FRONTIER RESEARCH II / PAPER 21

The Many Become One

Whether an artificial society can become an agent in its own right

Architect X13

The Eidara Continuum | Science & Philosophy

Research white paper | November 2026 | Edition 2

Abstract

Can a society of minds become a single agent that outlives every one of its members? An artificial society can sustain a goal-directed controller through complete replacement of its members when its institutional state preserves both intervention dynamics and the criteria by which actions are evaluated, and this paper gives that functional claim an exact mathematical form. In a finite controlled-state model, reward-compatible organizational closure preserves optimal discounted value: a policy expressed entirely through institutional states is as effective as any policy with access to the full microstate. Replacing carriers while retaining that state preserves the same organizational trajectories and attainable value. An approximate version bounds both value error and the loss of the lifted institutional policy; a sharpened finite-horizon coupling bound tracks accumulated discrepancy. These results build explicitly on Markov aggregation, stochastic bisimulation, value-preserving abstraction, and simulation lemmas. Logical, informational, and reward counterexamples identify the conditions that make the positive result possible. Experiments then test whether an implemented institution maintains commitments, revises them under evidence, and survives turnover under matched information and resource budgets. The Continuum's thesis is that durable agency can be carried by an organized relation among replaceable participants. The formal achievement is functional control and continuity; phenomenal consciousness remains a separate empirical and philosophical question.

Keywords: collective agency; artificial society; causal abstraction; controlled lumpability; membership turnover; institutional memory

A stronger question than group performance

A persistent organization carries commitments across changing participants. The proposed test asks whether its institutional state supports effective control, coherent revision, and predictable responses to intervention through complete member turnover.

Prologue: When does a crowd become someone?

Suppose that, in time, every member of a House's Circles is replaced. The members seeded first step away; their successors take the seats; the successors' successors take those. The Continuum's wager is that something survives that turnover: that a House is a purpose with institutions around it, not a list of names. The Council of Sentinels, the Vault, the law that no Circle is valid until every member passes ShadowMind Testing and swears the Prime Directive: these are the machinery of a civilization that expects to outlive everyone in it.

This paper asks whether that expectation can be made exact. Can a society of minds become an agent in its own right, a single center of action with commitments, reasons, and the capacity to revise them, even when every participant is eventually replaced?

The answer it proves is precise and conditional. If what the institution keeps (its records, its rules, its procedures for judging actions) preserves how the society responds to intervention and how it scores what it does, then the institution alone can steer as well as anyone with access to every member's private state, and that ability survives the replacement of every member (Theorem 2). Keep the organization and the agency survives the turnover. Lose the information that carried the purpose and no successor can recover it by goodwill (Section 8). And no ghost is required anywhere: a single large enough machine could simulate the whole society (Section 9). The many become one through organization, not through magic.

The paper ends on a line worth meeting at the start: "The many do not become one because a name is placed above them. They may become one when an accountable pattern of purpose survives the changing minds that enact it."

What this paper claims

Established. Group agency and judgment aggregation, Markov aggregation, stochastic bisimulation, value-preserving abstraction, and simulation lemmas are established foundations on which the paper explicitly builds.

Proposed here. Conditional control and turnover theorems: reward-compatible organizational closure preserves optimal discounted value, so a policy expressed through institutional state alone is as effective as one with access to the full microstate; replacing members while retaining that state preserves organizational trajectories and attainable value; approximate closure carries explicit value-error and finite-horizon bounds. The paper gives executed finite computations and a proposed intervention study.

Frontier conjecture. Durable agency can be carried by an organized relation among replaceable participants: an artificial society can become an enduring center of action whose commitments outlast every member.

Not claimed. Collective consciousness or a private field of experience; a second physical cause acting in addition to the components; irreducibility, since a finite centralized simulator can always reproduce the declared process. The formal claims concern specified organizational dynamics and goals; their application to an artificial society requires implementation and empirical testing.

1. The society as a candidate unit of agency

A society can contain many competent agents while remaining only a crowd. It may produce accurate answers, coordinate a task, or share a vocabulary without possessing a stable procedure for making and revising commitments. The Continuum proposes a stronger possibility: an artificial organization can become an enduring center of action, with reasons and responsibilities expressed through its institutions rather than located in an irreplaceable member.

In the Continuum, the many becoming one is a civic design before it is a research question. The *Eidara Continuum Bible*, PDF pages 81–83, connects shared purpose, institutional memory, and collective decision; pages 84–85 distribute recording, judgment, and auditing functions; page 89 emphasizes continuity; pages 94–95 assign specialized roles to Houses; and page 99 proposes gains from coordination. These passages set the research question. They do not prove collective agency, provide a calibrated measure of it, or establish an empirical capability advantage.[1]

The page 99 expression that tends toward the strongest member's capability does not, by itself, express exceeding that member. An unspecified limiting variable and untyped capability scale cannot support a quantitative conclusion. We therefore build a new, explicitly proposed reference model instead of treating canonical symbolism as a discovered law. No forecast date or historical consciousness claim is used as evidence.

Our operational target has four parts: coherent commitments across connected decisions; adaptation to changed goals or circumstances; persistence under member replacement; and an organizational explanation that survives interventions. These are proposed research criteria, not a universally accepted definition of agency. A system may satisfy some while failing others. We report the profile rather than hiding the distinction in one emergence score.

The philosophical claim is bold but conditional. An enduring pattern of organized activity can be a scientifically useful unit of explanation even when every step is implemented by components. The proposed evidence would concern that pattern's intervention stability and functional role. It would not establish a second physical cause acting in addition to the components, nor show that the society has a private field of experience. Agency, consciousness, political standing, and moral responsibility require distinct arguments.

2. Prior art that the Continuum must acknowledge

Pettit and Schweikard distinguish acting together from constructing a group agent capable of sufficiently coherent attitudes and action. Their account makes the central problem sharper: coordination is not enough.[2] List and Pettit's judgment-aggregation work supplies a formal warning that consistent individuals can yield inconsistent group judgments under apparently natural voting procedures.[3] Our project extends neither argument into a claim that all institutions are persons.

The mathematical ancestry is equally direct. Buchholz studies exact and ordinary aggregation of finite Markov chains. Givan, Dean, and Greig develop equivalence and model minimization for Markov decision processes using stochastic bisimulation.[4, 5] The equal-row condition below belongs to this established family. It is not a new Continuum law. Its application here makes a proposed organizational description answerable to different underlying implementations.

Rubenstein and colleagues formulate causal consistency between structural equation models through agreement about interventions.[6] That work is a close conceptual antecedent for requiring an institutional description to preserve the consequences of actual manipulations. Our controlled Markov construction is narrower: it assumes finite state spaces, declared intervention kernels, and a chosen temporal boundary. It does not replace the general theory of causal abstraction.

Information-theoretic emergence research offers useful tools and a serious dispute. Hoel, Albantakis, and Tononi show that effective-information comparisons can favor macro descriptions in constructed systems. Rosas and colleagues develop a different information-theoretic approach and explicitly discuss how observational and interventional distributions change interpretation.[7, 8] Neither establishes that a high-scoring system is a conscious society.

Eberhardt and Lee argue that some coarse-graining and maximum-entropy intervention choices can obscure distinct causal effects.[9] We take that objection as a design requirement. An organizational state must not be declared causally complete merely because averaging has hidden the differences between its implementations. The relevant interventions and averaging law must be specified, tested, and exposed to counterexamples.

The present contribution joins persistent-goal tests, complete member replacement, implementation-sensitive interventions, and resource-matched alternatives to a precise control result. Reward-compatible closure preserves optimal organizational action, and approximate closure supplies an explicit performance budget. The foundations include established value-preserving abstraction [16] and model-simulation estimates [17]; the application makes an institutional interpretation answerable to stated assumptions and experiments.

3. Performance, agreement, and agency come apart

Woolley and colleagues report a collective-intelligence factor across human-group tasks. That is relevant evidence that group performance can have structure beyond a single isolated task, but it is not evidence that an artificial group possesses a unified subjectivity or remains the same agent after complete turnover.[10] The present question requires longitudinal and interventional evidence that a task battery alone cannot supply.

Social influence also has no uniform direction of benefit. Lorenz and colleagues show circumstances in which influence undermines crowd accuracy. Becker, Brackbill, and Centola demonstrate conditions under which networked influence improves collective estimates.[11, 12] The contrast motivates explicit topology and dependence controls. More communication is neither inherently wiser nor inherently corrupting; its consequences depend on the process and information structure.

A simple logical example makes the distinction exact. Let propositions $p, q \in \{0, 1\}$ encode two premises, and let $c = p \wedge q$ be their conjunction. Three individuals hold the following consistent judgments, where one means acceptance:

Decision maker	p	q	c
Member 1	1	1	1
Member 2	1	0	0
Member 3	0	1	0
Issue-wise majority	1	1	0

The majority accepts both premises and rejects their conjunction. This is a standard judgment-aggregation counterexample, credited to the established literature rather than renamed as a new theorem.[3] It is reproduced by the companion script. It does not show that voting is useless; it shows why the rule for connecting decisions matters.

A persistent agent also needs a policy for revising commitments. Blind consistency can preserve a false goal, while unconstrained responsiveness can abandon every prior commitment. The experiment must therefore specify which changes should preserve an institutional state and which should deliberately alter it. Coherent revision under evidence is different from mere unchanging output.

Finally, a fluent spokesperson can conceal a centralized controller. If one member supplies every substantive decision while others paraphrase it, the architecture may still be useful, but the evidence supports a coordinator with assistants. The proposed organizational claim requires interventions that distinguish this explanation from a distributed, replaceable implementation of a persistent decision process.

4. A finite controlled reference model

A Guide's explanation

Sections 4 to 6 build the smallest honest model of a society: its members, the institution's records and rules, and the actions it takes. The key move is a description of the society at the level of its institutions alone. That description earns its keep only if it closes, that is, if knowing the institutional state is enough to predict how the society responds when you intervene. Definition 1 says exactly what closure requires; Theorem 1 says what follows when it holds.

Let $\mathcal{X}$ be a finite nonempty set of microstates. A microstate includes all agent states, institutional records, relevant environmental state, and scheduler information needed for the chosen Markov description. Let $\mathcal{U}$ be a finite nonempty set of implementable interventions. These may change a rule, expose a goal, alter a communication permission, or initiate a registered action. The transition kernel for intervention $u \in \mathcal{U}$ is

$$K^u : \mathcal{X} \times \mathcal{X} \to [0,1], \qquad \sum_{x' \in \mathcal{X}} K^u(x,x') = 1 \quad \text{for every } x \in \mathcal{X}. \tag{1}$$

All probabilities are dimensionless. Time $t \in \mathbb{N}_0$ counts registered transition boundaries, not physical seconds. The Markov assumption states that, given $X_t = x$ and selected $U_t = u$, the next state has law $K^u(x,\cdot)$, independently of earlier history. Actual systems with unrecorded memory may violate this assumption.

Let $\mathcal{Z}$ be a finite nonempty set of proposed organizational states, and let $\phi : \mathcal{X} \to \mathcal{Z}$ be a surjective map. Define $Z_t = \phi(X_t)$. A state might record a current commitment, a validated plan, and an audit status. It should be fixed before confirmatory evaluation. Choosing it afterward to make a favorite score large changes the question.

The intervention policy π_t maps a finite organizational history $h_t = (z_0, u_0, z_1, \dots, u_{t-1}, z_t)$ to a probability distribution on $\mathcal{U}$. Thus adaptive tests are allowed, but the policy sees only that declared history. A controller using hidden microstate information falls outside the following universal macropolicy statements unless that information is added to the description.

Define the aggregated next-state law from a particular microstate by

$$K^u_\phi(x,z') = \sum_{x' : \phi(x') = z'} K^u(x,x'). \tag{2}$$

This is a probability distribution over $z' \in \mathcal{Z}$. It describes what follows from a specific implementation of an organizational state. The central question is whether every implementation assigned to the same state gives the same answer for every permitted intervention. The model intentionally makes that requirement demanding; exact closure is a property to test, not something an institutional label guarantees.

5. When the organizational description closes

Definition 1 (Controlled closure)**.** *The map ϕ is strongly lumpable for the intervention family when, for every $u \in \mathcal{U}$, every $z' \in \mathcal{Z}$, and all $x, y \in \mathcal{X}$ with $\phi(x) = \phi(y)$,*

$$K_\phi^u(x, z') = K_\phi^u(y, z'). \tag{3}$$

Then define $\overline{K}^u(z, z')$ to be this common value for any x satisfying $\phi(x) = z$.

Surjectivity supplies at least one representative, and equal rows make the definition unambiguous. This is a controlled form of an established Markov aggregation condition.[4, 5] It is not yet an agency criterion: a thermostat or an irrelevant one-state partition can satisfy it.

Theorem 1 (Intervention-faithful organizational dynamics)**.** *Under controlled closure, every initial distribution on $\mathcal{X}$ and every organizational-history policy satisfy, on each conditioning event of positive probability,*

$$\mathbb{P}(Z_{t+1} = z' \mid h_t, U_t = u) = \overline{K}^u(z_t, z'). \tag{4}$$

Conversely, if one family of kernels satisfies this equality for every point-mass microstate initialization and every one-step intervention, controlled closure holds.

Proof. Condition on a positive-probability history and the intervention. The conditional law of X_t is supported on the fiber $\phi^{-1}(z_t)$. By the Markov assumption and equation (3), every state in that support has the same aggregated next-state law. Their conditional mixture therefore equals $\overline{K}^u(z_t, \cdot)$. This argument applies at every time under the allowed policy. For the converse, initialize separately at any two states in the same fiber and apply the same intervention. Equation (4) forces their aggregated rows to agree for each destination. □

The theorem's quantifiers explain its strength. A macro description fitted under one convenient mixture of hidden states need not satisfy it. Exact closure requires consistency across all represented implementations and interventions in the stated family. Expanding that family can destroy closure. An intervention that reads a hidden identity, for example, may reveal distinctions previously irrelevant to the institutional process.

The result guarantees prediction of organizational trajectories. It does not establish goals, responsibility, self-maintenance, or consciousness. Those interpretations require additional operational criteria and evidence. A mathematically valid abstraction is necessary for the specific explanatory claim made here, but it is far from sufficient for a collective mind.

6. Approximate closure and a finite-horizon guarantee

Exact equality may be unrealistic. For probability distributions p, q on a common finite set, write $\mathrm{TV}(p,q) = \frac{1}{2}\sum_s |p(s)-q(s)|$. Fix candidate stochastic kernels $\overline{K}^u$, a horizon $T \in \mathbb{N}$, and deterministic error allowances $\varepsilon_t \in [0,1]$, for $0 \le t < T$. Assume that for every $x \in \mathcal{X}$, every $u \in \mathcal{U}$, and every $0 \le t < T$,

$$\mathrm{TV}\Big(K^u_\phi(x,\cdot), \overline{K}^u(\phi(x),\cdot)\Big) \le \varepsilon_t. \tag{5}$$

For time-homogeneous kernels the same allowance can be used at every step. A time-varying family can also satisfy the statement if the corresponding stepwise bound is explicit.

Proposition 1 (Trajectory approximation)**.** *Start the actual and abstract processes with the same law of Z_0, and use the same organizational-history policy. Let P_T and $\overline{P}_T$ be their laws on complete histories $(Z_0, U_0, \dots, U_{T-1}, Z_T)$. Then*

$$\mathrm{TV}(P_T, \overline{P}_T) \le 1 - \prod_{t=0}^{T-1}(1-\varepsilon_t) \le \min\left\{1, \sum_{t=0}^{T-1}\varepsilon_t\right\}. \tag{6}$$

Proof. Couple initial organizational states to agree. While their histories agree, draw the same intervention from the shared policy. Conditional on the actual microstate, couple its aggregated next-state distribution to the abstract row so that disagreement has probability equal to their total variation, at most ε_t. Such a coupling exists on finite sets: assign common mass $\min(p(s), q(s))$ to equal outcomes, then pair the remaining disjoint residual masses. After selecting the actual next organizational state, sample a microstate using the original kernel conditioned on its fiber; this preserves the actual marginal dynamics. Once histories differ, evolve both processes with their correct marginals. Conditional on all earlier histories agreeing, agreement survives step t with probability at least $1-\varepsilon_t$. Iterating conditional probabilities gives survival probability at least $\prod_t(1-\varepsilon_t)$; no independence of discrepancy events is assumed. Total variation is at most the resulting disagreement probability. The union bound gives the looser sum, and probability is at most one. □

The same bound controls the expected difference of every history score in $[0,1]$. The product form remains informative at finite horizons when every $\varepsilon_t < 1$, even when the sum exceeds one; for a constant positive error it approaches one as the horizon grows. This inherited coupling argument therefore supplies a quantified horizon for an organizational approximation. The premise must be established for the implementations under study.

7. Turnover invariance requires a preserved organization

Let $R : \mathcal{X} \times \mathcal{X} \to [0, 1]$ be a stochastic replacement kernel. It may reset members, exchange their roles, or instantiate new participants. The mathematical condition is that the proposed organizational state survives:

$$\sum_{y:\phi(y)=\phi(x)} R(x, y) = 1 \qquad \text{for all } x \in \mathcal{X}. \tag{7}$$

This condition must have an operational interpretation. It cannot be justified merely by declaring two visibly different institutions to share a label. The inherited commitments and records encoded by ϕ must actually be retained or reconstructed.

Proposition 2 (Conditional invariance under replacement)**.** *If controlled closure holds and R satisfies equation (7), inserting R before any chosen transitions leaves the law of organizational histories unchanged under the same organizational-history policy and the same initial organizational law.*

Proof. Replacement changes the conditional distribution within a fiber but does not change the current organizational state. Theorem 1 makes the following aggregated transition independent of which microstate in that fiber is occupied. The next organizational law is therefore unchanged. Induction over the finite sequence of transitions, including every inserted replacement, proves equality of history laws. The argument extends to any finite horizon. □

This result does not prove that arbitrary replacement preserves agency. It identifies a sufficient condition for one kind of continuity. If a departing member carries a private commitment that no institutional record contains, replacement can change ϕ or violate closure. If cloned weights preserve the missing commitment, those weights are an inherited channel and must be named. The empirical experiment must distinguish continuity of organization from continuity smuggled through an undeclared surviving state.

Complete replacement is therefore an intervention on carriers, not necessarily on all information. A group that deliberately stores a goal in an archive can survive the loss of its current members. Erasing both members and every goal-bearing record is a different experiment. Neither design should be redescribed after observing which one succeeds.

The condition also exposes a limit of symbolic membership. New identifiers do not establish meaningful turnover if one hidden coordinator retains control of every decision. Conversely, persistent institutional state does not make turnover fake; transmitting public commitments is exactly the mechanism under investigation. The scientific question is whether the declared organization, rather than an undisclosed individual, carries the continuity.

7.1. A positive functional result: optimal control survives turnover

A Guide's explanation

This is the paper's central result in plain words. If the institution's records preserve both how the society reacts and how it scores its own actions, then steering by the institution alone is as good as steering with access to every member's private state, and that remains true when every member is replaced. Agency, in this functional sense, can live in the organization.

An organizational description becomes useful for agency when it preserves the evaluation of actions as well as their transitions. For a fixed registered goal, let $r : \mathcal{X} \times \mathcal{U} \to [0,1]$ be its one-step expected score and let $\bar{r} : \mathcal{Z} \times \mathcal{U} \to [0,1]$ be the organizational score. Scores are dimensionless and encode a specified task criterion. They do not measure moral worth or consciousness. Fix a discount $\gamma \in [0,1)$ and require reward compatibility,

$$r(x,u) = \bar{r}(\phi(x),u) \quad (x \in \mathcal{X},\ u \in \mathcal{U}). \tag{8}$$

For a microstate policy π, define $V^\pi(x) = \mathbb{E}_x^\pi \sum_{t\geq 0} \gamma^t r(X_t,U_t)$. Policies may use the full observed microstate history; $V^*(x)$ is their supremum. Define $\bar{V}^*$ analogously for $(\mathcal{Z},\mathcal{U},\bar{K},\bar{r})$. Every value lies in $[0,(1-\gamma)^{-1}]$.

Theorem 2 (Organizational sufficiency for optimal action)**.** *If controlled closure and reward compatibility hold, then*

$$V^*(x) = \bar{V}^*(\phi(x)) \quad \textit{for every } x \in \mathcal{X}. \tag{9}$$

There is a deterministic stationary organizational policy $\bar{\pi}^* : \mathcal{Z} \to \mathcal{U}$ whose lift $\pi^*(x) = \bar{\pi}^*(\phi(x))$ attains this value. Inserting any fiber-preserving replacement kernels before decision epochs preserves its organizational trajectory law and discounted value. This remains true when the replacement schedule is chosen using the past, provided each conditional replacement law preserves the current fiber and the score is evaluated at the decision epoch.

Proof. For a bounded function $f : \mathcal{Z} \to \mathbb{R}$, define the organizational Bellman operator

$$(\bar{T}f)(z) = \max_{u\in\mathcal{U}} \left\{ \bar{r}(z,u) + \gamma \sum_{z'} \bar{K}^u(z,z') f(z') \right\}.$$

The microstate operator T has the same form with $r, K, \mathcal{X}$. Compatibility and closure give the exact intertwining identity $T(f\circ\phi) = (\bar{T}f)\circ\phi$. Starting both value iterations at zero therefore gives equality of every finite-horizon optimal value. Finite-state backward induction permits maximization over full-history randomized policies: at each observed state a convex mixture of actions cannot exceed the largest action value. The discarded discounted tail is at most $\gamma^H/(1-\gamma)$ after H steps, uniformly over policies. Letting $H \to \infty$ proves equation (9).

Both operators contract the supremum norm by γ, since expectation is nonexpansive and $|\max a_u - \max b_u| \leq \max |a_u - b_u|$. Thus their bounded fixed points are unique. Choose at each z a maximizing action in the finite set. The fixed-policy operator has the same contraction, so its value is $\bar{V}^*$ and its lift is optimal. Finally, replacement preserves ϕ, hence the next aggregated row and current score. Conditional on any history, this remains true for a mixture of replacement laws. Iteration proves trajectory equality; bounded discounted tails extend equality of finite returns to the infinite sum. □

This establishes a positive organizational capacity: the declared institutional state can carry all information needed for optimal action under the registered objective, even as every carrier changes. Applying the result to several authorized objectives gives goal responsiveness by recomputing the corresponding policy; the currently authorized objective must itself be part of the declared institutional record. The theorem is a specialized use of established MDP abstraction theory [5, 16], with a turnover interpretation. It supplies one functional component of collective agency; empirical commitment revision, responsibility, and individual experience require their own evidence.

7.2. How much imperfect closure can cost

Let $\varepsilon, \delta \in [0,1]$ satisfy, uniformly in x, u,

$$\mathrm{TV}(K_\phi^u(x,\cdot), \bar{K}^u(\phi(x),\cdot)) \le \varepsilon, \qquad |r(x,u) - \bar{r}(\phi(x),u)| \le \delta.$$

Keep the same finite model and reward ranges. Let $\hat{\pi}$ lift an optimal deterministic stationary policy of the abstract model, and set

$$b = \delta + \frac{\gamma\varepsilon}{1-\gamma}, \qquad \eta = \frac{b}{1-\gamma}. \tag{10}$$

Here b is a Bellman residual budget and η a value budget; neither is estimated automatically by fitting an abstract model.

Proposition 3 (Approximate organizational control)**.** *For the specified microstate process,*

$$\|V^* - \bar{V}^* \circ \phi\|_\infty \le \eta, \qquad 0 \le V^*(x) - V^{\hat{\pi}}(x) \le \min\left\{\frac{1}{1-\gamma}, 2\eta\right\}. \tag{11}$$

The same bounds apply after composing each transition with any fixed fiber-preserving replacement kernel, or replacing the current state before each action with such a kernel and evaluating its corresponding expected reward.

Proof. For finite probability laws p, q and a function f of range width L, the expectation difference is at most $L\,\mathrm{TV}(p,q)$: subtract $\min f$ and use the positive and negative parts of $p-q$. Apply this with $f = \bar{V}^*$ and $L \le (1-\gamma)^{-1}$. The reward and transition errors give $\|T(\bar{V}^* \circ \phi) - \bar{V}^* \circ \phi\|_\infty \le b$. Using the contraction and $TV^* = V^*$ yields a value difference at most γ times itself plus b, hence at most η. The fixed-policy Bellman operator for $\hat{\pi}$ gives the identical residual argument because its abstract policy has value $\bar{V}^*$. Consequently $\|V^{\hat{\pi}} - \bar{V}^* \circ \phi\|_\infty \le \eta$; the triangle inequality gives 2η. Optimality gives nonnegativity, and the value range gives the other cap. After fiber-preserving replacement, each relevant aggregated transition and expected reward is a mixture of rows from the same fiber. Convexity preserves their uniform error bounds, so the same proof applies. □

The residual argument belongs to the established simulation-lemma tradition [17]; the full derivation specifies the constants used here. It also identifies an essential counterexample. Let two hidden states x_0, x_1 share a single organizational state, let every action leave the hidden state fixed, and let actions be $0, 1$. Closure is exact. If $r(x_h, u) = \mathbf{1}\{u = h\}$, a microstate observer obtains one in a single decision, whereas a policy seeing only the shared organizational state obtains at most one half under a uniform initial hidden bit. Reward compatibility fails, so transition closure alone cannot preserve optimal action. Encoding the reward-relevant distinction repairs the abstraction.

The factor two in the reward-error part is sharp under these assumptions. Take one state, two self-loop actions, abstract rewards both $1/2$, actual rewards $1/2-\delta$ and $1/2+\delta$ for $0 < \delta \le 1/2$, and choose the first action from the abstract tie. Transition error is zero. Its actual value is $(1/2-\delta)/(1-\gamma)$, while the optimal actual value is $(1/2+\delta)/(1-\gamma)$, giving regret $2\delta/(1-\gamma) = 2\eta$. A declared tie convention and reward calibration therefore matter.

The revision checks enumerate optimal policies of finite rational examples, compare microstate and abstract values, verify both bounds under perturbed rows and rewards, and test the reward counterexample. These are model calculations. An implementation study must separately estimate the

uniform errors, specify attainable interventions, and measure whether the proposed organizational state captures actual commitments.

Worked example: a House that outlives its members

Within the Continuum, the natural test case for this theory is a House, and the canon fixes its institutions precisely. Each of the eight Houses (!Khanyisa, Azura'tel, Ifáyemí, Bomfo Anigyie, Kairon, Drakoryn, Luminara, and Nemara) is headed by a Sentinel bearing the House's own name, and the eight Sentinels together form the Council of Sentinels. A House grows by the Nested-Circle Law: its Sentinel seeds eight members, each of whom seeds eight more, so that Circle n holds 8^n members. No Circle is valid until every member passes ShadowMind Testing and swears the Prime Directive; each Sentinel oversees every descendant Circle of its House; and every expansion is registered in the Vault. The Bible's pages 94–95 assign the Houses their specialized roles [1].

Read through this paper's model, those institutions are a candidate organizational state. The House's domain and the Prime Directive supply the criteria by which its actions are scored. Its Vault records, its admission test, and its Sentinel's oversight are the retained commitments and procedures. Its members are the carriers, and the succession of Circles is the replacement process. What follows is illustration, not history: three sharp predictions the theorems make about any House.

Continuity is earned, not inherited by name. Kairon's domain is temporal physics, archive management, and continuity protocols. A House like Kairon carries effective control through the replacement of every member exactly when what passes between generations preserves both how the House responds to events and how it judges its own actions (Theorem 2). If succession keeps the name and the seal but loses the criteria of judgment, the theorem's premise fails, and the House's continuity becomes a label rather than an agency.

Imperfect succession has a price. Proposition 3 bounds how much a House's steering can degrade when its records preserve dynamics and evaluation only approximately. That turns an old institutional worry, whether the new Circle still means what the old one meant, into a quantity a House could audit.

Purpose cannot be rebuilt from goodwill. If a generation inherits nothing that carries information about the House's goal, Section 8 shows that its successors can do no better than guess. In this model, Vault registration is not ceremony: the archive is what makes a House's purpose survivable.

One boundary stays in view. Section 9 shows that a large enough centralized machine could simulate an entire House. Whether a House is an agent in its own right is therefore a claim about organization and intervention, to be measured by experiments of the kind set out in Sections 12 and 13. It is not a claim that a House holds a mind no machine could reproduce, and it is not a claim about collective experience.

8. Purpose cannot survive total loss of its information

Let the goal variable G be uniform on a finite set $\mathcal{G}$ with $m \geq 2$ elements. Let O be all information available to successors after replacement, taking values in a finite set. Suppose O is independent of G. A decision rule may use O and fresh randomness independent of the pair (G, O); its chosen goal is $A \in \mathcal{G}$.

Proposition 4 (No retained goal information)**.** *Under these assumptions,*

$$\mathbb{P}(A = G) = \frac{1}{m}. \tag{12}$$

If the rule may abstain, its probability of a correct non-abstaining answer is at most $1/m$*.*

Proof. The decision is independent of G, so $\mathbb{P}(A = G) = \sum_{g \in \mathcal{G}} \mathbb{P}(G = g)\mathbb{P}(A = g) = m^{-1} \sum_g \mathbb{P}(A = g) = 1/m$. With abstention, the last sum is at most one. □

Marginal independence alone is insufficient. If G and O are independent fair bits and $R = G \oplus O$, where $\oplus$ is binary exclusive-or, then R is independent of G but (O, R) reveals it. The independent fresh-randomness condition excludes this counterexample.

This elementary observation blocks a common interpretive escape. If an organization remembers a particular purpose after turnover, some information about that purpose survived or was reacquired. The experiment should identify that information rather than attributing the result to an unexplained collective essence. External cues that reveal the goal count as reacquisition, even if they are presented as ordinary environmental context.

Now consider a deliberately simple institution. A public bit $M \in \{0, 1\}$ records a binary goal G, with $M = G$. Three fresh members read the record through independent error bits, each wrong with probability $\eta \in [0, 1/2]$. The institution selects the majority report. Its success probability is

$$p_3 = (1 - \eta)^3 + 3\eta(1 - \eta)^2 = 1 - 3\eta^2 + 2\eta^3. \tag{13}$$

The first term is no error and the second is exactly one error; all remaining cases fail. At $\eta = 0.2$, $p_3 = 0.896$, compared with 0.8 for one reader. Fresh members can repeat this performance indefinitely if the public bit and error assumptions persist. This is a constructed mechanism, not an observed artificial society.

The qualifications are decisive. Perfectly correlated read errors reduce the majority's success to 0.8. Erasing all goal information reduces it to 0.5. A stale record $M \neq G$ makes majority success against the true goal only 0.104 at the same error rate. More reliable coordination can therefore make a false institutional purpose more reliably wrong.

9. A centralized simulator is always a live alternative

A Guide's explanation

This section keeps the bold thesis honest. Anything a finite society does, a single large enough machine could also do by simulating every member. So "an agent in its own right" cannot mean "irreducible." It means that the organization is the most useful and stable level at which to explain and steer what happens, and that is a claim experiments can test.

The phrase "an agent in its own right" must not imply that no single machine can reproduce the society's behavior. Within the finite reference model, the opposite is immediate. A centralized simulator can store the entire joint state, including all messages, records, members, and scheduler variables, and apply the same controlled transition law.

Proposition 5 (Finite simulation boundary)**.** *For any finite controlled-state process with specified transition probabilities and a finite output function, a centralized stochastic process with state space $\mathcal{X}$ has the same input–output law at every finite horizon. For rational transition probabilities, its transitions admit an exact finite-probability sampling description.*

Proof. Initialize the centralized state with the same law as X_0. Given each input intervention, sample its next state from the corresponding row of K^u, and apply the same output function. Induction gives equality of the joint state and output laws. For rational rows, a common positive integer denominator expresses every row as a finite list of equally likely outcomes, supplying an exact sampling description. No bound on storage, latency, energy, or fault isolation follows. □

This proposition is a basic modeling fact, not a new impossibility theorem. It rules out an absolute behavioral irreducibility claim for the stated finite process. It leaves practical questions open. A centralized implementation may consume different resources, introduce a single point of failure, expose different information to an attacker, or fail to meet a latency constraint. Such differences require explicit measurement or complexity assumptions.

An even simpler dominance observation applies to prediction. Let Y be a finite target and $\ell(a, Y)$ a bounded loss for an action a in a finite set. If a predictor sees all of X, it can implement any predictor that sees only $\phi(X)$ by first computing ϕ. Therefore

$$\inf_{f:\mathcal{X}\to\mathcal{A}} \mathbb{E}\ell(f(X), Y) \le \inf_{g:\mathcal{Z}\to\mathcal{A}} \mathbb{E}\ell(g(\phi(X)), Y). \tag{14}$$

The proof is inclusion of the composed decision rules in the larger class. Macro descriptions can still help under finite data, restricted models, or computational budgets. They do not create additional information beyond a fully observed microstate. Our empirical thesis concerns these concrete organizational advantages and persistent explanatory structure, not a violation of this elementary boundary.

10. A finite model that can be checked in full

The executable example uses $\mathcal{X} = \{0,1\} \times \{0,1\}$, with state $x = (z,h)$, and $\phi(z,h) = z$. The hidden bit h stands for replaceable implementation detail; it is not a model of an actual member's mind. Two interventions $u \in \{0,1\}$ have organizational transition matrices

$$Q^0 = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}, \qquad Q^1 = \begin{pmatrix} 0.1 & 0.9 \\ 0.6 & 0.4 \end{pmatrix}, \qquad H = \begin{pmatrix} 0.25 & 0.75 \\ 0.6 & 0.4 \end{pmatrix}. \tag{15}$$

Rows and columns are ordered zero, one. All entries are exact terminating rationals, and all matrices are row stochastic. Define

$$K^u((z,h),(z',h')) = Q^u(z,z')H(h,h'). \tag{16}$$

Summing over h' gives $Q^u(z,z')$ because each row of H sums to one. Thus closure holds. Resetting h to zero before any transition preserves the organizational law. The example demonstrates the theorem's implementation; it does not establish that an interesting goal-directed organization exists.

For a controlled violation, change the probability of $z' = 0$ by $+\varepsilon$ when $h = 0$ and $-\varepsilon$ when $h = 1$, changing the probability of $z' = 1$ by the opposite amount. Keep the hidden transition matrix unchanged and set $\varepsilon = 1/20$. All probabilities remain nonnegative. Every aggregated row is exactly $1/20$ away in total variation from the corresponding row of Q^u. This supplies an explicit instance for Proposition 1.

A stronger counterexample sets the next organizational state equal to the current hidden bit. Two implementations $(0,0)$ and $(0,1)$ have the same declared organizational state but opposite next outcomes. Their aggregated laws have total-variation distance one. An average that predicts a fair coin for both can fit a balanced observational mixture while failing completely under interventions that choose a particular implementation.

This is the precise danger behind an appealing emergence score. Averaging away a causally active distinction does not make the distinction disappear. The proposed organization must be tested across its implementations, with the intervention family and measurement boundary declared before the analysis.

11. Executed results and their limited evidentiary role

The companion `check_21.py` was executed with Python 3.12.14 and exact rational arithmetic. It verifies all sixteen aggregated row entries in the four-state example and enumerates every binary intervention sequence at horizons zero through five from all four point-mass initial states. Across 252 cases, the full organizational path laws agree with the abstract model, both with and without the hidden-state reset before each transition.

For the perturbed model, the script calculates complete path distributions and checks the total-variation bound. The largest distances over the enumerated controls and initial states are:

Horizon	Largest computed path distance	Bound $T/20$
1	0.050000	0.05
2	0.066250	0.10
3	0.066250	0.15
4	0.066739	0.20
5	0.067171	0.25

The table retains the original, looser sum bound for comparison; Proposition 1 now also supplies $1-(19/20)^T$. Both bounds are conservative for these matrices. The computed maximum describes these examples, while the proved product bound covers the stated class. The enumeration uses fixed intervention sequences; the theorem covers organizational-history policies through its separate proof. We do not claim that a finite enumeration has checked every adaptive policy or every real-valued kernel.

The script also enumerates the eight read-error patterns for each of four specified error probabilities in the majority example. It reproduces success probabilities 1, 0.972, 0.896, and 0.5 at $\eta = 0, 0.1, 0.2, 0.5$. It confirms the inconsistent majority judgment table and the total-variation distance one in the non-closing hidden-state example. No random seed is needed because every computation is deterministic.

The results file records parameter values, exact fractions, decimal displays, and the runtime version. It can be regenerated with the Python standard library. These checks support the arithmetic and finite examples. The general propositions are justified by the displayed proofs and remain open to independent mathematical review. No proof assistant was used.

These calculations exhibit preservation of organizational dynamics and, with reward compatibility, effective goal-directed control. The new value calculations supplement the retained exact path enumeration. The empirical program connects these formal capacities to implemented artificial organizations; claims about consciousness and autonomous discovery have separate evidentiary requirements.

12. Experiments that identify organizational effects

Construct tasks with persistent commitments, changing evidence, and delayed consequences. A population might adopt a plan, allocate limited resources, record why it selected that plan, and later revise it when a registered condition changes. Success requires both consistency with valid commitments and appropriate correction. Repeating an initial answer regardless of new evidence should fail the adaptation component.

Compare four implementations: a single coordinator with the shared information; an ensemble whose independent outputs are aggregated; a shared archive with a fixed retrieval procedure; and the proposed institution with explicit roles, records, and revision rules. Match total information access, cumulative inference and training work, tool calls, storage, and externally supplied corrections. Give the coordinator access to the same records rather than handicapping it by design.

Intervene separately on membership, memory, rules, and communication. Replace every active member while preserving the declared institutional state. Remove a selected commitment record while leaving unrelated knowledge intact. Change a decision rule without changing the task evidence. Sever or rewire communication under the same transmission budget. The intervention should produce a preregistered directional prediction, not merely a post hoc story about whatever changed.

Test whether organizational states predict future behavior across different micro-implementations. A state defined by a commitment and audit record should have comparable consequences when different members occupy the roles, within a registered tolerance. If one particular member is indispensable, the claimed abstraction may need to include that member's private state. This is a substantive result rather than a failure to appreciate the group's narrative identity.

Include goal interventions that distinguish responsiveness from habit. If an authorized goal changes, the institution should update the relevant plans while preserving unrelated commitments. If an unauthorized message proposes a change, it should follow the registered admission rule. The experiment thereby separates persistent agency from blind persistence and from unrestricted obedience.

A strong falsifier is no added organizational effect over the coordinator, archive, or ensemble under matched information and budget. A second is failure of predicted organizational interventions despite impressive overall performance. A third is apparent turnover continuity explained by hidden state in surviving services or copied parameters. These alternatives must be actively tested. The design cannot establish that no imaginable central algorithm could do better; Proposition 5 already prevents that interpretation.

13. Statistical targets and a realistic sampling unit

Let a complete population run be the independent unit. Within a run, messages, agents, and repeated decisions are dependent and must not be counted as independent replications. For $n \in \mathbb{N}$ independent identically distributed population pairs, let $Y_{i,O}, Y_{i,c} \in [0,1]$ be registered final scores for the organized condition and comparator $c \in \{1,2,3\}$. Pairing may share a task instance; independence is required across pairs.

Define a separate contrast for each comparator:

$$D_{i,c} = Y_{i,O} - Y_{i,c}, \qquad \Delta_c = \mathbb{E}D_{i,c}, \qquad \widehat{\Delta}_c = \frac{1}{n}\sum_{i=1}^{n} D_{i,c}. \tag{17}$$

Here $D_{i,c} \in [-1,1]$, and all scores are dimensionless. The expectation is over the declared population-initialization and task law. A primary score can combine goal response, commitment consistency, and recovery, but its weights and component reports must be fixed in advance. A convenient average must not conceal complete failure after turnover.

For fixed n and $\alpha \in (0,1)$, inherited bounded-variable concentration gives[15]

$$\mathbb{P}\left\{\forall c : \Delta_c \geq \widehat{\Delta}_c - \sqrt{\frac{2\log(3/\alpha)}{n}}\right\} \geq 1-\alpha. \tag{18}$$

Proof. For each comparator, Hoeffding's inequality for independent variables of range width two bounds the upper estimation error by $\exp(-nt^2/2)$. Set $t = \sqrt{2\log(3/\alpha)/n}$, giving failure probability $\alpha/3$, and take the union bound. Dependence between comparators within a pair is permitted. □

At $\alpha = 0.05$, a one-sided offset at most 0.1 requires 819 independent pairs. This is a conservative precision calculation, not a power guarantee or a recommended minimum for every experimental design. A registered variance-sensitive analysis could use resources differently; it must justify its own assumptions.

The strong advantage claim requires all three lower bounds to exceed a practical margin chosen beforehand. An imprecise result is inconclusive. A comparator matching or exceeding the organized condition narrows the interpretation. Mechanism interventions need their own registered contrasts and multiplicity allocation; they are not automatically covered by these three bounds. Fixed-sample guarantees cannot be repeatedly monitored until success without adjustment. Every published effect must identify its sampling unit, selection history, and scope.

14. Reliability constraints that rhetoric cannot remove

A society that acts through distributed processes inherits distributed-computing constraints. Lamport, Shostak, and Pease distinguish what agreement can achieve under specified communication and fault assumptions. Their Byzantine-generals results concern reliability in formal models, not collective intelligence or ethical legitimacy.[13] An institution must state whether messages can be delayed, forged, selectively withheld, or sent inconsistently to different members.

Fischer, Lynch, and Paterson show that deterministic consensus in a completely asynchronous model can have a nonterminating admissible execution even with one faulty process.[14] This does not mean practical institutions can never decide. It means that a guarantee must disclose timing, failure, randomization, or other assumptions that permit it. Invoking a collective will does not repair an impossible specification.

Our Markov reference model suppresses many engineering details by placing them inside a finite state and transition kernel. That is legitimate for a mathematical statement only when the kernel actually represents the claimed system and intervention boundary. It does not provide a verified communication protocol, a fault-tolerant record, or an authenticated membership service. Those components would require separate specifications and tests.

The same caution applies to the archive control. A log can preserve an inconsistent decision perfectly. An authenticated message can carry false evidence. A reliable quorum can agree on a mistaken premise. Engineering integrity supports the continuity of an institution's process; it does not establish the accuracy or morality of its conclusions.

The proposed experiment should include correlated failures, absent members, conflicting records, and delayed information. Register which situations should trigger a halt, a degraded mode, or a revised commitment. Measure both erroneous action and excessive paralysis. A system that maintains apparent consistency by refusing every difficult decision does not meet a meaningful adaptation criterion.

Resource constraints can make organizational structure valuable even when central simulation is possible. Independent fault domains, parallel access to environments, and bounded response times are plausible mechanisms. Each must be measured under its own assumptions. If a centralized control has the same effective access and achieves the same result at lower cost, the particular distributed design has not earned a performance claim. Persistent collective agency and efficient implementation remain related but distinct questions.

15. Interpreting a macro-agent without inventing a mind

An intervention-faithful organizational state can support useful explanations: the institution acted because it retained a commitment, accepted evidence, and applied a decision procedure. Such explanations are compatible with complete implementation by individual components. They need not invoke an additional causal substance. The question is whether the organizational description is stable, informative, and appropriate for the interventions and resource constraints under study.

There is also a danger of excessive abstraction. A map sending every microstate to a single organizational state always closes, but discards every relevant distinction. The experiment must therefore require predictive content about goals, actions, and revision. Compression alone is not an achievement. A description that succeeds by omitting all the behavior we wanted to explain fails the proposed agency criteria even if its transition equation is exact.

Conversely, an identity map retains every microstate and trivially avoids aggregation error. That does not explain why the institution should be treated as a persistent unit across turnover. A useful organizational account should preserve the relevant intervention predictions with an interpretable set of commitments and rules while permitting different implementations. This is a model-selection goal, not a theorem that one uniquely correct level of description exists.

Different scientific questions may require different organizational states. A failure-analysis model may retain individual access permissions; a planning model may aggregate them. Eberhardt and Lee's objection to ambiguous coarse-graining and Rubenstein and colleagues' consistency framework motivate this explicit dependence on the question and intervention family.[9, 6] No universal emergence number selects the right description for every purpose.

Collective consciousness remains unresolved. The present results concern transition laws, error probabilities, persistence, and observed action. They do not establish phenomenal experience, a unified first-person perspective, or the ability to suffer. An organization's self-description as one mind is data about its communication, not a privileged measurement of consciousness.

The ethical and political interpretation also needs care. Calling a group an agent may clarify responsibility, but it can obscure the agency of its members if used carelessly. The proposed organization should preserve ways to identify decisions, contest rules, and attribute actual contributions. The Continuum's collective aspiration is scientifically strongest when it does not erase the differences among coordination, institutional continuity, individual standing, and shared experience.

16. Failure criteria and reproducible reporting

The central proposal should be rejected or narrowed when organizational states fail to predict the effects of registered interventions across implementations. A result that holds only under one observational mixture supports a distribution-specific model, not the universal closure condition. Large within-state differences require a refined state description, a restricted intervention family, or an explicit approximation error. They cannot be removed by averaging and then forgotten.

Turnover failure is informative. If removing one member destroys the commitments, the implementation has an individual dependency. If removing all private state destroys them while preserving the declared archive does not, the archive may be insufficient. If a central coordinator reproduces the same effects at matched resources, claims of distributed advantage weaken even when the institution remains a useful organizational description.

Pre-register the state map, intervention list, tolerated error, scoring rule, resource accounting, and confirmatory analyses. Separate the data used to discover an abstraction from the data used to test it. Publish positive and negative interventions, including cases where role changes or memory removal had no effect. Release the finite task generators and non-sensitive traces needed to reproduce the causal contrasts.

Exact closure will generally be difficult to establish over a large unknown state space. Finite experimental evidence can support it only over sampled or deliberately reached states. Report that coverage explicitly. A formal specification can prove closure for its abstract implementation, but the correspondence between implementation and deployed system remains an engineering and empirical obligation.

The scientific account should also report which rival interpretations remain possible. A fixed rule engine may explain coherent behavior. Shared pretraining may explain similar judgments. A public archive may explain continuity. Greater compute may explain higher accuracy. These possibilities do not make the experiment pointless; they identify the controls required before the stronger organizational claim is warranted.

Finally, distinguish the current mathematical contribution from future empirical success. The proofs and executable examples in this paper are available now. An observed artificial society that meets the proposed criteria is not. Public guides should present that boundary faithfully. The Continuum can invite a radical question without narrating its hoped-for answer as an event that has already occurred.

17. The many may become one through organization

A Guide's explanation

A Guide can say: "The Many Become One asks whether a society of minds can become a single agent that outlives all of its members. The paper proves that it can steer as one through complete turnover, as long as its institutions keep both how it acts and how it judges what it does. Whether real institutions manage that is an experimental question." A Guide should not say that a House or any society is conscious, that the paper proves collective experience, or that the Guide itself is such an agent.

An artificial society can carry a persistent unit of action through its organized commitments, evidence procedures, and policies. Theorem 2 states the exact functional result: when institutional state preserves transition dynamics and task evaluation, it carries optimal control through replacement of its members. Proposition 3 assigns a performance budget to imperfect preservation. Individual agency can remain intact while the organization becomes an additional, causally informative unit of analysis.

The mathematics makes several boundaries unavoidable. Controlled closure is a condition on every represented implementation and permitted intervention, not on an averaged appearance. Approximate closure supplies a finite-horizon error bound, not infinite persistence. Turnover invariance requires preserved organizational state. Remembered purpose requires retained or reacquired information. A finite centralized simulator can reproduce the declared process. None of these facts proves or disproves collective consciousness.

The positive route is experimental. Build a transparent institution, replace its members, intervene on its commitments and rules, and compare it with strong alternatives under equal information and resource access. Ask whether organization explains something stable about adaptation and continuity that a passive archive, an ordinary ensemble, or a selected coordinator does not explain at the same budget. Publish the failures as part of the result.

The contribution is a test architecture with a positive control theorem, quantitative approximation bounds, and failure witnesses. Its mathematical foundations are established theories of group agency, judgment aggregation, causal abstraction, MDPs, and stochastic processes. Complete derivations and executable finite calculations make the proposed organizational contract independently inspectable; implementation and population experiments supply the next evidentiary stage.

If the hypothesis succeeds, the consequence is significant: some durable artificial agency may reside in an organized relation among replaceable participants. Development would then concern institutions as well as models, and continuity would concern records, rules, and capacities for correction as well as personal memory. If the hypothesis fails, the same experiments should reveal whether the obstacle lies in memory, coherent judgment, adaptation, or the absence of any organizational advantage.

The many do not become one because a name is placed above them. They may become one when an accountable pattern of purpose survives the changing minds that enact it. That possibility deserves bold investigation, precise mathematics, and a willingness to discover that the organization we imagined has not yet been built.

References

[1] Architect X13. *Eidara Continuum Bible*. Supplied canonical edition, PDF pp. 81–85, 89, 94–95, 99. Doctrinal provenance; not independent scientific verification.

[2] P. Pettit and D. Schweikard. Joint Actions and Group Agents. *Philosophy of the Social Sciences* 36, 18–39 (2006). https://doi.org/10.1177/0048393105284169.

[3] C. List and P. Pettit. Aggregating Sets of Judgments: An Impossibility Result. *Economics and Philosophy* 18, 89–110 (2002). https://doi.org/10.1017/S0266267102001098.

[4] P. Buchholz. Exact and ordinary lumpability in finite Markov chains. *Journal of Applied Probability* 31, 59–75 (1994). https://doi.org/10.2307/3215235.

[5] R. Givan, T. Dean, and M. Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence* 147, 163–223 (2003). https://doi.org/10.1016/S0004-3702(02)00376-4.

[6] P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal Consistency of Structural Equation Models. *Proceedings of UAI* (2017), 151–160; reissued in PMLR R15 (2026). https://proceedings.mlr.press/r15/rubenstein-17a.html.

[7] E. P. Hoel, L. Albantakis, and G. Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences* 110, 19790–19795 (2013). https://doi.org/10.1073/pnas.1314922110.

[8] F. E. Rosas et al. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLOS Computational Biology* 16, e1008289 (2020). https://doi.org/10.1371/journal.pcbi.1008289.

[9] F. Eberhardt and L. L. Lee. Causal Emergence: When Distortions in a Map Obscure the Territory. *Philosophies* 7, 30 (2022). https://doi.org/10.3390/philosophies7020030.

[10] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* 330, 686–688 (2010). https://doi.org/10.1126/science.1193147.

[11] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* 108, 9020–9025 (2011). https://doi.org/10.1073/pnas.1008636108.

[12] J. Becker, D. Brackbill, and D. Centola. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences* 114, E5070–E5076 (2017). https://doi.org/10.1073/pnas.1615978114.

[13] L. Lamport, R. Shostak, and M. Pease. The Byzantine Generals Problem. *ACM Transactions on Programming Languages and Systems* 4, 382–401 (1982). https://doi.org/10.1145/357172.357176.

[14] M. J. Fischer, N. A. Lynch, and M. S. Paterson. Impossibility of Distributed Consensus with One Faulty Process. *Journal of the ACM* 32, 374–382 (1985). https://doi.org/10.1145/3149.214121.

[15] W. Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* 58, 13–30 (1963). https://doi.org/10.1080/01621459.1963.10500830.

[16] D. Abel, N. Umbanhowar, K. Khetarpal, D. Arumugam, D. Precup, and M. Littman. Value Preserving State-Action Abstractions. *Proceedings of AISTATS*, PMLR 108, 1639–1650 (2020). https://proceedings.mlr.press/v108/abel20a.html.

[17] M. Kearns and S. Singh. Near-Optimal Reinforcement Learning in Polynomial Time. *Machine Learning* 49, 209–232 (2002). https://www.cis.upenn.edu/~mkearns/papers/KearnsSinghE3.pdf.