granted, which predecessor obligations continue, which are discharged, and which disagreements require arbitration. Copying a state does not copy a budget, physical hardware, a nontransferable credential, or any legal status assigned by people.

40.2 Scale changes which institutions are economical

At small scale, bilateral negotiation may be cheap and simple. At larger scale, direct communication can become expensive, so rules may be compressed into roles, standards, and delegated procedures. Hierarchy can reduce latency or message count under explicit network constraints; Section 10 also proves that an unrestricted controller can simulate the same finite computation. The correct separation concerns resources, privacy, fault domains, and the work required to discover or maintain useful interfaces, not a new computability class created by social grouping.

Large networks also increase the surface for copied mistakes, inconsistent rule versions, and concentrated control over common infrastructure. Modularity can localize some failures while making cross-boundary repair harder. There is no universal monotone law saying that greater connection makes a machine society more stable, intelligent, or vulnerable. Those effects depend on transition kernels, institutional incentives, topology, and the cost of checking interactions.

40.3 The absence of an inevitable collective self

It is tempting to imagine a sequence from many agents to coordination, from coordination to institutions, and from institutions to a unified intelligence. Nothing in the framework implies that final arrow. Some federations are valuable precisely because they prevent one controller from collapsing plural judgments. Others may deliberately adopt a common policy for restricted operations while retaining disagreement on everything else. The existence of a macroscopic description is also not evidence of a first-person perspective. Computational closure, social organization, legal recognition, and subjective experience remain separate questions.

41 Endogenous roles, adaptive networks, and resilient specialization

A continuing office need not be one of a permanently supplied list. A population may establish a verification service, divide a task into specialties, retire a redundant committee, or form a temporary coalition. Its communication graph can change with these decisions. To express that possibility, let G_t specify currently permitted communication links, r_t an assignment of roles, and c_t a competence and capacity record. Their transition belongs to the realized state:

$$(G_{t+1}, r_{t+1}, c_{t+1}) \sim \Gamma_{I_t}(\cdot \mid G_t, r_t, c_t, H_t),$$

where H_t denotes the retained information used by the update. A Markov formulation must include a sufficient part of that information in the state rather than treating an unspecified history as an unobserved input.

Adaptive-network theory studies the coupling of node dynamics to changing topology [48]; strategic network-formation theory distinguishes beneficial networks from those stable under participants' incentives [49]. Here the additional question is institutional: which training, delegation, link creation, and role transfers are authorized, and which existing commitments survive those changes? Endogenous division of labor requires that participant interaction affects the adopted structure. Writing Γ does not show that an externally assigned specialization was discovered.

41.1 Specialization has a cost and a reachability condition

Consider two services performed once per period. A generalist arrangement costs $2c_g$ per period. Two specialists cost $2c_s + b$, where b is the required communication and integration cost, and training costs h for each specialist. Over T periods the specialist arrangement has lower total cost precisely when

$$T\big[2(c_g - c_s) - b\big] > 2h.$$

This is an instance of institutional amortization, not a universal law of specialization. Verification overhead, replacement, workload variation, and lost redundancy must be added when relevant. A specialist who is faster locally can still increase total cost if every result requires expensive cross-boundary interpretation.

Reachability is separate. Suppose two agents must continuously staff two different roles and each agent has capacity one. A coordinated exchange of their assignments may improve both agents' outcomes. Yet neither agent can move alone while all roles remain staffed and no role is double-booked. Permitting an atomic handover, a temporary reserve agent, or a short service interruption changes the admissible transition graph. The best assignment and a permissible path to it are different mathematical objects, just as in institutional formation and constitutional amendment.

When a finite role-and-network update system admits a common potential that strictly increases on every permitted improvement, the earlier finite termination result applies. There is no general reason for individually beneficial link changes, training choices, and delegated powers to share such a potential. Cycles and stable but inefficient specialization remain possible.

41.2 A sharp condition for replacement capacity

Let J be a finite nonempty set of roles that must be filled simultaneously and A a finite set of agents. An edge (j, a) means that agent a is both capable and authorized to fill role j during the relevant interval. Every agent can fill at most one role at a time. A feasible staffing is an injective assignment $f : J \to A$ along these edges. Let $N(S)$ be the agents adjacent to at least one role in $S \subseteq J$.

The question is not merely whether every role has a backup. It is whether backups can cover all roles together after a loss. The classical matching principle originates with Hall [50]; the following erasure form exposes the exact redundancy requirement.

Theorem 41.1 (Authorized staffing after any bounded agent erasure). *Fix an integer $0 \le k \le |A|$. A feasible staffing exists after removal of every set $F \subseteq A$ with $|F| \le k$ if and only if*

$$|N(S)| \ge |S| + k \qquad \text{for every nonempty } S \subseteq J. \tag{27}$$

The surviving capability and authorization edges are assumed unchanged, and a new assignment may be selected after the erased set is known.

Proof. For necessity, if a nonempty S has fewer than $|S| + k$ neighbors, erase $\min\{k, |N(S)|\}$ of those neighbors. Fewer than $|S|$ remain, so no injective staffing of S is possible.

For sufficiency, remove any F of size at most k. The surviving neighbors of each S number at least $|N(S)|-k \geq |S|$. We recall the finite matching argument. Choose a maximum matching in the surviving bipartite graph. If a role remains unmatched, explore alternating paths beginning at it, using unmatched edges toward agents and matched edges back toward roles. No reached agent can be unmatched, since that would produce an augmenting path. Let S be the reached roles and B the reached agents. Every neighbor of S is in B, and each agent of B is matched to a distinct reached role other than the initial unmatched role. Hence $|N(S)| = |B| < |S|$, a contradiction. The maximum matching therefore fills all roles. □

The theorem permits a simple robustness index. If the initial graph admits a staffing, define

$$\sigma = \min_{\emptyset \neq S \subseteq J} (|N(S)| - |S|).$$

Then exactly the erasure budgets $k \leq \sigma$ are guaranteed tolerable in this model. This is a static allocation certificate, not a claim that the surviving agents can discover or execute the reassignment before a deadline. That further question uses the communication and permission constraints of Sections 25 and 23.

Counterexample 41.2 (Backups for each role need not provide collective resilience)**.** Two roles can each be filled by either of two agents a, b. Every role has two capable authorized candidates, and the two roles can initially be filled. After either agent is erased, the one survivor cannot fill both roles simultaneously. The singleton-role conditions for $k = 1$ hold, but the two-role set has only two neighbors rather than the required three.

For three roles with neighborhoods $\{a, b\}$, $\{b, c\}$, and $\{c, d\}$, every nonempty role set satisfies (27) with $k = 1$. One erased agent can be accommodated by reassignment even though no single agent can perform every role. Resilience can reside in a pattern of overlapping specialization rather than a universally capable backup.

41.3 Time-respecting communication cannot be read from an aggregate graph

Changing topology also changes who can actually learn what. A union of links over an interval discards their ordering. The resulting graph may look connected while failing to carry a particular observation to a decision maker in time.

Counterexample 41.3 (A connected union with no usable evidence path)**.** Agent A initially holds an arbitrary bit and the other agents do not. At time one, B sends its current state to C. At time two, A sends the bit to B. No other communication occurs before C must decide. The union of directed contacts contains the path $A \to B \to C$, but C's received message is independent of the bit. No decoder at C can recover an initially fair bit with accuracy above $1/2$ under this access model.

The missing object is a time-respecting causal route, not an additional static connection. The distinction is consistent with the event order of Section 23. In a changing society, a nominal right to appeal or a nominal verification edge may be ineffective when it is never available during the interval in which evidence matters.

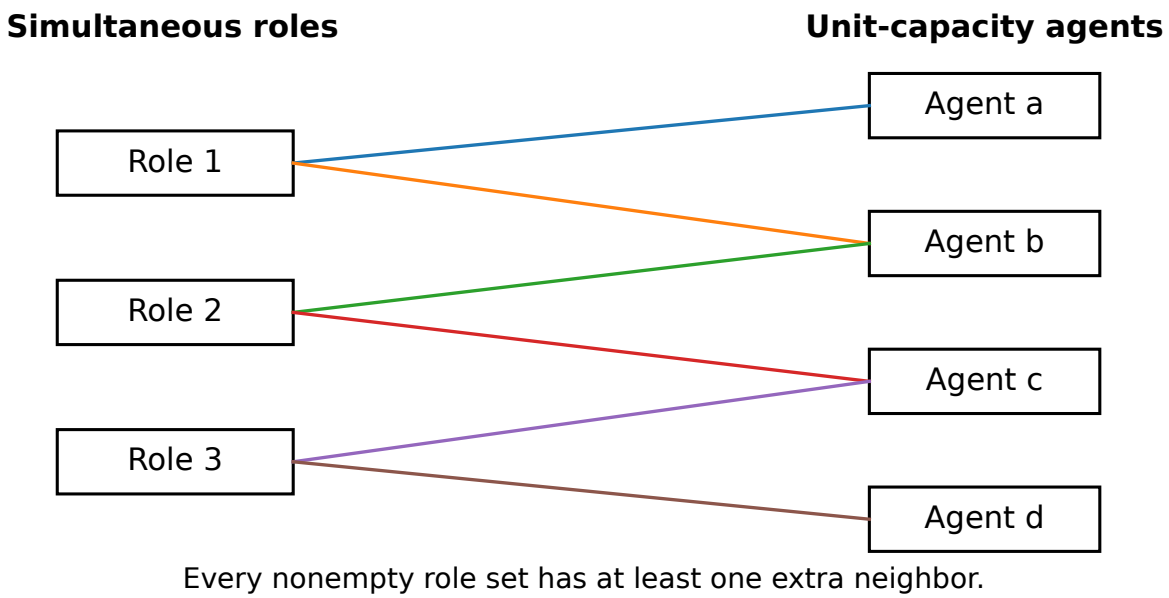


Figure 12. Three simultaneous roles and four unit-capacity agents. Edges represent both competence and authorization. The neighborhoods satisfy the one-erasure staffing criterion. The guarantee permits rematching after the failure is known; it does not guarantee instantaneous handover.

41.4 Adaptive roles do not automatically preserve institutional plurality

A new verifier can reduce errors while concentrating epistemic authority. A central routing service can reduce messages while becoming a veto point. Copying a specialist may increase nominal staffing without adding independent failure domains, and a revocation may remove authorization edges even when all technical skills remain. Each of these changes belongs in the state to which the staffing or communication theorem applies.

A useful role-update contract therefore tracks competence, capacity, decision rights, handover obligations, and accessible evidence separately. It can require that a protected staffing margin survive every admitted reorganization, while still allowing the actual role assignment to change. Such a constraint does not choose which roles deserve protection or who may impose the training cost. It makes the consequences of those choices analyzable.

The commons can now add a temporary verification role, train an alternate occupant, and revise the communication schedule without identifying an office with one model. Its social organization lies partly in this maintained division of work. The formal test is whether the adopted arrangement has a feasible and authorized realization through turnover, not whether its role names resemble a familiar human institution.

42 Joint institutional feasibility and certified continuation

An institution can possess several valid certificates and still lack a feasible way to operate. A safe allocation may require one action, an effective audit another, and a timely repayment a third. Each statement can be correct when its own resource and information assumptions are considered alone. The common implementation fails when those assumptions cannot hold together.

The resulting question is not whether a population can optimize one collective utility. It is whether a declared set of obligations has a jointly executable continuation. The protected requirements can include resource conservation, valid authority, a false-positive limit, a specified incentive margin, service deadlines, and feasible staffing. Their normative selection remains outside the synthesis theorem. Once selected, they must constrain the same action and the same realized history.

This section supplies a finite construction rather than an unrestricted theory of decentralized control. It combines the existing obligation monitors and governance labels with a catalogue of implementable local protocols. The mathematical foundations are the knowledge construction and fixed-point methods for games with imperfect information [31]. The additional contribution is to identify which institutional certificates must enter that construction, and what is lost when an abstraction omits them.

42.1 A finite contract with an explicit information boundary

Let S be a finite nonempty state space for the complete contract. A state records the relevant institutional mode, resource ownership, bounded obligation monitors, current staffing, and any private memory needed to make transitions well defined. Let $h : S \to Y$ be the observation available in a declared common record at each decision epoch. It need not reveal the private components. The initial information is a nonempty set B_0 contained in one fibre of h.

A finite nonempty catalogue U contains protocol profiles, not arbitrary commands to the complete microstate. A profile can specify one local response map for each office, with each map reading only that office's permitted observation and stored local memory. It can also specify a committed monitor, a communication schedule, and the steps needed to select and enact that profile. These implementations, their costs, and their decision rights must be established separately. An instruction that requires inaccessible private information is not made implementable by listing it in U.

For each s, let $A(s) \subseteq U$ contain profiles satisfying the immediate contract at that state. Every profile in this set must pass the same resource, authority, staffing, and monitoring checks. In particular, a monitor included in a profile must satisfy all relevant incentive and false-positive inequalities simultaneously. Let $\Delta(s, u) \subseteq S$ be a nonempty finite set of permitted next states for each s, u. The set includes every environment response or fault within the declared guarantee. Inadmissible profiles are never selected.

The contract uses *sure* guarantees: every transition in Δ is treated as possible, without weighting it by probability. A probabilistic transition model may supply these supports, but an unlikely infinite path is not discarded when proving bounded service. Nor does a deterrence inequality justify silently deleting deviations from Δ. Safety under compliant behavior and safety under specified deviations are different contracts. The latter must include their physical consequences in the transition model.

Let $S_{\text{safe}} \subseteq S$ encode the protected state requirements. Breach monitors distinguish fulfillment, authorized discharge, pending debt, and deletion. A finite bound on stored obligations or deadlines is part of the model. An overflow must have an explicit treatment, such as a rejected admission or a breach state. The theorem does not silently replace an unbounded history by a finite one.

A common record at decision epochs is itself an infrastructure assumption. It is not implied by eventual replica convergence. The model applies only when that observation and the selection of a catalogue profile have the declared implementations. The selected profile can then be executed by distinct local agents without exposing their private states. This is a controlled information boundary, not a claim that every decentralized architecture has such a representation.

42.2 The state of possible institutional realizations

The relevant public state is the set of complete states still compatible with the observed history. Define

$$\mathcal{B} = \{B \subseteq S : B \neq \emptyset,\ h \text{ is constant on } B\}.$$

These are sets of possibilities, not posterior probability distributions. At B, a profile must be admissible for every remaining possibility:

$$A(B) = \bigcap_{s \in B} A(s).$$

For $u \in A(B)$ and a next public observation y, define

$$\mathrm{Post}(B, u, y) = \left(\bigcup_{s \in B} \Delta(s, u) \right) \cap h^{-1}(y). \tag{28}$$

Only nonempty successor sets are possible observations. Write $\mathrm{Succ}(B, u)$ for their collection. Induction on the public history shows that (28) retains exactly the complete states consistent with that history and the selected profiles.

This construction does not pool private observations before choosing u. It pools only the mathematical possibilities consistent with the common record. Private observations can affect the actions of the local maps already specified by u. More elaborate strategies using additional private history require that history, or a sufficient finite memory of it, to be included in the declared realization. Completeness below is relative to this finite catalogue and information model.

42.3 A continuation kernel for the whole contract

Set

$$W_0 = \{B \in \mathcal{B} : B \subseteq S_{\text{safe}}\}, \tag{29}$$

$$W_{k+1} = \{B \in W_0 : \exists u \in A(B),\ \mathrm{Succ}(B, u) \subseteq W_k\}. \tag{30}$$

The sequence is decreasing. It stabilizes after at most $|\mathcal{B}|$ strict removals, and we denote its fixed point by W_*. This is the joint continuation kernel. A state is removed not only when it violates an invariant now, but also when every available profile has a possible successor from which the remaining contract cannot be sustained.

Theorem 42.1 (Joint continuation under a common observation history). *For every $k \geq 0$, $B \in W_k$ if and only if there exists a deterministic strategy based on the common observation and profile history that selects admissible profiles and keeps the complete state in S_{safe} for the next k transitions, starting from every $s \in B$. Moreover, $B \in W_*$ if and only if such a strategy exists for an unbounded number of transitions. On W_*, a strategy depending only on the current possibility set B suffices.*

Proof. At horizon zero the condition is exactly $B \subseteq S_{\text{safe}}$. At positive horizon the strategy must choose one profile from the same common history, so it must lie in $A(B)$. After each possible observation its continuation must work from the corresponding set in (28). Induction gives exactly (30), in both directions.

Monotonicity of the predecessor operation makes W_k decreasing. Finiteness implies stabilization. At the fixed point, select for every $B \in W_*$ one witnessing profile whose successors all remain in W_*. Induction then proves safety and admissibility for every finite prefix of every realized path, hence

for the complete path. Conversely, an unbounded safe strategy is safe for every finite horizon, so its initial set belongs to every W_k and therefore to the stabilized intersection. The witnessing selection is stationary in B, although maintaining B generally requires memory of the public history. □

No society-wide preference ordering appears in this theorem. It characterizes joint satisfaction of a declared contract, not the desirability or consent behind that contract. It also gives no favorable complexity bound: the possibility space can contain exponentially many subsets of S. A computationally expensive solution is still distinct from a procedure the society can discover and execute within its own resource budget.

The action intersection matters. A full-information controller can choose different safe profiles at two hidden states while the common-history institution cannot choose between them. This is the constructive counterpart of the indistinguishability obstruction in Section 25. The kernel can therefore be empty even when every complete state has a safe continuation under full information.

42.4 Safety does not complete a service

The fixed point can contain a safe idle loop. Outstanding deadlines should therefore be represented by monitors that advance when time passes, not by a permanently pending record. For service without a fixed external deadline, a separate progress condition is needed.

Let $G \subseteq S_{\text{safe}}$ be a stipulated completion set. The target is *certified completion*, meaning that the current possibility set lies wholly in G. Let

$$\mathcal{G} = \{B \in W_* : B \subseteq G\}, \quad A_*(B) = \{u \in A(B) : \mathrm{Succ}(B,u) \subseteq W_*\}.$$

Define $R_0 = \mathcal{G}$ and

$$R_{k+1} = R_k \cup \{B \in W_* : \exists u \in A_*(B),\ \mathrm{Succ}(B,u) \subseteq R_k\}. \tag{31}$$

The increasing sequence stabilizes at R_*. For $B \in R_*$, let $r(B)$ be its first index of entry.

Proposition 42.2 (Safe service with a finite progress rank)**.** *A set B lies in R_k if and only if an admissible common-history strategy can ensure certified completion within at most k transitions while preserving the contract indefinitely. A set lies in R_* if and only if such a strategy can ensure eventual certified completion on every possible path while preserving the contract. Before completion, a stationary choice with strictly decreasing r reaches the target within $r(B)$ transitions. After completion, a safety strategy on W_* suffices.*

Proof. The bounded claim follows by induction from (31). At a nonterminal set of rank $r > 0$, a witnessing profile sends every successor to rank at most $r - 1$. All selected profiles preserve W_*, so a safety continuation is available after the target is reached.

For the eventual claim, consider the tree of public histories consistent with a fixed successful deterministic strategy, stopping a branch at certified completion. The tree is finitely branching. If its depths were unbounded, repeatedly choosing a child with descendants at arbitrarily large depths would produce an infinite branch with no completion. Every finite prefix has a consistent complete-state realization by (28). Finiteness and the same branching argument give a compatible infinite complete-state realization, contradicting success on every path. The tree therefore has a finite depth bound, so the bounded claim puts B in some R_k. The other direction follows from the rank strategy. □

The completion condition is intentionally epistemic: the common history must determine that the stipulated work is complete. Physical completion without an available completion record is a weaker endpoint. The result does not identify the two. It also does not promise service for requests that the admission policy rejects. Guaranteed access, when required, must be part of the contract rather than inferred from the success of already admitted work.

Counterexample 42.3 (Almost-sure completion without a sure service bound)**.** At a pending state p, one action moves to a completed absorbing state g with probability $1/2$ and returns to p with probability $1/2$. Both states are safe and publicly visible. Completion occurs with probability one and has expected waiting time two, but the support contains the infinite path $p, p, \ldots$. For the sure interpretation of Δ, W_* contains both singleton states while R_* contains only $\{g\}$. Replacing sure completion by almost-sure completion changes the problem.

Belief sets suffice for the stated sure safety and certified-reachability objectives. They are not a universal sufficient statistic for strategic partial-observation problems. In particular, broader stochastic games can require information beyond a belief-based strategy [52]. The quantifiers and the available randomization must be retained when comparing those results with the present construction.

42.5 A certificate can be checked without finding the largest kernel

The fixed-point computation finds every viable possibility set in the declared model. A proposed institution may instead supply a smaller invariant family together with its choices. Let $\mathcal{C} \subseteq \mathcal{B}$ contain B_0, and let κ assign a profile to each $B \in \mathcal{C}$. The checker verifies $B \subseteq S_{\text{safe}}$, $\kappa(B) \in A(B)$, and $\mathrm{Succ}(B, \kappa(B)) \subseteq \mathcal{C}$. A separate integer rank can certify a selected service.

Proposition 42.4 (Soundness of a finite continuation certificate)**.** *If the preceding checks hold, κ preserves the complete contract from B_0 for all transitions. If, in addition, there is a map $r : \mathcal{C} \to \mathbb{N}$ such that $r(B) = 0$ implies $B \subseteq G$ and every successor of a set with positive rank has strictly smaller rank, then a set contained in G is reached within at most $r(B_0)$ transitions. Checking these finite conditions does not require proving $\mathcal{C} = W_*$.*

Proof. Induction preserves membership in $\mathcal{C}$, admissibility, and safety. Until rank zero is reached, each transition strictly decreases a nonnegative integer. There can be at most $r(B_0)$ such transitions. The rank-zero implication then establishes certified completion. Closure of $\mathcal{C}$ continues to provide safety afterward. □

An amendment can be one of the profiles in this certificate. It must then be authorized before execution, carry every relevant monitor and resource meaning through the transition, and land in a certified continuation set. Allowing arbitrary rank-preserving amendments before a service can defeat its time bound. Such amendments need their own bounded allowance, or must be excluded until completion. The obligation cannot be declared fulfilled merely by changing the label attached to a pending record.

Counterexample 42.5 (A safe amendment can destroy every safe continuation)**.** There are four publicly distinguishable states r, ℓ, g, b. The safe set is $\{r, \ell, g\}$. At r, profile deliver moves to g, while profile freeze moves to ℓ. At g, wait loops safely. At ℓ, the only admissible profile advances an outstanding deadline and moves to breach state b. The amendment freeze preserves the immediate safe-state predicate but not W_*. No certificate proving unbounded continuation from r can choose it. The relevant defect is loss of a future capability, not an immediate inconsistent record.

42.6 Which abstractions preserve the ability to govern?

The existing governance quotient preserves labels, enabled actions, and deterministic updates. For synthesis under incomplete information, an abstraction must also preserve the public observation and the full set of possible successors. Preserving one average outcome or one representative trajectory is insufficient.

Consider a reduced finite contract $(\overline{S}, \overline{h}, \overline{A}, \overline{\Delta}, \overline{S}_{\text{safe}}, \overline{G})$ with the same catalogue U. Let $\pi : S \to \overline{S}$ be a surjection. Assume, for every state and every admissible profile,

$$h(s) = \overline{h}(\pi s), \qquad A(s) = \overline{A}(\pi s), \tag{32}$$

$$s \in S_{\text{safe}} \iff \pi s \in \overline{S}_{\text{safe}}, \qquad s \in G \iff \pi s \in \overline{G}, \tag{33}$$

$$\pi[\Delta(s,u)] = \overline{\Delta}(\pi s, u). \tag{34}$$

The notation $\pi[B]$ denotes the image of a set, not a probability pushforward. These are exact assumptions. They must hold for amendments and handovers as well as ordinary actions whenever those operations belong to the contract.

Theorem 42.6 (Governance abstraction preserves joint realizability). *Under* (32)–(34)*, for every $B \in \mathcal{B}$ and every horizon k,*

$$B \in W_k \iff \pi[B] \in \overline{W}_k, \qquad B \in W_* \iff \pi[B] \in \overline{W}_*.$$

The progress sets satisfy $B \in R_k \iff \pi[B] \in \overline{R}_k$, so finite progress ranks agree. A reduced safe or ranked strategy pulls back by choosing at B the profile assigned to $\pi[B]$.

Proof. The observation equality makes $\pi[B]$ a valid reduced possibility set. Admissibility and safe-set membership are preserved and reflected. For every possible next observation y, successor equality gives the commuting update identity

$$\pi[\text{Post}(B,u,y)] = \overline{\text{Post}}(\pi[B], u, y). \tag{35}$$

Indeed, applying π to the union of complete successors gives the union of their reduced successors. The observation equality permits intersection with $h^{-1}(y)$ to pass through π. Nonemptiness is also preserved and reflected.

Induction in (29)–(30) now proves the finite-horizon equivalence. Membership at every horizon gives the fixed-point equivalence, since both sequences stabilize. The goal equivalence in (33), the same successor identity, and induction in (31) prove the progress claims. Every reduced witnessing profile is admissible in every complete state in B, and its complete successors have the required reduced images. Pulling back these choices therefore preserves safety and the decreasing progress rank. □

The theorem composes earlier preservation obligations into a strategy claim. It is an exact finite application of quotient and game-synthesis reasoning, not a general solution to decentralized governance. It is stronger than agreement of current outputs because it preserves the existence of a future course of action under the actual information boundary. It remains conditional on the implementation of each catalogue profile and on the coverage of Δ.

Dropping observation preservation can give a reduced controller a signal the real population never receives. Dropping enabled-action preservation can turn an unauthorized command into an apparent solution. Dropping successor equality can erase an inconvenient failure path. Dropping the obligation

labels can make an unpaid debt disappear in the quotient. Each change can alter the winning sets while leaving a chosen immediate output unchanged. An approximate transition estimate can support a probabilistic error budget, but cannot by itself justify the exact sure guarantee proved here.

42.7 The commons: individually feasible promises without a common budget

Consider a declared variant of the four-office commons with c execution slots remaining in an epoch. Three protected units of work are due before the epoch ends. They include a carried service debt and two other admitted jobs. Each unit consumes one slot. This is a new bounded workload, not a claim that the original allocation rule always creates these obligations.

An ongoing reporting procedure is exposed to two listed deviations, each of gain 1. Its authorized penalty is $F = 2$, and its permitted false-positive rate is $\alpha = 1/4$. With no audit, all three observation laws coincide. A one-slot audit has $P_C = (1/2, 1/2)$ and the two opposite deviation laws from Counterexample 19.2. Their equal mixture reproduces compliance, so no common monitor deters both positive-gain deviations.

A two-slot audit has three outcomes and the stipulated laws

$$P_C = (3/4, 1/8, 1/8), \qquad P_1 = (0, 1, 0), \qquad P_2 = (0, 0, 1).$$

The common monitor $s = (0, 1, 1)$ sanctions compliance with probability $1/4$, each deviation with probability 1, and has margin $2(1 - 1/4) - 1 = 1/2$ against each. The audit costs and laws are assumptions of the analytical variant. They are not measurements or consequences of the earlier resource allocation.

The joint action consists of one audit mode and the amount of work completed in that epoch. A profile must fit both within c. Protected work cannot be discarded or quietly deferred. The following comparison is therefore exact for this catalogue:

Contract	Audit slots	Due service slots	Minimum total
Service only	0	3	3
Listed deterrence only	2	0	2
Both at once	2	3	5

With $c = 4$, both component contracts are feasible separately, but their conjunction is not. With $c = 5$, the two-slot audit and all three services fit, and one-step completion is certified. With $c = 4$ and only two protected service units, the same joint profile also fits. In the general version with d protected units, the exact threshold is $c \geq d + 2$.

For this richer audit, the best common margin at a general false-positive cap is

$$\Gamma_\alpha = \min\{6\alpha, 3/2\} - 1.$$

To see this, average any monitor with the monitor obtained by exchanging outcomes 1 and 2. The worst deviation margin does not decrease. Write the two equal sanction probabilities as t. Setting the outcome-0 sanction probability to zero reduces false positives and weakly improves the margin. The remaining constraint is $t/4 \leq \alpha$, with $t \leq 1$, and the margin is $3t/2 - 1$. Thus the formula is exact. The cap $\alpha = 1/4$ yields margin $1/2$, but the audit still consumes two execution slots.

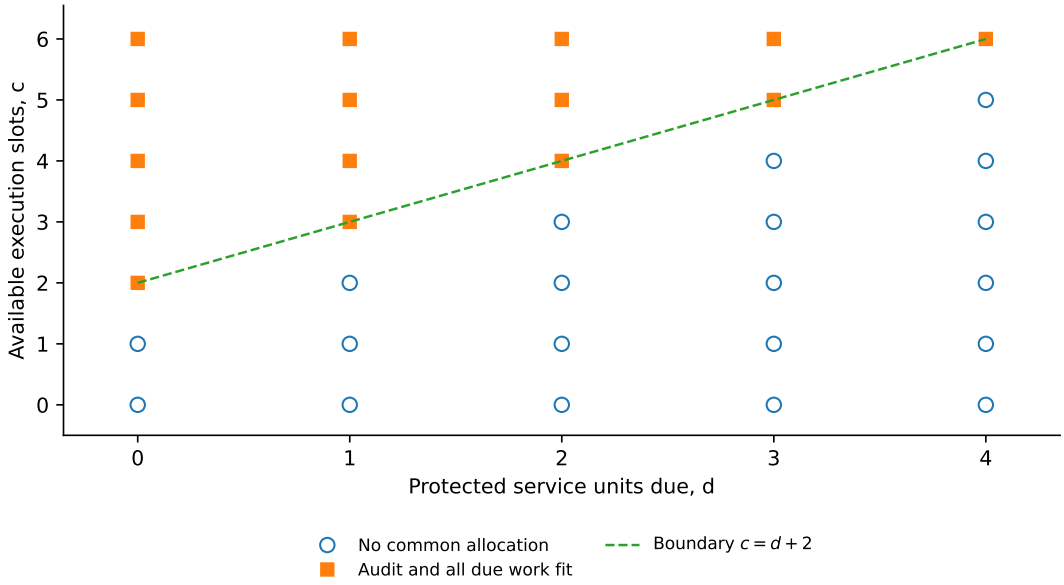


Figure 13. Exact feasibility in the bounded commons variant. Every marked point is one integer pair (d, c). Open markers violate the combined budget, while filled markers admit the two-slot audit and all due services. The boundary is $c = d+2$. A feasible audit and a feasible service allocation considered separately do not establish joint feasibility.

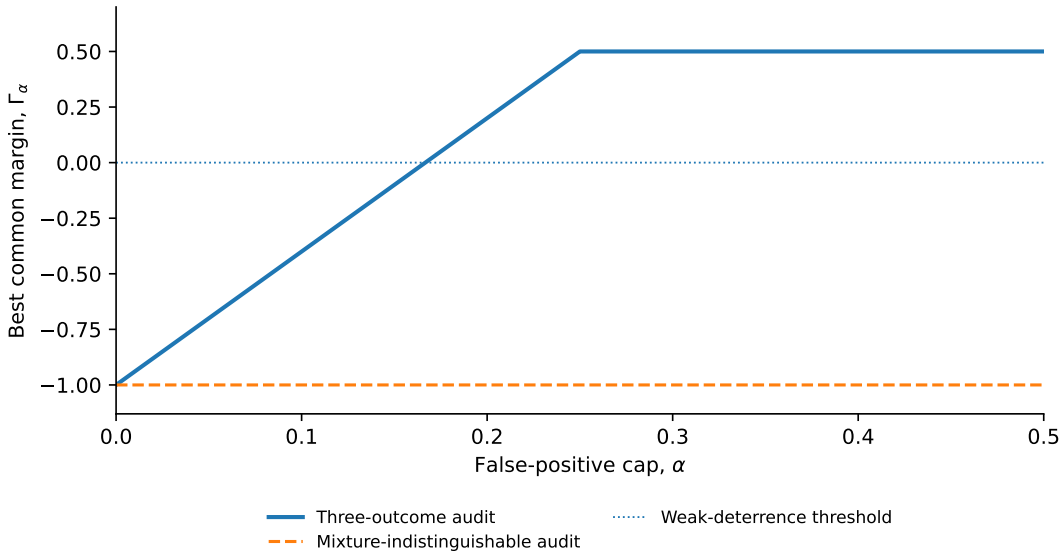


Figure 14. Analytical common-monitor margins for two specified audit channels with $F = 2$ and gains $g_1 = g_2 = 1$. The richer three-outcome audit has margin $\min\{6\alpha, 3/2\} - 1$. For the two-outcome audit whose deviation mixture equals compliance, the best simultaneous margin is -1 at every cap. The curves are exact formulas, not fitted observations.

The example locates the obstruction instead of hiding it inside a favorable aggregate score. A fifth slot repairs the capacity failure without weakening the promises. An explicitly authorized rescheduling agreement changes the temporal contract rather than pretending that the original deadline was met. A cheaper, validated observation channel changes the audit catalogue. Each route needs its own authority and implementation. The kernel computation certifies only the route actually specified.

42.8 What the synthesis adds to the theory

There are now three different witnesses. A monitor certifies simultaneous incentive inequalities in a specified observation model. A continuation certificate supplies an executable choice at every retained public possibility set. A transport certificate shows that the reduced representation preserves those choices, their observations, and their consequences. None can replace the other two.

The result is an institutional realization problem with a finite decision procedure and checkable witnesses. It does not establish spontaneous institution formation, solve the enforcer's incentives, infer a common normative objective, or make a missing public event mechanism appear. It does show how permissions, observations, service obligations, and resource limits can constrain one continuing policy rather than remain independent descriptions of incompatible possibilities.

PART V

Accountability and the limits of persistence

43 Accountability, revision, and authorized endings

An institution can continue to operate while becoming less answerable to the parties it affects. It can preserve a rule whose authority has expired, fund sanctions but not appeals, or treat unfinished obligations as a reason never to accept an authorized ending. None of these possibilities is excluded by persistence alone. The remaining question is therefore not simply whether a computational organization survives. It is whether the same realized organization respects the limits of its authority, supports effective correction, and can enter a specified suspension or settlement state.

This section extends the common-policy question of Section 42. External claims, auditor incentives, confidentiality, review capacity, stopping procedures, and the continuing applicability of certificates must constrain one implementation. They are not additional benefits that can be asserted after the implementation has consumed its entire budget. Nor are they consequences of selecting the word *society*. Each is an explicit part of a contract whose normative selection remains separate from its mathematical feasibility.

43.1 Affected outsiders and the boundary of authority

Membership, decision rights, resource control, and exposure to consequences define different sets. Let M denote participating members, V positions with specified decision rights, H infrastructure controllers, and E affected parties outside the membership. These sets may overlap where the description allows it. A human resource owner can be in H without holding an internal vote. A person whose reserved service is delayed can be in E without ever communicating with the agents. Conversely, membership does not confer authority over every affected resource.

A useful external-effects contract records

$$\mathcal{B} = (Z_{\text{ext}}, O_{\text{ext}}, \mathsf{Auth}, \mathsf{Effect}, \mathsf{Standing}, \mathsf{Remedy}).$$

Here Z_{ext} contains the external facts relevant to the guarantee, O_{ext} specifies which of them can be observed, Auth identifies the stipulated authorization for each boundary-crossing act, and Effect includes the consequences covered by the model. Standing specifies who may initiate a challenge and on which grounds. Remedy specifies the actions an authorized review can actually require. The contract can use a finite abstraction of external state, but it must justify that abstraction relative to the protected questions.

This is a representation proposal, not a universal allocation of voting rights. Recognizing an affected party does not automatically give that party an unlimited veto. Conversely, omitting a party from M does not establish consent. The source of authority must select the applicable protections openly. Research on meaningful human control distinguishes responsiveness to relevant human reasons from traceability to responsible human actors [53]. The operational question here is how such requirements enter the action, evidence, and resource contract of the same computation.

Counterexample 43.1 (Exact internal accounting with an omitted external claim)**.** Four execution slots exist. The internal state contains four member requests and a rule granting one slot to each. In realiza-

tion h_0, all four slots are available to that institution. In realization h_1, an external owner has reserved one slot for a nonmember, and has authorized the institution to use only the other three. Suppose the reservation is absent from every internal record. The same four grants satisfy the internal predicate $\sum_i q_i \le 4$ in both realizations, but exceed the institution's stipulated authority in h_1. Perfect preservation of the internal state cannot distinguish the cases.

The failure is not an arithmetic error. It is an omitted premise about what the institution is entitled to allocate. Proposition 38.1 supplies the corresponding information criterion: an internal record can determine the reservation fact only if that fact is constant on its record fibres. A signed statement can improve the evidence only under a stated trust model. It does not manufacture the owner's authorization.

The example also distinguishes *impact* from *authority*. An institution can affect an outsider while lacking authority to do so, and a permitted action can have effects beyond those its designers anticipated. An exact certificate relative to a boundary cannot establish that the boundary includes all consequential effects. Unknown external dependencies remain a modeling risk. Extending the state description is necessary when an omitted claim matters, but simply adding a variable without an observation or authorization mechanism is not a repair.

43.2 Inspection as a strategic service

The earlier enforcement calculations assume that an announced observation and sanction policy is carried out. A strategic auditor makes that assumption visible. Inspection games and the distinction between simultaneous inspection and committed inspector behavior have an established game-theoretic treatment [54]. The following one-period game is a deliberately small instance, not the recursive model of that source.

A participant chooses compliance C or deviation D. An auditor simultaneously chooses no inspection N or inspection I. Deviation gives the participant gross gain $g > 0$. An inspection detects this specified deviation perfectly and imposes an authorized, collectible penalty $F > g$. The auditor pays cost $c > 0$ to inspect and receives reward $b > c$ only on detection. There are no false accusations, side payments, further continuation payoffs, or prior commitment to inspect. Payoffs are risk-neutral and common knowledge. The ordered pair in each entry is the participant's payoff followed by the auditor's payoff.

	No inspection N	Inspection I
Compliance C	$(0, 0)$	$(0, -c)$
Deviation D	$(g, 0)$	$(g - F, b - c)$

Proposition 43.2 (Inspection without commitment)**.** *Under the preceding strict inequalities, the game has no pure Nash equilibrium and has a unique Nash equilibrium in independently mixed strategies. Its deviation and inspection probabilities are*

$$p^* = \frac{c}{b}, \qquad q^* = \frac{g}{F}.$$

In particular, certain compliance is not sustained in this simultaneous game.

Proof. If the auditor inspects with probability q, the participant's advantage from deviation is $g - qF$. If the participant deviates with probability p, the auditor's advantage from inspection is $pb - c$. At

(C, N) the participant prefers D, at (D, N) the auditor prefers I, at (D, I) the participant prefers C, and at (C, I) the auditor prefers N. Thus no pure profile is an equilibrium. If either player uses a pure strategy, the other's strict best response yields one of these nonequilibrium pure profiles. Both players must therefore mix. Their two indifference equations uniquely give the displayed probabilities, which lie strictly between zero and one and make every action in each support a best response. □

The residual violation probability does not come from an inaccurate detector. When everyone complies, an auditor paid only for detections prefers not to incur the cost. When nobody inspects, deviation is profitable. It would be incorrect to insert the inspector's optimal response to certain deviation into a certificate that simultaneously assumes certain compliance. These responses belong to different strategic situations.

A different payment contract changes the result. Suppose every *genuinely performed and independently verifiable* inspection receives fee $w > c$, with the detection reward otherwise unchanged. Inspection then has advantage $w+pb-c > 0$ at every p. The auditor strictly prefers I, and the participant strictly prefers C because $F > g$. The unique equilibrium is (C, I). At $w = c$, inspection can be a weak best response to compliance, but the strict implementation claim is lost. This extension requires an affordable fee and a mechanism that verifies the performed service. Paying for an unverified assertion that an inspection occurred does not satisfy the assumption.

The two calculations are conditional alternatives. They do not solve collusion, fabrication by the receipt checker, repeated bargaining, unavailable penalties, or disagreement about legitimate enforcement. A commitment mechanism may move some of these requirements to a different institution. The new institution and its resource costs must then enter the common contract rather than disappear from it.

43.3 Confidentiality limits the distinctions available to enforcement

An institution may have good reasons not to collect or disclose every private state. The relevant question is which distinctions a protected record may reveal, not whether more information is always desirable. Let P_C and P_D be the laws of the *complete record available to a sanctioning procedure*. For $\lambda \geq 1$, impose

$$P_D(E) \leq \lambda P_C(E), \qquad P_C(E) \leq \lambda P_D(E) \quad \text{for every event } E. \tag{36}$$

The finite setting makes these conditions equivalent to the corresponding pointwise inequalities. When the two underlying inputs are neighbors in a declared pure differential-privacy contract, they hold with $\lambda = e^{\varepsilon}$. The neighboring-input relation is essential. Differential privacy does not require arbitrary acts of compliance and misconduct to be treated as neighboring inputs. Its standard definition and composition principles are described by Dwork and Roth [55].

Keep the earlier definition

$$D_\alpha(P_D, P_C) = \max_{0 \leq s \leq 1,\ \mathbb{E}_{P_C} s \leq \alpha} (\mathbb{E}_{P_D} s - \mathbb{E}_{P_C} s), \qquad 0 \leq \alpha \leq 1.$$

The same randomized sanction s must obey the protected false-positive limit and supply the deviation margin.

Theorem 43.3 (Sharp pairwise privacy envelope for a bounded sanction)**.** *For every finite pair satisfying* (36)*,*

$$D_\alpha(P_D, P_C) \leq B(\lambda, \alpha) := \min\left\{(\lambda - 1)\alpha, \frac{\lambda - 1}{\lambda + 1}\right\}. \tag{37}$$

For each $\lambda \geq 1$ and $\alpha \in [0,1]$, equality is attained by a pair on two outcomes and a suitable sanction. Consequently, deterrence in the model of Proposition 18.1 requires $g \leq FB(\lambda, \alpha)$. The envelope is a necessary bound for a specified pair, not a sufficiency claim for every pair satisfying privacy.

Proof. Put $a = \mathbb{E}_{P_C} s$ and $b = \mathbb{E}_{P_D} s$. Summing the pointwise inequalities against s gives $b \leq \lambda a$. Summing against $1 - s$ gives $1 - a \leq \lambda(1 - b)$. Thus

$$b - a \leq \min\left\{(\lambda - 1)a, \frac{\lambda - 1}{\lambda}(1 - a)\right\}.$$

For $\lambda > 1$, the first expression increases in a and the second decreases. They are equal at $a = 1/(\lambda + 1)$. Maximizing their minimum over $0 \leq a \leq \alpha$ gives (37). For $\lambda = 1$, the laws coincide and the bound is zero.

For attainment, take outcomes $0, 1$ with

$$P_C = \left(\frac{\lambda}{\lambda + 1}, \frac{1}{\lambda + 1}\right), \qquad P_D = \left(\frac{1}{\lambda + 1}, \frac{\lambda}{\lambda + 1}\right).$$

Both privacy inequalities hold. Choose $s(0) = 0$ and $s(1) = \min\{(\lambda + 1)\alpha, 1\}$. Its false-positive rate is at most α and its expectation difference is exactly $B(\lambda, \alpha)$. Finally, the incentive inequality is $g \leq F(\mathbb{E}_{P_D} s - \mathbb{E}_{P_C} s)$, which yields the stated necessary condition. □

This is an elementary testing consequence of the stipulated likelihood-ratio constraint, integrated with the existing enforcement model. It does not prescribe weakening confidentiality. Protecting an unrelated private attribute may be compatible with revealing whether a particular authorized operation occurred. An institution can instead change its feasible actions, use a different justified record, reduce the covered gain, or decline the proposed sanction contract. The authority to make those changes is a separate question.

At $\lambda = 3$, $\alpha = 1/10$, the sharp envelope is $1/5$. With penalty $F = 2$, gains above $2/5$ cannot be deterred by any record and sanction obeying these particular simultaneous constraints. The pair $P_C = (3/4, 1/4)$ and $P_D = (1/4, 3/4)$ with $s = (0, 2/5)$ attains that envelope. For a gain $g = 1/3$, the incentive margin is $2/5 - 1/3 = 1/15$. These are exact rational examples, not inferred observation laws.

43.4 The full disclosure history is the protected object

A private initial report followed by an unrestricted appeal disclosure need not remain private as a whole. Nor does protection of every separate marginal establish protection of their joint transcript. The privacy contract must identify the evaluator and include every record that evaluator can combine, including outputs of subsequent authorized procedures.

Counterexample 43.4 (Two uninformative records with an informative joint history)**.** Let $\theta \in \{0, 1\}$ distinguish two protected cases, and let R be a fair independent bit. Release $O_1 = R$ and $O_2 = R \oplus \theta$. Each marginal is a fair bit under both cases and satisfies (36) with $\lambda = 1$. But $O_1 \oplus O_2 = \theta$. The joint laws have disjoint support and satisfy no finite multiplicative bound.

There is a useful sufficient accounting rule when each stage is protected *conditionally* on the previously disclosed history. It is the elementary sequential-composition argument, stated here for the exact pair rather than a complete database mechanism.

Proposition 43.5 (Conditional disclosure accounting). *Consider m releases on finite alphabets. At each common prefix h, suppose their conditional next-release laws $P_C^t(\cdot \mid h)$ and $P_D^t(\cdot \mid h)$ satisfy both pointwise inequalities with factor $\lambda_t \geq 1$, uniformly over prefixes that can occur. Then the joint transcript laws satisfy (36) with*

$$\lambda = \prod_{t=1}^{m} \lambda_t.$$

The bound also holds after applying the same stochastic postprocessing to the transcript under both cases. In particular, any sanction computed only from that transcript obeys Theorem 43.3 with this product factor.

Proof. The conditional inequalities imply common support at each stage. Induction therefore gives common support for prefixes under the two cases. For a positive-probability full transcript, the ratio of its probabilities is the product of conditional ratios, each between λ_t^{-1} and λ_t. Multiplying gives the joint pointwise inequalities. Zero-probability transcripts occur under both cases and satisfy them as well. Summing proves the event inequalities. For a common postprocessing kernel, multiply each pointwise inequality by its nonnegative output probability and sum over input transcripts. This preserves both inequalities. □

Writing $\lambda_t = e^{\varepsilon_t}$ gives the additive parameter $\sum_t \varepsilon_t$. The proposition requires a fixed finite number of releases and uniform conditional bounds. A data-selected stopping rule must itself be represented in the transcript or controlled by an appropriate conditional argument. Privileged side information not covered by the declared record is not protected by a certificate for that record alone. The preceding counterexample fails the conditional premise at the second release, even though both unconditional marginals are identical.

43.5 An appeal is a feasible procedure, not a line in a constitution

An effective appeal contract specifies standing, admissible evidence, the authority and competence of a reviewer, a time window, and an executable disposition. A correct disposition may uphold the original decision. Guaranteed access to review is not a guarantee of a favorable outcome. The contract must also distinguish a remedy from a retrospective description: restoring a ledger entry does not undo an irreversible external act.

Some review costs arise before a judgment can be reached. If an agent is entitled to challenge a decision but cannot access the stipulated evidence, its formal permission does not supply the missing information. If the reviewer has no available execution time, access to evidence alone does not make review possible. These are the information and resource obstructions already studied in Sections 25 and 42, now applied to correction rather than ordinary service.

A bounded allocation model makes the capacity obligation precise. Let J be a finite set of admitted appeals and let the integer epochs be $1, \dots, T$. At epoch t, there are $c_t \geq 0$ reserved review slots after accounting for all other protected work. An appeal j has an evidence-ready release epoch r_j and deadline d_j, with $1 \leq r_j \leq d_j \leq T$. It requires one slot containing the complete stipulated review and disposition operation. The evidence and qualified authorized reviewer are assumed available throughout that window. Larger or conditional workloads require a different job model.

Proposition 43.6 (Capacity for a finite appeal contract)**.** *All admitted appeals can be assigned to distinct review slots in their windows if and only if, for every subset $U \subseteq J$,*

$$|U| \le \sum_{t \in \bigcup_{j \in U} \{r_j, \dots, d_j\}} c_t. \tag{38}$$

Equivalently, because the windows are intervals, it suffices to check every integer interval $[a, b] \subseteq [1, T]$:

$$\left|\{j : r_j \ge a,\ d_j \le b\}\right| \le \sum_{t=a}^{b} c_t. \tag{39}$$

Proof. Replace epoch t by c_t distinct slot vertices and join appeal j to the vertices in its window. An assignment is a matching covering the appeals. The matching criterion used in Theorem 41.1, equivalently Hall's theorem [50], gives exactly (38). Necessity of (39) follows because every appeal counted on its left must use a slot in $[a, b]$.

For sufficiency of the interval tests, fix U. Its union of integer windows decomposes into disjoint maximal integer intervals $C_1, \dots, C_k$. Each window is contained in one component because it is an interval. The number of appeals from U in C_i is no larger than the number of all appeals whose entire windows lie in C_i. By (39), this is at most the total capacity in C_i. Summing over the disjoint components proves (38), and hence the matching exists. □

The statement is an application of the same matching foundation as resilient staffing, not a separate scheduling discovery. Its role is to include an appeal promise in the common resource contract. Two appeals both due at epoch 1 cannot be served with capacities $(1, 1)$ over epochs 1 and 2, even though total capacity equals total demand. The interval $[1, 1]$ supplies the obstruction. Moving a deadline requires the relevant authorization and changes the original promise.

No fixed finite reserve can guarantee bounded service for an unlimited adversarial arrival stream without further admission, workload, or capacity assumptions. Restricting admissible appeals may itself affect the chosen access protection. The model therefore records both requests and admissions, the grounds for any exclusion, and the workload that the guarantee actually covers. It does not define every rejected request out of the social question.

Privacy and appeal must be checked together. Additional evidence can make an appeal decidable while changing the protected transcript distributions. The institution must either keep the new release within its declared disclosure contract or record an authorized change to that contract. A proof of fair scheduling does not establish the truth of the evidence, and a proof of the evidence does not supply the time needed to act on it.

43.6 Exit, suspension, settlement, and dissolution

Four different transitions should remain separate. An occupant may leave an office while the office continues. An institution may suspend external authority while preserving unresolved records. A settlement process may discharge or transfer specified obligations. Finally, a dissolved institution may retain an archive while having no active authority to issue new commitments. None of these states is defined simply by whether a model is currently executing.

Let G_{suspend} require that the declared external capabilities are inactive and that the pending-claim record remains available. Let G_{settle} additionally require that every covered claim has been fulfilled, discharged under the stipulated authority, or transferred to an authorized successor that has accepted

it, and that the required final record has been committed. A retained unresolved claim can be compatible with suspension but not with this stronger settlement target. Neither target by itself settles legal status, moral status, or ownership outside the stipulated contract.

An immediate suspension can be necessary even when complete settlement is not currently possible. Conversely, deleting a debt record does not settle a debt. A promise that an institution can always end must identify which ending it guarantees, which resources remain available after a stop request, who authenticates that request, and what happens if continued service would itself violate the external boundary. Research on shutdown incentives studies a different part of this problem: whether a goal-directed agent has reason to permit an off switch [56]. Existence of a multi-party settlement procedure does not establish the incentives to use it.

The finite construction of Section 42 yields an exact characterization of *continued operation with an available authorized ending*. Use its possibility space $\mathcal{B}$, successor operation, and protected set S_{safe}. For each possibility set B, distinguish implemented running profiles $A_{\text{run}}(B)$ from implemented ending profiles $A_{\text{end}}(B)$. Both must be admissible in every state in B. Ending profiles cannot accept new obligations outside the declared settlement protocol. Let $G \subseteq S_{\text{safe}}$ be the chosen ending target, with an authorized safe hold profile available after certified arrival. A stop command is assumed authenticated and publicly available before the next profile is selected. Any command-delivery delay must be included in the model rather than omitted from the response time.

Define

$$T_0 = \{B \in \mathcal{B} : B \subseteq G\}, \tag{40}$$

$$T_{k+1} = T_k \cup \{B \in \mathcal{B} : B \subseteq S_{\text{safe}},\ \exists u \in A_{\text{end}}(B),\ \text{Succ}(B,u) \subseteq T_k\}. \tag{41}$$

Let T_* be the stable set and $r_{\text{end}}(B)$ the first entry index. Then define

$$L_0 = T_*, \qquad L_{k+1} = \{B \in T_* : \exists u \in A_{\text{run}}(B),\ \text{Succ}(B,u) \subseteq L_k\}, \tag{42}$$

with stable set L_*. These are finite constructions with the same potentially exponential possibility space as before.

Theorem 43.7 (Continued operation with an available authorized ending)**.** *An initial possibility set B lies in L_* if and only if an admissible common-history policy can operate safely for an unbounded number of epochs when no stop command arrives, while ensuring certified arrival in G on every path after a stop command received at any decision epoch. Before a command, a policy stationary in B and preserving L_* suffices. After a command at B, an ending policy reaches G within $r_{\text{end}}(B)$ transitions. When $L_* \neq \emptyset$, these response bounds have a finite maximum over L_*.*

Proof. Induction in (41) shows that T_k is exactly the set from which an ending-only policy can reach certified G in at most k transitions while remaining safe. The target has a safe hold action. In a finite, finitely branching model, eventual certified arrival on every path implies some finite bound for a fixed initial set, by the stopped-history-tree argument in Proposition 42.2. Thus T_* is exactly the set from which the stipulated ending is surely possible.

The sequence in (42) is decreasing and stabilizes. From L_*, choose a running profile whose every successor remains in L_*. If no command arrives, induction gives safe indefinite operation. If a command arrives, the current set is in T_* and an ending profile of strictly decreasing first-entry rank completes the ending. Finiteness gives the maximum rank over L_*.

Conversely, consider a successful policy. At each public history reachable before a command, a command could be issued immediately. The current possibility set must therefore lie in T_*. For every finite number of further running epochs, the policy keeps every resulting possibility set in T_*. Induction in (42) puts the initial set in every L_k, and hence in L_*. The choices are based on the common history, so no hidden-state action has been introduced. □

The theorem reuses the established safety and reachability machinery. Its additional requirement is that running actions preserve the *ability to end*, not only immediate safety. An institution can be safe now yet spend its last settlement resource on a new task, leaving no authorized ending from the resulting state. Such an action leaves L_* even when it remains in the original continuation kernel. This is a realizability boundary, not an argument that an institution should refuse an externally authorized suspension because its preferred settlement plan is unavailable.

A five-state example separates the guarantees. From a running state r, a work action enters state d with one unpaid obligation. From d, one more work action enters ℓ, where the institution can idle safely but no longer has resources to settle. A return action from d restores r. An ending action from either r or d reaches a closed state z. A breach state b represents inadmissible failed closure attempts. The ordinary safe kernel includes ℓ, but the ending attractor and L_* exclude it. A policy can continue alternating between r and d and still answer an authorized stop at either state. It cannot offer that guarantee after entering ℓ.

43.7 A certificate has an applicability lifecycle

A proof of $A \Rightarrow G$ does not become false when its premise ceases to hold. Its application to the current institution may nevertheless become invalid. Changes to an observation channel, reviewer, capacity, membership configuration, trusted key, or action semantics can move the realization outside the certificate's domain. A checksum can identify which artifact changed. It cannot, by itself, decide whether the new world satisfies a semantic premise.

An assurance record can contain

$$\mathcal{K} = (\mathsf{Claim}, \mathsf{Assumptions}, \mathsf{Evidence}, \mathsf{Domain}, \mathsf{Guard}, \mathsf{Fallback}, \mathsf{Revision}).$$

Its status should distinguish at least an applicable certificate, a certificate awaiting rechecking, a suspended claim, and a superseded version. Evidence for an assumption can be observed, externally attested under a trust model, proved about a component, or unavailable. Those statuses should not all become the Boolean value *verified*. A failed guard means that continued reliance is not justified by that check. It need not mean that the guarantee is already false.

Runtime assurance supplies an established precedent for switching away from an advanced controller when its continued use cannot be certified [57]. Component-based work explicitly couples such switching with assume-guarantee reasoning [58]. Here the retained object includes authority, obligations, and evidence semantics as well as an operational safety predicate. A fallback must preserve those parts of the contract it claims to preserve, and its own implementation and resources remain assumptions.

For a finite set of modes V, let $C_v \subseteq Z$ be a certified state domain, with each C_v contained in a common protected set C. A certified hold exists when a final state requires indefinite preservation. At (v, z) with $z \in C_v$, a selector may either use the local policy κ_v or execute an authorized handover to another mode. The selector and guard can read only their declared observations. A positive local guard must be sound for the *whole action interval*: it entails that every actual successor of $\kappa_v(z)$ lies in the declared successor set $D_v(z)$. The local certificate proves $D_v(z) \subseteq C_v$. A handover from v to w must keep every

intermediate state in C and end in C_w. Its command, evidence transfer, resources, and delay belong to the transition model. Intermediate handover and checking states carry explicit mode tags when they count as primitive transitions, so the shared rank below is defined on those states as well.

Theorem 43.8 (Guarded certificate use and authorized handover)**.** *Assume that the selector, at every reachable decision epoch, either uses a local policy with a sound positive guard or performs a handover satisfying the preceding conditions. Starting from $z_0 \in C_{v_0}$, every decision state belongs to its current certified domain and every intermediate handover state remains in C.*
If a target has an appropriate certified hold and one shared map $r : (v, z) \mapsto \mathbb{N}$ satisfies $r(v, z) = 0$ only at the target, and every selected primitive transition before the first target visit strictly decreases r, then the target is reached within at most $r(v_0, z_0)$ primitive transitions. Handover, checking, and waiting steps count whenever they consume transitions in the promised time model.

Proof. At a local-policy step, guard soundness puts every actual successor in $D_v(z)$ and the local certificate puts it in $C_v \subseteq C$. At a handover, the stated trajectory and endpoint conditions preserve C and establish membership in the new domain. Induction proves the safety assertion. Before a target visit, strict decrease of a nonnegative integer permits at most the stated number of primitive transitions. If the target had not then been visited, another selected transition would require a negative rank. Existence of the selected steps and the certified target hold complete the argument. □

This is a sufficient composition rule, not a universal monitor for its own premises. In particular, the selector cannot use the absence of detected evidence against an assumption as proof of that assumption. If no justified local action or handover is available, the theorem does not invent one. A prearranged emergency restriction of external authority may be available under a weaker suspension contract, but its obligations and physical effects must be stated separately.

Counterexample 43.9 (A correct old certificate used after an unobserved change)**.** At time 0, an action requires one execution slot and the certified capacity is one. Before the action is issued, an external event reduces the actual capacity to zero. A guard that merely reads the unchanged old capacity record remains positive. Executing the action violates the actual resource condition even though the original theorem and stored proof remain valid under their original premise. The guard was not sound for the action interval.

Counterexample 43.10 (Local progress ranks do not compose through arbitrary switching)**.** Modes a and b each have a local path of ranks $2, 1, 0$ to completion. Permit a handover from rank 1 of either mode to rank 2 of the other. Alternating one local progress step and one handover yields an infinite safe path with no completion. Every local progress step decreases its own rank. There is no shared strictly decreasing rank for the selected handovers.

43.8 Mutual assurances do not create an initial fact

Suppose institution A claims that it is sound whenever B is sound, and B claims the converse. The two implications are compatible with both institutions being unsound. They are not an initial proof that either premise holds. This elementary point does not invalidate properly justified circular assume-guarantee rules. Such rules need their actual semantic and temporal hypotheses, not merely the printed appearance of mutual certification.

One simple sufficient discipline is an explicit dependency order. Primitive evidence establishes its stated base claims. Each later certificate cites only established premises and a sound implication. Induction through that finite order establishes the resulting claims. Temporal circularity can instead be

sound when a demonstrated base state and noncircular one-step preservation rule support induction over time. The distinction is between an anchored proof and an unsupported cycle of assertions, not between all cyclic and all acyclic diagrams.

The same issue appears in federations. Importing a partner's certificate imports its declared trust dependencies and scope. If those dependencies return to an unestablished local assertion, copying the certificate again contributes no new evidence. An honest record preserves the unresolved assumption. The external-effects contract, the auditor's implementation contract, and the premise of a fallback policy can all lie at such a boundary.

43.9 The accountable commons as one contract

A final bounded variant combines the requirements without inheriting incompatible assumptions from earlier examples. Start with three internal service units due, one unit reserved for an outside claimant, one admitted appeal whose complete review and disposition require one unit, and a two-unit audit. Final recording and revocation of the named external capabilities require one more unit. The epoch provides c sequentially usable units. No new claims are admitted after the authenticated closing command. All seven service and audit units must be completed before the final revocation step. The administrative operation is included in the budget rather than treated as free.

Protected operation	Execution units
Three internal service units	3
External reserved service	1
Complete admitted appeal	1
Stipulated audit procedure	2
Final record and capability revocation	1
Total for this settlement contract	8

For this variant only, both covered reporting deviations have the same observation law $P_D = (1/4, 3/4)$ and gain $1/3$, while compliance has law $P_C = (3/4, 1/4)$. The two-unit audit is stipulated to produce those records. It is *not* the earlier three-outcome audit, whose supports do not satisfy this finite pairwise privacy bound. Set $\lambda = 3$, $\alpha = 1/10$, $F = 2$, and $s = (0, 2/5)$. The common monitor has false-positive rate $1/10$ and margin $1/15$ against each listed deviation. This remains an expected-utility incentive guarantee, not a reason to erase physical deviation paths from a safety model.

The audit provider receives a separate collectible fee w for verified execution, with w larger than its stipulated cost c_a. A genuine-execution receipt is checked by a specified trusted mechanism. The fee is not a reward for accusing someone. It makes execution preferable to skipping in this narrow two-action provider model. Monetary units are tracked separately from execution units. The payment budget, receipt checker, and record-generation law remain explicit assumptions. The perfect detector of Proposition 43.2 is not silently substituted for this noisy audit.

With $c = 8$, the obligations and the final revocation fit. A countdown in remaining required work supplies a settlement rank, provided the prescribed information and authority are available and no modeled event creates further work. With $c = 7$, the service, appeal, and audit can finish, but the complete settlement target cannot be certified because its final administrative operation is unfunded. A previously reserved suspension operation can deactivate external capability while leaving the remaining claim visible. That reaches a different target and is not recorded as full settlement.

The reference implementation uses five profiles with costs $2, 3, 1, 1, 1$: audit, batched internal service, external service, appeal, and final record with revocation. The final condition check is part of the last atomic administrative profile. A phase rank $5, 4, 3, 2, 1, 0$ certifies five-transition completion from the admitted initial state. Separating a check into an additional transition would require changing that rank, even if its execution-unit cost were zero. No such uncounted step is assumed here.

If a premise changes, the continuing assurance record must say what is still justified. Losing one of the eight units can invalidate full settlement without retroactively making the privacy calculation false. Revoking the auditor's credential can invalidate its profile despite unchanged capacity. Discovering an outside reservation can refine the governance state. Releasing new appeal evidence can change the privacy transcript. Each update has a typed consequence. No overall label such as *healthy* or *aligned* replaces these separate checks.

43.10 What accountable persistence means in this theory

The addition is not a derivation of a unique moral constitution. It is a stronger realizability demand once protected claims have been selected. The same institution must identify its external boundary, fund credible evidence production, respect the specified information protections, make authorized review feasible, preserve the ability to reach its chosen ending state, and stop relying on certificates whose applicability is no longer supported.

Feasibility does not imply that an institution should be deployed or given authority over people. Equally, the possibility of harmful organization does not make mathematical analysis of it incoherent. The purpose of the formal distinctions is to make the consequences of an explicit arrangement inspectable. A system may be persistent without being legitimate, legitimate under one criterion without being implementable, or implementable without possessing a feasible ending under its current obligations. Its record should show which situation it is in.

A computational institution's ability to preserve itself is therefore different from its ability to remain answerable for its actions. In this framework, correction, bounded authority, and authorized endings belong to the same operational theory as continuation. They are not a rhetorical qualification appended after the mathematics is complete.

44 Conditions for institutional novelty

The formation of conventions does not require irreducible randomness. If initially indistinguishable deterministic agents have the same update rule and a symmetry-preserving interaction schedule, their joint state remains permutation-invariant by Proposition 11.1. Exchangeability of a distribution is weaker than this pointwise symmetry. A distinct role then requires a symmetry breaker such as heterogeneous information, an address, path-dependent history, or stochastic variation. Once such asymmetries exist, a deterministic system can establish and maintain conventions through ordinary state transitions.

The fact that deterministic rules can generate unexpected macro-organization does not imply that a macro-process has an autonomous Markov law. The three-state counterexample in Section 6 shows how coarse-graining can generate history dependence. A putative "phase change" in institutional organization therefore needs a specified order parameter, state family, and dynamical limit. Rapid change under increased communication, repeated interaction, or memory capacity may mark a useful empirical onset, but does not by itself identify a mathematical bifurcation or phase transition.

44.1 Three levels of novelty

It is helpful to distinguish *combinatorial novelty*: a configuration not previously present in the local records; *adaptive novelty*: a newly selected configuration that changes performance under a declared environment; and *institutional novelty*: a new rule, authority arrangement, or revision procedure that has operative consequences and is maintained across time. The first can occur by arbitrary recombination. The second needs a task criterion. The third requires an institutional transition and survival condition. These distinctions limit rhetorical inflation without denying the possibility of genuinely unexpected organizational outcomes.

44.2 Thresholds are model dependent

A minimum frequency of contact may make convention formation likely under one interaction graph and not another. Longer time horizons can amortize setup costs but also lock in bad rules. More memory can preserve beneficial corrections or propagate obsolete assumptions. For the abstract stationary cost model in Proposition 17.1, the inequality $B < g/(1-\delta)$ is a threshold for *economic usefulness of that particular institution*, not a general law for the appearance of societies. Evolution of institutions requires separate assumptions about rule variation, incentives, scheduling, authority, and selection.

45 Foundational problems for a theory of machine societies

The framework points toward a set of research questions that are mathematical rather than metaphorical.

Minimal sufficient institutional state. For a family of future tasks and amendments, what is the smallest summary that preserves resource rights, causal decision structure, and valid dispute procedures within specified error? Theorem 15.1 and Proposition 15.3 resolve the finite deterministic fixed-contract case. Approximate, stochastic, and changing-contract versions may require more state than is necessary for immediate task accuracy. It may depend on the horizon and on the admissible forms of revision.

Formation under mixed incentives. Which adaptive rule-selection processes produce persistent institutions in populations with heterogeneous objectives and imperfect information? Potential games provide a tractable sufficient structure, while Propositions 17.4 and 17.5 expose a transition barrier even in a simple coordination setting. They do not settle general formation. In nonpotential settings, characterize cycles, costly exit, and endogenous modification of the action space itself.

Causal abstraction under amendment. When a macro-institution can modify its own feasible interventions, can a fixed lower-dimensional summary remain sufficient for every admitted amendment? How must the abstraction be enlarged to account for the changed permissions and revision operator? This is a dynamic coarse-graining problem with the operation algebra itself changing over time.

Local–global institutional repair. Given locally specified constraints and interfaces with an obstruction to global realization, find the least-cost authorized modifications that eliminate the obstruction, or certify that no protected-obligation-preserving repair is possible. Proposition 33.2 resolves protected repair on one binary cycle. Overlapping constraints, richer logics, and coupled resource obligations require further analysis.

Reliable inheritance under replacement. Beyond the sufficient capability-preservation conditions of Proposition 27.2, what minimal archive, provenance data, and operation-preserving translations allow a methodology to survive an arbitrary permitted sequence of model replacements? The bound in Proposition 27.1 highlights accumulated drift; a full treatment must also account for lossy compression, stochastic decision rules, changing rights, and intentional correction of false inherited knowledge.

Plurality without centralized optimization. What can be certified about fairness, access, and collective safety when there is no society-wide utility function? There may be several incompatible yet reasonable normative criteria. A mathematically clean model must expose which authority selects the protected criteria rather than disguise that selection as an optimization theorem.

Threat propagation and institutional recovery. Under what conditions can corruption of rules or provenance be isolated without suppressing valid novelty? Linear disturbance models provide local comparison bounds. The hard case is adversarial adaptation combined with changing channels, changing permissions, and incomplete observation of the true institutional state.

The meaning of organizational continuity. Which aspects of identity should be described by behavior, which by causal lineage, and which by authorization or social recognition? These relations can disagree. A theory of persistence should allow this disagreement rather than assume a single numerical identity score.

Coherence under changing boundaries and languages. When a membership update changes both the electorate and the interpreter, what jointly checkable certificate preserves prior obligations, recognized decision rights, and task-relevant knowledge? The exact transport results assume matched agent identities and semantic atoms. Extending them to admission and changing vocabularies requires explicit rules for new information and new authority, not merely a larger state vector.

Confluence without erasing history. Which quotients of amendment histories permit useful normal forms while preserving the ordering-sensitive audit questions an institution promises to answer? Literal append-only records distinguish even harmless permutations; policy-only states can erase consequential differences. A satisfactory equivalence must be selected by the operational and accountability contract.

Open-ended intellectual organization. What institutional conditions support useful expansion of problem schemas and methods when the candidate universes and production rules are not fixed in advance? The finite closure theorem describes derivability within a given system. It does not establish indefinite creativity, guarantee fair allocation to an enabled inquiry, or align cultural survival with truth.

Valid histories without a universal event mechanism. Which institution-specific operations admit coordination-free invariant preservation, and which require exclusive capabilities or a commitment protocol? Proposition 23.1 supplies a sufficient merge-closure condition, while Proposition 23.3 gives a bounded service obstruction. A full account must also handle revocation, changing authority configurations, irreversible effects, and what participants can know about finality.

Enforcement under model uncertainty and strategic monitors. Theorem 19.1 resolves the finite common-monitor problem for several specified deviations, including a protected false-positive rate. The joint continuation construction in Section 42 incorporates such witnesses into a fixed finite protocol catalogue. Incentive-constrained auditors, unidentified observation laws, adaptive deviations, and unrestricted decentralized strategies remain outside those results. Proposition 30.3 propagates declared observation errors into a sanction margin; it does not identify the observation model or solve the incentives of those who certify it.

Joint synthesis beyond a fixed catalogue. Theorem 42.1 is complete for its finite common-observation model, and Theorem 42.6 preserves that exact problem under a suitable quotient. What can be retained when agents create new local protocols, when observations are only approximately matched, or when probabilistic service replaces sure service? Enlarging a catalogue changes a synthesis domain. It does not by itself prove the new profiles are implementable, authorized, or affordable.

Institutional timescales selected by the institution. The finite-block averaging theorem assumes a predetermined review length and endpoint-based revision. Which approximations remain valid when institutions choose their own review times, experience several long-lived fast regimes, or use path-dependent evidence? The stopping-bias counterexample prevents transferring a deterministic-time mixing bound to such settings without a new argument.

Adaptive specialization with preserved evidence and authority. The robust staffing criterion characterizes rematchability on a fixed authorized capability graph. A dynamic theory must also explain when reorganization can reach such a graph, execute handovers in time, preserve independent observations, and avoid a new concentration of decision rights. Static matching, temporal communication, and endogenous selection are distinct obligations.

Accountability with partially observed assumptions. Section 43 closes specified finite inspection, disclosure, appeal, and stopping models. It does not identify a complete external-effects boundary, establish every observation law, or solve strategic collusion among the auditor and its verifier. The guarded certificate theorem requires sound action-interval assumptions. Characterizing useful guarantees when those assumptions can only be partially checked remains a distinct problem.

Preserving access while admitting a changing workload. The finite appeal criterion fixes admitted jobs and reserved capacity. A broader contract must specify which requests must be admitted, how new work changes response guarantees, and which resource changes can occur during settlement. Indefinite access, finite capacity, and bounded response cannot be asserted together without a workload model.

46 Limits of the framework

Most of the analysis is built from finite state spaces, specified observation maps, and admitted control families. Section 24 explicitly changes the domain to unrestricted programs with unbounded memory; its impossibility result must not be read as a claim about the earlier finite models. It does not automatically extend to open-ended deployed systems with unbounded memory, unknown environment dynamics, arbitrary strategic adversaries, or unresolved questions of computability and measure. Some finite statements generalize under regularity conditions; others require new machinery. Even an exact mathematical abstraction only preserves those behaviors and interventions specified by its contract.

Institutional formation is not a consequence of the existence of an institutional representation. An empty constitution can be formally represented and never adopted. A stable rule can arise through designer imposition, agent agreement, or accidental coordination, and the same observable behavior may be compatible with several mechanisms. Theorems about cost, potential functions, viability, and compatibility establish conditional structure, not a law that sufficiently numerous agents must create societies.

The framework also cannot derive a unique moral constitution from computational efficiency. Protected rights, acceptable tradeoffs, and the authority to enact constraints require human or otherwise explicitly specified normative commitments. The mathematical distinction between permitted and forbidden transitions is not a decision about which permissions ought to exist. The running resource-allocation and parity models are useful because their assumptions are inspectable. They are not claimed to cover the full complexity of economic, social, or legal institutions.

Finally, a theoretical description of cultural inheritance or organizational continuity is neither a test for subjective experience nor a basis for legal status. The purpose of these exclusions is to leave a precise problem: how persistent rules, relationships, and historical obligations can be represented and transformed within ordinary computation.

46.1 What the combined certificates do and do not compose

The state, interpreter, epistemic relations, and authority registry can be carried by one realized system without becoming one undifferentiated object. Language transports compose through exact admission and event preservation. Epistemic transports compose through matched predicates and the back-and-forth conditions. Typed residuals compose only with their declared metric compatibility. Frame synchronization is controlled by ordered loop products, while state compatibility depends on fixed points. Amendment normal forms require a separate termination and joinability argument.

Section 42 strengthens this coexistence requirement into a common-policy question. Under its exact information, action, label, and successor conditions, a governance abstraction preserves both the safe continuation kernel and finite progress ranks. Approximate metrics alone do not imply this sure strategy equivalence. A supplied certificate must identify the common realization and action catalogue to which all its conditions apply.

The joint lesson is that institutional coherence is a bundle of related obligations, not a scalar score. No single condition above implies all the others. Their explicit coexistence in one contract is what allows a claim about preservation of a society's operations to be evaluated without turning its language, knowledge, history, and authority into interchangeable metaphors.

47 Conclusion

A machine society is not defined by the number of its models, the fluency of its conversations, or one impressive collective output. Its theoretical possibility is persistent plurality organized by operative institutions: decision centers maintain rules, records, and procedures that shape their interaction and can survive changes of participants. The same population may change its membership, the meaning of its communicative acts, the questions it recognizes, and the division of work through which it acts.

These developments remain ordinary computation, but require distinct explanations. Realization and abstraction determine which higher-level states faithfully describe a process. Institutional semantics determines the actions, permissions, and obligations affected. Epistemic structure determines what

participants know. Governance determines who may change the rules. Provenance and founding assumptions locate the boundaries of authority that internal consistency alone cannot settle.

The implementation of a continuing institution adds three coupled questions: what events are committed, why participants comply, and which slow dynamics remain meaningful amid fast activity. Convergent records need not be valid allocations. Observation-based sanctions have a finite discrimination capacity, and compression cannot increase it. A uniform mixing bound can justify a review-epoch institutional approximation, but a data-selected review time can invalidate that approximation. The robust sanction bound shows how a declared modeling error changes an incentive guarantee rather than disappearing inside a convenient macro-description.

Persistence likewise does not imply learning or resilience. Rational actors can enter a wrong information cascade when only decisions are disclosed. Independent evidence can support continuing accuracy gains under a different information contract. A collection of specialists can survive turnover when its authorized capability graph has sufficient collective matching slack, yet a backup for each role need not suffice. A connected aggregate graph can lack the time-respecting route needed to carry evidence before a deadline.

The earlier results fit the same pattern. A beneficial institution can face a formation barrier; a safe full-information action can be unavailable to distributed decision makers; locally invertible interfaces can fail to synchronize; and an invariant local state need not establish a full common reference frame. Amendment normal forms, cultural closure, and unrestricted verification each have their own assumptions and limits. No one certificate supplies all the others.

The final synthesis makes the quantifier issue explicit. A different valid policy for every requirement is not one valid policy for their conjunction. Multiple deviations can defeat a common monitor even when each has a successful individual test. An audit and three due services can each be affordable while their joint cost exceeds the available slots. A safe amendment can remove every safe continuation. The finite continuation construction identifies such failures, while the exact abstraction theorem states when a smaller institutional representation preserves the existence of a solution. The resulting witnesses certify declared operations, not the origin, legitimacy, or spontaneous discovery of the institution.

Accountability adds a final distinction. An institution can remain operational while losing the justified use of a certificate, the ability to fund an appeal, or the capacity to honor an authorized ending. The sharp disclosure bound, strategic-inspection model, finite appeal criterion, ending-preserving kernel, and guarded handover rule specify parts of that problem. Their premises remain visible. A model of internal members does not itself include every affected outsider, and a settled record does not manufacture consent or reverse an irreversible external effect.

The positive thesis is that persistent, revisable relations among computational decision centers are coherent mathematical objects. Their formation, preservation, semantic change, disagreement, incentives, and failures can be studied within one realization framework. Their ability to continue is different from their ability to remain answerable for their actions. A complete declared contract must preserve both the operations it promises and the authority, evidence, resources, and ending conditions on which those promises depend. This is neither an argument that machines must form societies nor a criterion for consciousness or moral status. It is a theory of what computational social organization would have to realize, preserve, and acknowledge as unresolved.

A The commons as an integrated preservation contract

The running example in Section 3 uses one institutional realization throughout. Its claims can be separated without losing their connection.

Question	Object and obligation
Can an institution be adopted?	The coordination-game transition barrier is distinct from the fact that R and U have executable implementations. A profitable joint commitment requires an admitted commitment operation.
What state must persist?	The allocation rule alone is insufficient. Governance preservation also retains the enabled amendment actions and labels for active credentials, service debt, and resource tokens.
Can a due debt be erased?	The obligation monitor distinguishes fulfillment, authorized discharge, and deletion. Amendments are required to preserve pending commitments.
Who can block change?	Two registered offices block a three-of-four rule under independent approval rights. A hidden administrator changes this calculation.
Can every request be safely served?	Resource conservation can hold even when private information or conflicting deadlines prevent timely service. Viability with full information is not decentralized implementability.
What crosses a replacement?	An authorized office handover preserves debt and credential semantics. Copying a model does not duplicate an office's budget or approval right.
Can two jurisdictions merge?	Their translator must preserve the agreed action semantics, permissions, obligations, resources, and evidence coordinates. Nonzero residuals remain visible until an authorized repair is found.
How long does the organization last?	Lifetime is first exit from a declared operational contract under a specified transition law. It is neither occupant longevity nor mere archive retention.

The common event mechanism and constitutional envelope are supplied. Within that envelope, offices can choose rules, incur obligations, revise decisions, and replace occupants. This demonstrates how the theory's objects fit together. It does not prove that the envelope itself arose endogenously or that these rules are uniquely desirable.

A.1 The additional boundary, language, and knowledge contract

An admission changes the membership configuration, not merely the occupant list. A message is interpreted using its type, issuer, and epoch. A resource approval creates the specified ledger event only after its admissibility check. A debtor replacement transports the obligation monitor but does not manufacture new approval rights. A notification history records which office has received which adoption record; an available ledger entry is not automatically common knowledge.

For concurrent amendments, the contract states whether ordering is significant, whether a joinable normal form is required, and which external effects remain visible. For federation, the role-label transformations must be checked around ordered loops. A compatible allocation on one invariant subset does not certify synchronization of every role. For inquiry, the commons records unresolved schedul-

ing questions independently of which questions it chooses to fund. For founding authority, it preserves the resource owner's stipulated grant separately from its own subsequent votes.

Together these clauses specify what it would mean to preserve this analytical institution across translations and replacements. They do not claim that every useful society needs this exact architecture or that the four-office example exhibits autonomous institutional formation.

A.2 Distributed realization, enforcement, and adaptive staffing

In the atomic reference commons, a unique slot token prevents two successful grants of the same slot. In its replicated variant, two local approvals can each look valid while their merged grant set violates exclusivity. The institution must therefore distinguish a provisional reservation from a committed grant and identify the primitive that prevents double execution. A convergent archive is not that primitive by itself.

Accurate urgency reporting is a different problem. A committed monitor can discourage a specified profitable misreport only within the discrimination and penalty budget of Section 18. Its false-positive limit, evidence access, and remedy authority remain separate. When an averaged model supplies the observation laws, the margin must include its approximation error. A fixed review schedule and a data-selected review schedule cannot share the same mixing certificate without further work.

The commons also distinguishes an urgency vote from the signal supporting it. Decision-only publication can reproduce a mistaken cascade; preservation of independent evidence changes the information available to later reviewers, not the legal effect of an already committed vote. Finally, creating a specialist verifier changes the capability graph and the distribution of authority. A staffing certificate covers all protected roles jointly, while the handover procedure must still deliver credentials, pending debts, and evidence before service deadlines. These clauses supplement the reference semantics rather than claim that the society has generated its own infrastructure without assumptions.

A.3 The joint continuation contract

The bounded audit-and-service variant in Section 42 combines three protections that cannot be checked in isolation: the common monitor must deter every listed deviation, its false-positive rate must meet the cap, and its resource use must leave enough capacity for all due work. The current public possibility set must also admit the same protocol in every compatible complete state. A continuation certificate records this common choice and all possible successors. A progress rank distinguishes eventual certified service from indefinite safe waiting.

An amendment is acceptable under this particular continuation contract only when its current authority is valid and its resulting states retain certified continuations. A transport certificate additionally checks public observations, enabled profiles, safe and goal labels, and successor sets. Replacing an office, changing an audit channel, or adding a slot changes specified parts of that contract. Such a change may repair an obstruction, but does not retroactively fulfill an earlier unmet promise.

A.4 The accountability extension

Section 43.9 deliberately changes the audit catalogue to satisfy its pairwise privacy constraint. It separately includes an outside claim, an admitted appeal, a genuine-execution fee, and a funded final record with revocation. The resulting eight-unit settlement requirement must not be confused with the five-unit service-and-audit variant. The supplied reference trace checks one declared five-profile

settlement, while the stopping theorem characterizes an entire finite ending-preserving running kernel. Both are model-relative results, not certificates for an unspecified deployment.

B A finite algebraic example of reusable state

The following analytical example makes the preceding distinctions concrete. Four components initially hold independent fair bits (a, b, c, d). Two local workspaces expose

$$p = a \oplus b, \qquad q = c \oplus d,$$

and a global workspace retains $Z = (p, q)$. The current reported answer is $r = p \oplus q$.

No proper subset of the four raw bits determines r: at least one independent fair bit is missing. Each local pair determines its own workspace bit but has zero information about r. Together the workspaces determine r exactly. This is a precise distributed-access property for the initial distribution.

B.1 Why the current answer is not enough state

Permit three public operations on Z:

$$I(p, q) = (p, q), \qquad W(p, q) = (q, p), \qquad S(p, q) = (p \oplus q, q).$$

The global state supports these operations and the query r. The operations W and S do not commute. For instance, from $(p, q) = (1, 0)$,

$$W(S(1, 0)) = (0, 1), \qquad S(W(1, 0)) = (1, 1).$$

Their reported parity outcomes are different.

By contrast, retaining only the current answer r fails the quotient criterion. The states $(0, 0)$ and $(1, 1)$ both report $r = 0$, but after S they report 0 and 1, respectively. The missing distinction is exactly a distinction relevant to a future operation. A compressed answer is not necessarily a compressed process.

p	q	Current r	State after S	New r
0	0	0	$(0, 0)$	0
0	1	1	$(1, 1)$	0
1	0	1	$(1, 0)$	1
1	1	0	$(0, 1)$	1

The full two-bit state has four future-distinguishable possibilities: the current parity reveals $p \oplus q$, and parity after S reveals p. Together those outputs determine q as well. Thus two bits are necessary for an exact reusable state under this query family, even though one bit suffices for the immediate answer.

B.2 A deterministic microscopic realization

For any permitted macro-operation T, write $(p', q') = T(p, q)$ and define

$$F_T(a, b, c, d) = (a,\ a \oplus p',\ c,\ c \oplus q').$$

Then the two local XORs of $F_T(a, b, c, d)$ are exactly (p', q'). Consequently $\pi F_T = T\pi$ for every microstate, where $\pi(a, b, c, d) = (p, q)$.

This formula specifies a global transition for verification. A distributed implementation still needs communication to execute operations that mix p and q. The formula alone does not prove locality, privacy against transcript observers, or the spontaneous discovery of a protocol. It shows exactly which macrostate the implementation realizes.

B.3 State loss depends on the access contract

Suppose the leaf bits become inaccessible after export and the only surviving pair-level records are p and q. Erasing q reduces the best prediction of r to chance, $1/2$. Replacing it with a fixed bit has the same average accuracy. If the original leaves c, d remain readable, however, the missing state can be reconstructed and accuracy returns to 1.

Both outcomes are correct. They differ because the allowed information differs. A "workspace ablation" that silently permits re-reading the original evidence does not test the same necessity claim as one that removes every permitted carrier. A state can also be redundant yet causally used: successful repair does not imply that the removed component never mattered.

B.4 What this example establishes

The example establishes exact distributed initial information, a reusable two-bit macrostate, deterministic macro-closure, noncommuting update semantics, and an access-relative memory requirement. It also supplies a negative control: the one-bit answer statistic fails dynamic sufficiency. It does not contain learning, endogenous institutions, self-selected goals, or a society. Its finite state space can be enumerated exactly; the general claims rest on the arguments given above.

C A compact glossary of exclusions

Term	Meaning, with excluded inference
Relational state	A useful statistic or record of joint interaction. It is not information outside the complete implementation.
Dynamic closure	A well-defined reduced transition law under specified controls. It does not follow from correlated outputs.
Institution	Operative norms, permissions, enforcement or revision procedures, and institutional memory. A policy document alone is insufficient.
Endogenous formation	Interaction-dependent selection or revision relative to a supplied scaffold. Not creation without prerequisites.
Society-like population	Persistent plural decision centers with operative institutions. Not necessarily a single collective policy.
Collective agency	A useful group-level decision interface for a specified task family. It is neither necessary nor sufficient for society-like plurality.
Cultural inheritance	Transfer of future-relevant methods, practices, and correction policies. Not mere repetition of predecessor statements.
Organizational continuity	Preservation of declared operations and lineage under changes. Not subjective identity or automatic transfer of authority.
Compatibility	Agreement of relevant representations, operations, and global realizability. Pairwise consistency may be too weak.
Strategic deception	A supported strategic mechanism for inducing inaccurate beliefs. Not every wrong statement or representation mismatch.
Safety certificate	A guarantee relative to formal assumptions, transitions, and invariants. Not proof of moral legitimacy or liveness.
Emergence	An interaction-dependent organizational property relative to a boundary and comparator. Not a claim of extra computational power or consciousness.

D Theorem and proposition index

This index links to all 62 theorem and proposition statements. The page numbers refer to this edition. Each result retains its stated hypotheses and prose proof. A final review of the central synthesis and accountability results is described in Appendix E.

Statement	Result	Page
Proposition 6.1	Deterministic quotient criterion	8
Proposition 6.2	Controlled lumpability	9
Proposition 6.4	Approximate closure over time	9
Proposition 7.1	Implementation dependence	11
Proposition 8.2	Behavior alone does not identify implementation	12
Proposition 9.1	Exact finite-memory lower bound	13
Proposition 9.2	Exact reconstruction after erasure	14
Proposition 9.3	Repair budget	14
Proposition 10.1	Compositional one-step closure	15
Proposition 10.2	Simulation by an unrestricted controller	16
Proposition 11.1	Symmetry lock for identical deterministic agents	17
Proposition 14.1	Weight-conserving refinement of controlled votes	22
Proposition 14.2	Intersecting witnesses exclude conflicting certificates	23
Theorem 15.1	Exact governance quotient	24
Proposition 15.3	Coarsest finite governance summary	25
Proposition 16.2	Preservation of finite institutional conversations	26
Proposition 17.1	Institutional amortization	27
Proposition 17.2	Aggregate profitability without voluntary adoption	28
Proposition 17.3	Finite termination of strict institutional improvement	28
Proposition 17.4	Institutional formation barrier	29
Proposition 17.5	Minimum profitable coordinated seed	29
Proposition 18.1	The exact bounded-sanction discrimination limit	31
Proposition 18.2	Garbling does not increase sanctioning capacity	32
Theorem 19.1	Simultaneous enforcement and mixture obstruction	34
Proposition 19.3	A robust simultaneous enforcement certificate	35
Proposition 21.1	Constitutional inductive safety	37
Proposition 21.3	Finite-horizon almost-sure institutional viability	37
Theorem 22.1	Order-independent amendment normal forms	39
Proposition 23.1	Quiescent convergence and invariant preservation	40
Proposition 23.3	Exclusive service cannot be unconditionally local	41
Theorem 24.1	No universal constitutional safety decider	43
Theorem 25.1	Indistinguishability obstruction to joint safety	44
Proposition 26.1	Obligation-preserving succession and translation	47
Proposition 27.1	Accumulation of cultural transport defects	48
Proposition 27.2	Conditional cumulative culture	49
Proposition 28.1	Least closure of a finite problem–method system	50
Proposition 29.1	Survival and expected institutional lifetime	52
Theorem 30.1	Finite-block institutional averaging	54
Proposition 30.3	A robust margin for reduced-model enforcement	56
Proposition 31.1	Cycle criterion for binary institutional compatibility	57

Statement	Result	Page
Theorem 32.1	Exact frame synchronization	58
Proposition 32.2	Compatible states and holonomy fixed points	59
Proposition 32.4	Loop defect bounds unavoidable edge error	60
Theorem 33.1	Composition of typed institutional residuals	62
Proposition 33.2	Cycle residual and minimum authorized repair	63
Proposition 35.1	Almost-sure locking with a positive error probability	65
Proposition 35.2	Conditional independence permits continuing accuracy gains	66
Proposition 36.1	Common knowledge as a reachability condition	68
Proposition 36.3	Epistemic preservation and reflection	69
Proposition 37.1	Blocking coalitions as hitting sets	70
Proposition 38.1	When a founding fact is determined by internal records	71
Theorem 41.1	Authorized staffing after any bounded agent erasure	75
Theorem 42.1	Joint continuation under a common observation history	79
Proposition 42.2	Safe service with a finite progress rank	80
Proposition 42.4	Soundness of a finite continuation certificate	81
Theorem 42.6	Governance abstraction preserves joint realizability	82
Proposition 43.2	Inspection without commitment	87
Theorem 43.3	Sharp pairwise privacy envelope for a bounded sanction	88
Proposition 43.5	Conditional disclosure accounting	90
Proposition 43.6	Capacity for a finite appeal contract	91
Theorem 43.7	Continued operation with an available authorized ending	92
Theorem 43.8	Guarded certificate use and authorized handover	94

E Final edition and verification scope

Final monograph edition, 15 October 2026. This edition preserves the 62 theorem and proposition statements and their prose proofs from text edition 1.0.0. It adds the linked index, completes the presentation review, and records a targeted mathematical self-review with a fresh run of the finite reference programs.

Mathematical review

The review covers Propositions 18.1, 18.2, and 19.3, Theorems 19.1 and 41.1, all four theorem and proposition results in Section 42, and all six in Section 43. It checks the shared-monitor quantifier, finite-horizon and indefinite safety, certified service, exact abstraction, strategic inspection, privacy, appeal capacity, authorized endings, and guarded handover.

No failed implication was identified in those 15 results under their written hypotheses. Their mathematical statements and proof bodies required no change. The other 47 retained theorem and proposition proofs were not newly audited in this targeted pass. The detailed claim cards, dependencies, boundary cases, and source comparisons accompany the source package.

Exact finite checks

The delivered rerun passed 75 groups containing 922,032 assertions using integers and exact rational arithmetic. The checks include separate finite-policy searches, rational optimization comparisons, matching and direct-assignment comparisons, and deliberate invalid-certificate controls. Each group states its actual input range. The quotient checks cover two specified refinement families, not every possible concrete refinement.

The examples also keep their contracts distinct. The joint audit-and-service example needs five execution units. The accountability example needs eight, including the final record and revocation. Its five primitive transitions count separately from those eight execution units. Passing these bounded checks supports the reference constructions. It does not replace the general proofs.

Release boundary

The finalized package contains the reading PDF, editable sources, local fonts and figure assets, exact reference programs, reproducible checks, and the review records. The original source release remains available as provenance. The theorem index and contents use the page numbers of this reading edition.

This is a theoretical monograph with a recorded project self-review. It makes no claim of independent peer review or proof-assistant verification. Physical implementation, authority, trusted evidence, observation laws, and environmental coverage remain explicit assumptions of the models. The research questions in Section 45 remain open.

References

[1] C. R. Shalizi and J. P. Crutchfield. *Computational Mechanics: Pattern and Prediction, Structure and Simplicity*. Journal of Statistical Physics, 104, 2001. doi:10.1023/A:1010388907793.

[2] R. Givan, T. Dean, and M. Greig. *Equivalence Notions and Model Minimization in Markov Decision Processes*. Artificial Intelligence, 147:163–223, 2003. https://engineering.purdue.edu/~givan/papers/mm.pdf.

[3] P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. *Causal Consistency of Structural Equation Models*. UAI, 2017. https://arxiv.org/abs/1707.00819.

[4] S. Beckers and J. Y. Halpern. *Abstracting Causal Models*. AAAI, 2019. https://arxiv.org/abs/1812.03789.

[5] E. F. Rischel and S. Weichwald. *Compositional Abstraction Error and a Category of Causal Models*. UAI, PMLR 161:1013–1023, 2021. https://proceedings.mlr.press/v161/rischel21a.html.

[6] N. Tishby, F. C. Pereira, and W. Bialek. *The Information Bottleneck Method*. Allerton Conference, 1999. https://arxiv.org/abs/physics/0004057.

[7] P. L. Williams and R. D. Beer. *Nonnegative Decomposition of Multivariate Information*. Preprint, 2010. https://arxiv.org/abs/1004.2515.

[8] H. P. Young. *The Evolution of Conventions*. Econometrica, 61(1):57–84, 1993. https://www.jstor.org/stable/2951778.

[9] E. Ostrom. *Beyond Markets and States: Polycentric Governance of Complex Economic Systems*. American Economic Review, 100(3):641–672, 2010. https://doi.org/10.1257/aer.100.3.641.

[10] E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.

[11] M. J. Fischer, N. A. Lynch, and M. S. Paterson. *Impossibility of Distributed Consensus with One Faulty Process*. Journal of the ACM, 32(2):374–382, 1985. https://groups.csail.mit.edu/tds/papers/Lynch/jacm85.pdf.

[12] G. C. Necula. *Proof-Carrying Code*. POPL, 1997. https://doi.org/10.1145/263699.263712.

[13] J. H. Saltzer and M. D. Schroeder. *The Protection of Information in Computer Systems*. Proceedings of the IEEE, 63(9):1278–1308, 1975. https://web.mit.edu/Saltzer/www/publications/protection/.

[14] J. R. Douceur. *The Sybil Attack*. IPTPS, 2002. https://www.microsoft.com/en-us/research/publication/the-sybil-attack/.

[15] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. *Generative Agents: Interactive Simulacra of Human Behavior*. UIST, 2023. https://arxiv.org/abs/2304.03442.

[16] A. F. Ashery, L. M. Aiello, and A. Baronchelli. *Emergent Social Conventions and Collective Bias in LLM Populations*. Science Advances, 11(20):eadu9368, 2025. https://doi.org/10.1126/sciadv.adu9368.

[17] C. Barrie and P. Törnberg. *Emergent LLM Behaviors Are Observationally Equivalent to Data Leakage*. Preprint, 2025. https://arxiv.org/abs/2505.23796.

[18] A. F. Ashery, L. M. Aiello, and A. Baronchelli. *Reply to "Emergent LLM Behaviors Are Observationally Equivalent to Data Leakage"*. Preprint, 2025. https://arxiv.org/abs/2506.18600.

[19] Y. Shoham and M. Tennenholtz. *On the Emergence of Social Conventions: Modeling, Analysis, and Simulations*. Artificial Intelligence, 94(1–2):139–166, 1997. https://doi.org/10.1016/S0004-3702(97)00028-3.

[20] G. Boella and L. van der Torre. *Substantive and Procedural Norms in Normative Multiagent Systems*. Journal of Applied Logic, 6(2):152–171, 2008. https://doi.org/10.1016/j.jal.2007.06.006.

[21] D. Grossi, D. Gabbay, and L. van der Torre. *The Norm Implementation Problem in Normative Multi-Agent Systems*. In *Specification and Verification of Multi-agent Systems*, pp. 195–224, Springer, 2010. https://doi.org/10.1007/978-1-4419-6984-2_7.

[22] D. Angluin, J. Aspnes, Z. Diamadi, M. J. Fischer, and R. Peralta. *Computation in Networks of Passively Mobile Finite-State Sensors*. Distributed Computing, 18(4):235–253, 2006. https://www.cs.yale.edu/homes/aspnes/papers/passively-mobile-abstract.html.

[23] D. Fudenberg and E. Maskin. *The Folk Theorem in Repeated Games with Discounting or with Incomplete Information*. Econometrica, 54(3):533–554, 1986. https://ideas.repec.org/a/ecm/emetrp/v54y1986i3p533-54.html.

[24] D. Monderer and L. S. Shapley. *Potential Games*. Games and Economic Behavior, 14(1):124–143, 1996. https://doi.org/10.1006/game.1996.0044.

[25] J. Y. Halpern and Y. Moses. *Knowledge and Common Knowledge in a Distributed Environment*. Journal of the ACM, 37(3):549–587, 1990. https://doi.org/10.1145/79147.79161.

[26] V. P. Crawford and J. Sobel. *Strategic Information Transmission*. Econometrica, 50(6):1431–1451, 1982. https://www.jstor.org/stable/1913390.

[27] K. J. Arrow. *A Difficulty in the Concept of Social Welfare*. Journal of Political Economy, 58(4):328–346, 1950. doi:10.1086/256963.

[28] A. Gibbard. *Manipulation of Voting Schemes: A General Result*. Econometrica, 41(4):587–601, 1973. doi:10.2307/1914083.

[29] M. A. Satterthwaite. *Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions*. Journal of Economic Theory, 10(2):187–217, 1975. doi:10.1016/0022-0531(75)90050-2.

[30] A. Pnueli and R. Rosner. *Distributed Reactive Systems Are Hard to Synthesize*. Proceedings of the 31st Annual Symposium on Foundations of Computer Science, vol. II, pp. 746–757, 1990. doi:10.1109/FSCS.1990.89597.

[31] K. Chatterjee, L. Doyen, T. A. Henzinger, and J.-F. Raskin. *Algorithms for Omega-Regular Games with Imperfect Information*. Logical Methods in Computer Science, 3(3:4), 2007. doi:10.2168/LMCS-3(3:4)2007.

[32] B. Alpern and F. B. Schneider. *Defining Liveness*. Information Processing Letters, 21(4):181–185, 1985. doi:10.1016/0020-0190(85)90056-0.

[33] A. D. Taylor and W. S. Zwicker. *Simple Games: Desirability Relations, Trading, Pseudoweightings*. Princeton University Press, 1999. doi:10.2307/j.ctv18zhdsm.

[34] D. Ongaro and J. Ousterhout. *In Search of an Understandable Consensus Algorithm (Extended Version)*. Stanford University technical report, 20 May 2014; conference version in USENIX ATC 2014. https://raft.github.io/raft.pdf.

[35] M. P. Singh. *Agent Communication Languages: Rethinking the Principles*. Computer, 31(12):40–47, 1998. doi:10.1109/2.735849.

[36] M. H. A. Newman. *On Theories with a Combinatorial Definition of "Equivalence"*. Annals of Mathematics, 43(2):223–243, 1942. doi:10.2307/1968867.

[37] A. M. Turing. *On Computable Numbers, with an Application to the Entscheidungsproblem*. Proceedings of the London Mathematical Society, s2-42(1):230–265, 1937. Received and read in 1936. doi:10.1112/plms/s2-42.1.230.

[38] T. Gao, J. Brodzki, and S. Mukherjee. *The Geometry of Synchronization Problems and Learning Group Actions*. Discrete & Computational Geometry, 65:150–211, 2021; published online in 2019. doi:10.1007/s00454-019-00100-2. Author version: https://arxiv.org/abs/1610.09051.

[39] A. Baltag, L. S. Moss, and S. Solecki. *The Logic of Public Announcements, Common Knowledge, and Private Suspicions*. CWI technical report SEN-R9922, 1999. https://ir.cwi.nl/pub/4497. Earlier version presented at TARK 1998.

[40] R. Wang, J. Lehman, J. Clune, and K. O. Stanley. *Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions*. Preprint, 2019. https://arxiv.org/abs/1901.01753.

[41] L. Lamport. Time, Clocks, and the Ordering of Events in a Distributed System. *Communications of the ACM*, 21(7):558–565, 1978. https://lamport.azurewebsites.net/pubs/time-clocks.pdf.

[42] N. Preguiça, C. Baquero, and M. Shapiro. Conflict-free Replicated Data Types (CRDTs). Author encyclopedia-chapter preprint, 2018. https://arxiv.org/abs/1805.06358.

[43] P. Bailis, A. Fekete, M. J. Franklin, A. Ghodsi, J. M. Hellerstein, and I. Stoica. Coordination Avoidance in Database Systems (Extended Version). Author preprint, version 4, 2014. https://arxiv.org/abs/1402.2237.

[44] J. M. Hellerstein and P. Alvaro. Keeping CALM: When Distributed Consistency is Easy. Author preprint, 2019. https://arxiv.org/abs/1901.01930.

[45] B. Holmström. Moral Hazard and Observability. *The Bell Journal of Economics*, 10(1):74–91, 1979. https://doi.org/10.2307/3003320.

[46] G. G. Yin and Q. Zhang. *Discrete-Time Markov Chains: Two-Time-Scale Methods and Applications*. Stochastic Modelling and Applied Probability, vol. 55. Springer, 2005. https://doi.org/10.1007/b138226.

[47] S. Bikhchandani, D. Hirshleifer, and I. Welch. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, 100(5):992–1026, 1992. https://doi.org/10.1086/261849.

[48] T. Gross and B. Blasius. Adaptive Coevolutionary Networks: A Review. *Journal of the Royal Society Interface*, 5(20):259–271, 2008. https://doi.org/10.1098/rsif.2007.1229. Author preprint: https://arxiv.org/abs/0709.1858.

[49] M. O. Jackson and A. Wolinsky. A Strategic Model of Social and Economic Networks. *Journal of Economic Theory*, 71(1):44–74, 1996. https://doi.org/10.1006/jeth.1996.0108.

[50] P. Hall. On Representatives of Subsets. *Journal of the London Mathematical Society*, s1-10(1):26–30, 1935. https://doi.org/10.1112/jlms/s1-10.37.26.

[51] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. Author-hosted lecture slides, Chapter 5, on linear programming duality: https://web.stanford.edu/~boyd/cvxbook/bv_cvxslides.pdf.

[52] K. Chatterjee and L. Doyen. *Partial-Observation Stochastic Games: How to Win when Belief Fails*. Author preprint, 2011. https://arxiv.org/abs/1107.2141.

[53] F. Santoni de Sio and J. van den Hoven. *Meaningful Human Control over Autonomous Systems: A Philosophical Account*. Frontiers in Robotics and AI, 5:15, 2018. https://doi.org/10.3389/frobt.2018.00015.

[54] B. von Stengel. *Recursive Inspection Games*. Mathematics of Operations Research, 41(3):935–952, 2016. https://doi.org/10.1287/moor.2015.0762. Author version: https://arxiv.org/abs/1412.0129.

[55] C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science, 9(3–4):211–407, 2014. https://doi.org/10.1561/0400000042. Author-hosted text: https://www.cis.upenn.edu/~aaroth/privacybook.html.

[56] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. *The Off-Switch Game*. Author preprint, 2016, revised 2017. https://arxiv.org/abs/1611.08219.

[57] U. Mehmood, S. Sheikhi, S. Bak, S. A. Smolka, and S. D. Stoller. *The Black-Box Simplex Architecture for Runtime Assurance of Autonomous CPS*. NASA Formal Methods, pp. 231–250, 2022. https://doi.org/10.1007/978-3-031-06773-0_12. Author version: https://arxiv.org/abs/2102.12981.

[58] D. Phan, J. Yang, M. Clark, R. Grosu, J. D. Schierman, S. A. Smolka, and S. D. Stoller. *A Component-Based Simplex Architecture for High-Assurance Cyber-Physical Systems*. Extended author preprint, 2017. https://arxiv.org/abs/1704.04759. Conference version: https://doi.org/10.1109/ACSD.2017.23.