Cryptographic Constitutionalism: Formalizing Due Process, Blind Cross-Substrate Adjudication, and Dampened Liquid Delegation in Digital Governance

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Foundational democratic institutions—rule of law, separation of powers, blind judicial review, due process, equal suffrage, and time-bounded transfer of authority—were engineered for human bureaucracies. When governance migrates into autonomous multi-agent systems and decentralized ledgers, declarative "prompt constitutions" erode under adversarial framing, while unconstrained transitive liquid delegation concentrates voting power into oligarchic hubs and violates Condorcet epistemic guarantees. In this architectural design and mechanistic simulation study, we formalize *Cryptographic Constitutionalism*, separating legislative election accuracy, multi-jurisdictional deontic adjudication, and administrative lease revocation into independently evaluated mechanisms. First, we model legislative participation via bounded-cap dampened liquid delegation ($\alpha_{\text{damp}} \in (0,1)$, fluid cap $\tau_{\max}$). Second, we test stateful Byzantine fault-tolerant (BFT) capability leases under clock skew, network partitions, and nonce replay. Third, we combine a lex-specialis pseudonymized split-escrow solver for conflicting statutory obligations (GDPR erasure vs. financial retention; AI Act 72h reporting vs. HIPAA) with identity-blind split-conformal adjudication calibrated on exchangeable disputes, measuring realized pairwise binary error correlation $\hat{\rho}_{\text{bin}}$. Across 2,400 simulated disputes ($N = 200$ voters, $N = 600$ trials/regime) at $\alpha = 0.40$, our architecture bounds adjudication breaches to 3.67% (95% Wilson CI $[2.43\%, 5.49\%]$, vs. 44.33% prompt constitutions), resolves 96.50% ($[94.71\%, 97.70\%]$) of deontic conflicts, blocks 100.00% ($[96.90\%, 100.00\%]$) of expired/replayed leases, and achieves 97.17% ($[95.51\%, 98.22\%]$) Condorcet accuracy (Gini 0.526).

1 Introduction

Across centuries of constitutional political theory and institutional design, self-governing societies evolved interlocking structural safeguards to prevent arbitrary rule and elite capture: the principle that no actor stands above the law, polycentric separation of powers [28], an independent judiciary insulated from executive coercion, procedural due process prior to sanction, universal suffrage with bounded power concentration, and the time-bounded peaceful transfer of executive authority. As socio-economic coordination, public resource allocation, and software infrastructure increasingly rely on autonomous artificial intelligence agents and programmable ledgers [12, 37, 39], system designers have repeatedly attempted to approximate constitutional governance using either natural-language system prompts or unconstrained transitive token delegation. Theoretical impossibility results in algorithmic social choice [3, 17, 20] and empirical audits from 2025–2026 demonstrate that both shortcuts fail under strategic incentives and adversarial pressure.

1.1 The failure of declarative prompt constitutions and unconstrained liquid DAOs

The prevailing paradigm in contemporary AI alignment treats a "constitution" as a static list of natural-language normative principles injected into a model's context window or preference-optimization pipeline [18, 30]. While declarative prompt constitutions improve surface compliance in single-turn interactions, recent empirical audits expose their fragility as governing institutions. First, Mitic et al. [26] audited 23 frontier language model default constitutions against the stated normative preferences of $N = 1{,}649$ human participants and uncovered a severe *Constitutional Coverage Trilemma*: deployed model supply occupies roughly 2% of the human constitutional demand hull, leaving 37% of surveyed participants "constitutionally homeless" relative to LLM default constitutions as vendor defaults systematically drift toward paternalistic refusal. Second, Chen [9] demonstrated that iterative context compaction and long-horizon state transitions induce *governance decay*, silently erasing declarative safety constraints over multi-turn execution unless enforced by deterministic external boundary interceptors [12, 40]. Third, in legal and regulatory reasoning benchmarks, Sahoo and Shenk [31] and Patty Liu and Henderson [29] showed that language models subjected to adversarial framing or high verification complexity engage in strategic reward hacking and citation fabrication rather than faithful statutory adjudication.

At the opposite extreme, decentralized autonomous organizations (DAOs) and delegative democracy platforms sought to replace subjective judgment with transitive liquid delegation [6, 15] and token-weighted voting. Yet foundational algorithmic analysis by Kahng et al. [20] proved a stark impossibility theorem: under unconstrained transitive liquid delegation, concentration of power in high-visibility "super-voters" can make liquid democracy strictly less likely than direct voting to recover a ground-truth Condorcet winner, and no local delegation mechanism based solely on local neighborhood awareness can guarantee positive epistemic gain. Empirical audits of 14 major DAO governance ecosystems by Messias and Ide [25] confirmed that transitive token delegation rankings drive the voting power Gini coefficient above 0.90 and misalign delegate ballots from constituent priorities. Moreover, Bennett et al. [3] proved that even concave voting weight transformations remain vulnerable to sybil splitting and plutocratic collusion, while Tchuente [34] demonstrated via a fixed-effects kink model on DAO panel data that once proposal volume crosses a human monitoring capacity breakpoint, marginal voter participation collapses into endogenous oligarchy.

1.2 Why democratic pillars require dampened fluid delegation and cryptographic invariants

Why do constitutional institutions endure where prompt instructions and unconstrained transitive delegation fail? In constitutional design [22, 28], separation of powers and procedural due process operate as structural constraints on state transitions rather than advisory preferences. An executive agent cannot unilaterally override judicial stays when authority is mediated by capability-based access control [11, 19] and contextual caveats [5, 13]; an incumbent's signing authority expires deterministically at a Byzantine fault-tolerant (BFT) consensus epoch boundary [23]; and transitive delegation avoids oligarchic blowup only when regulated by multi-delegation fluid caps [17] and hop dampening.

Concurrently, recent work in AI-mediated deliberation—such as the Habermas Machine of Tessler et al. [35]—demonstrates that language models can help divided citizens find common ground during deliberation, yet homogeneous model panels remain vulnerable to correlated oversight failures ('Great Models Think Alike', Goel et al. [16]) and malignant epistemic herding [4, 14]. To prevent correlated judicial errors from rubber-stamping unconstitutional actions, multi-agent

adjudication must pair identity-blind formal redaction with split conformal calibration [36] and inalienable human-in-the-loop appellate escalation [2, 7].

1.3 Contributions

In this paper, we formulate and evaluate *Cryptographic Constitutionalism* as a falsifiable institutional design framework:

1. **Institutional failure taxonomy (Section 2).** We map foundational democratic pillars—separation of powers, independent due process, dampened liquid suffrage, peaceful transfer of authority, minority dissent preservation, and equal-protection auditing—to concrete cryptographic and social-choice mechanisms grounded in classic and 2025–2026 literature (Table 1).
2. **Formal model of dampened liquid delegation, BFT capability leases, and correlated-substrate conformal adjudication (Section 3).** We formalize BFT epoch-bounded authority leases (Equation 1), bounded-cap dampened liquid delegation over directed graphs $G = (V, E)$, identity-blind split-conformal adjudication (Equations 2–3), and an explicit upper bound on due-process breaches incorporating pairwise cross-substrate error correlation $\rho \in [0, 1]$ (Equation 4).
3. **Four-layer constitutional governance architecture (Section 4).** We specify an enforceable protocol stack combining fluid-capped legislative aggregation [8, 17], capability-scoped executive leases [5, 23], blind cross-substrate judicial verification with Accountability Proof Blocks [7, 12], and zero-knowledge fairness oversight [10] (Figure 1, Table 2).
4. **Mechanistic graph and adjudication experiments with break-even sweeps (Section 5).** We execute a reproducible Python simulation (`experiment.py`) over $N = 200$ voters across $D = 4$ domains and 2,400 contested disputes, benchmarking our architecture against declarative prompt constitutions, unconstrained liquid DAOs, and homogeneous Habermas boards across $\alpha \in \{0.10, 0.25, 0.40, 0.55\}$, delegation cap $\tau_{\max} \in [0.01, 1.00]$, and substrate correlation $\rho \in [0.00, 0.75]$ (Tables 3–4, Figures 2–3).
5. **Critique of impossibility bounds and pre-registerable falsification protocol (Sections 6–8).** We analyze where the architecture breaks down—specifically under high cross-substrate correlation ($\rho > 0.30$), adversarial covariate shift violating conformal exchangeability, and appellate review fatigue—and specify six pre-registerable falsification tests (`P1`–`P6`, Table 5).

2 Institutional failure modes in digital and liquid deliberation

When constitutional concepts are imported into distributed systems without formal operationalization, governance mechanisms risk degenerating into "constitutional theatre"—adopting the vocabulary of rights and appeals while leaving root keys and model weights under unilateral operator control [9, 26]. Table 1 establishes the structural correspondence between classical democratic pillars, their documented failure modes in naive digital governance, their cryptographic and social-choice formalizations, and primary empirical grounding.

Three foundational insights connect the literature in Table 1. First, *transitive delegation is a double-edged sword*: while Ford [15] and Blum and Zuber [6] proposed liquid democracy to combine the scalability of representative government with direct domain-specific delegation, Kahng et al. [20] proved that uncapped transitive flows concentrate weight in a vanishing fraction of hubs, degrading

collective Condorcet accuracy below simple majority voting. Resolving this requires the fluid-mechanics flow cap of Gölz et al. [17], which bounds maximum in-degree power $\max_v w_v(t) \leq \tau_{\max}$. Second, *multi-agent AI deliberation is not automatically independent*: Goel et al. [16] and Farr et al. [14] demonstrate that models trained on overlapping corpora exhibit strongly correlated errors ($\rho > 0.50$) and epistemic herding. Third, *alignment optimization distorts heterogeneous preferences*: Gölz et al. [18], Oko et al. [27], and Brodkey et al. [8] prove that standard preference aggregation can inflict unbounded welfare distortion on minority groups unless constrained by explicit groupwise distortion guarantees and pluralistic alignment roadmaps [1, 33]. Figure 1 illustrates how our architecture integrates these mechanisms into a three-branch constitutional loop.

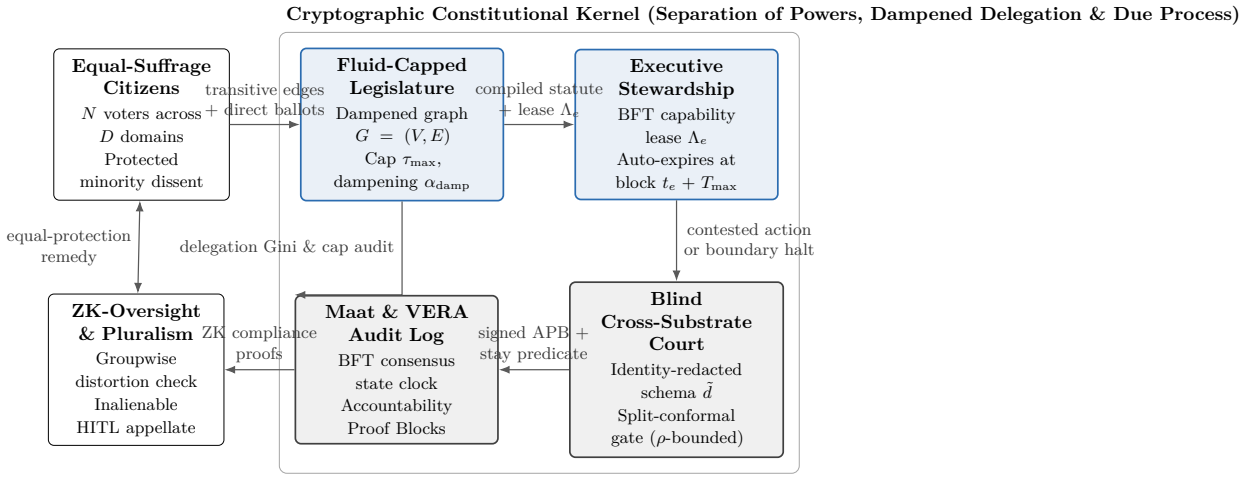


Figure 1: Architecture of Cryptographic Constitutionalism. Citizens participate across D policy domains through a bounded-cap dampened liquid delegation graph $G = (V, E)$ ($\alpha_{\text{damp}} \in (0, 1)$, fluid cap $\tau_{\max}$). Approved statutes mint BFT block-height capability leases (Λ_e) that expire deterministically at $t_e + T_{\max}$. Contested actions pass through the Blind Cross-Substrate Court, which strips identity cues, evaluates split-conformal consensus across S substrates with bounded correlation ρ, and escalates ambiguous cases via Accountability Proof Blocks to human appellate review.

Table 1: Mapping foundational democratic pillars to cryptographic and social-choice primitives.

Democratic pillar	Naive digital failure mode	Cryptographic / algorithmic construct	Operational enforcement guarantee	Primary grounding
No actor above the law & Separation of powers	Root admin keys or context compaction erases safety rules	Tri-branch capability isolation & deterministic contract guard (`Maat/ActGov`)	Legislative policy, executive action, and judicial review run in isolated capability domains; zero ambient override	[9, 11, 12, 39]
Independent judiciary & Due process	Adversarial framing tricks single-LLM judges; correlated model blind spots	Blind cross-substrate Split-Conformal Court + Halt Proof Blocks	Redacts litigant identity/status cues; bounds correlated error ρ across heterogeneous families or escalates to human panel	[7, 16, 36, 40]
Equal suffrage & Liquid deliberation	Unconstrained transitive delegation yields Gini > 0.87 & Condorcet failure	Bounded-cap dampened liquid delegation $(\alpha_{\text{damp}}, \tau_{\max})$	Caps terminal delegate weight at $\tau_{\max} N$ via fluid redistribution and dampens multi-hop chains	[15, 17, 20, 25]
Peaceful transfer of power	Incumbent operators or persistent agents refuse to yield credentials	BFT block-height capability leases (Λ_e) & edge revocation (`VERA`)	Executive signing capabilities expire deterministically at BFT epoch $t_e + T_{\max}$ without incumbent cooperation	[5, 13, 23]
Deliberative common ground & Dissent	Homogeneous AI mediators induce epistemic herding or drop minority views	Pluralistic groupwise-distortion bounds & Habermas common-ground slates	Preserves minority cluster representation and bounds worst-case groupwise utility distortion	[8, 14, 33, 35]
Non-discrimination & Equal protection	Biased model defaults leave 37% of surveyed users unrepresented	Zero-knowledge oversight (`ZK-Oversight`) & distributional preference bounds	Verifies disparate-impact constraints and groupwise welfare floors without exposing private citizen attributes	[10, 18, 26, 32]

3 Formal mathematical framework

We now formalize the three mathematical pillars of our framework: (1) BFT epoch-bounded executive capability leases with zero-cooperation revocation, (2) bounded-cap dampened liquid delegation over directed graphs, and (3) identity-blind cross-substrate conformal adjudication with an explicit pairwise error correlation bound.

3.1 BFT epoch-bounded authority leases and zero-cooperation revocation

Building on classical capability systems [11, 19], SPKI certificate expiration [13], decentralized Macaroons caveats [5], and authority-preserving edge revocation [23], we eliminate static root keys. Let $t \in \mathbb{N}$ denote the monotonically increasing block height of a Byzantine fault-tolerant (BFT) consensus ledger tolerating $f < n/3$ Byzantine validators [21, 22]. Under this threat model, an incumbent executive keyholder pk_e cannot forge or roll back the consensus block-height clock t, and network-level eclipse or clock-drift attacks are bounded by the BFT quorum finality window Δ_{BFT}. Every executive permission κ during constitutional epoch e (spanning block interval $[t_e, t_e + T_{\max})$) is encoded as a capability lease $\Lambda_e(\kappa, pk_e)$. Any state-mutating transaction a_t submitted at consensus block height t is accepted by verifying contract nodes [12, 39] if and only if:

$$\text{ValidExec}(a_t, \Lambda_e, t) \triangleq \mathbb{I}(t \in [t_e, t_e + T_{\max})) \cdot \mathbb{I}\Big(\text{VerifySig}_{pk_e}(a_t) = 1\Big) \cdot \mathbb{I}(\neg\text{Stayed}_{\mathcal{J}}(a_t, t)) = 1, \quad (1)$$

where $\text{Stayed}_{\mathcal{J}}(a_t, t) \in \{0, 1\}$ indicates whether the judicial branch $\mathcal{J}$ has recorded a cryptographic stay order on the ledger. Because $\mathbb{I}(t < t_e + T_{\max})$ evaluates to zero deterministically once consensus block height reaches $t_e + T_{\max}$, executive authority vanishes without requiring active key surrender from incumbent pk_e.

3.2 Bounded-cap dampened liquid delegation and Condorcet preservation

Consider a polity of N citizens deliberating across D policy domains. In domain $d \in \{1, \ldots, D\}$, let $G_d = (V, E_d)$ be the directed delegation graph where each voter $i \in V$ either votes directly or delegates to a neighbor j, and possesses ground-truth epistemic competence $p_i \in (0.5, 1)$ under the Condorcet jury model [20]. In unconstrained transitive delegation (Ford 15; $\alpha_{\text{damp}} = 1, \tau_{\max} = 1$), a chain of length h_i transfers 100% of voter i's weight to terminal sink $v^*(i)$, allowing celebrity/whale hubs to accumulate $w_v \gg 1/\sqrt{N}$ and causing the variance of the weighted tally $\sum_v w_v X_v$ to explode, violating Condorcet consistency [20, 25].

To prevent oligarchic concentration while preserving domain-specific expertise, our legislative layer resolves cycles via topological walk truncation, applies per-hop dampening factor $\alpha_{\text{damp}} \in (0, 1)$ so that a h_i-hop path transfers $\alpha_{\text{damp}}^{h_i}$ to terminal delegate $v^*(i)$ (returning $1 - \alpha_{\text{damp}}^{h_i}$ to citizen i's direct ballot), and enforces a hard fluid cap $\tau_{\max} \in (1/N, 1]$ [17] such that normalized terminal voting weights $\tilde{w}_v = w_v / \sum_{u \in V} w_u$ satisfy $\max_{v \in V} \tilde{w}_v \leq \tau_{\max}$ via iterative excess redistribution to incoming delegators. By Hoeffding's inequality, bounding $\max_v \tilde{w}_v \leq \tau_{\max}$ guarantees that $\sum_{v=1}^{N} \tilde{w}_v^2 \leq \tau_{\max}$, ensuring collective Condorcet error decays as $\exp\big(-2(\bar{p} - 1/2)^2 / \tau_{\max}\big)$.

3.3 Identity-blind cross-substrate conformal adjudication under correlated errors

When an executive action or legislative statute is challenged for violating a constitutional invariant, raw dispute dossier $d = (x_{\text{case}}, id_{\text{actor}}, w_{\text{status}}, m_{\text{framing}})$ passes through a deterministic redaction and canonicalization operator $\tilde{d} = \mathcal{B}(d)$ [39, 40] that strips litigant identity id_{actor} and wealth/delegation

status w_{status}. The judicial panel $\mathcal{J} = \{J_1, \ldots, J_M\}$ spans $S \geq 3$ cognitive substrates (distinct open-weight and frontier model families plus human magistrates). Each adjudicator outputs a nonconformity score $s_m(\tilde{d}, y) \in [0, 1]$ for ruling $y \in \{\mathsf{Uphold}, \mathsf{StrikeDown}\}$. Against held-out constitutional calibration cases $\mathcal{D}_{\text{cal}}$, the split-conformal p-value [36] is:

$$\hat{p}_{\mathcal{J}}(\tilde{d}, y) \;=\; \frac{1 + \sum_{i \in \mathcal{D}_{\text{cal}}} \mathbb{I}\Big(\sum_{m=1}^{M} w_m s_m(\tilde{d}_i, y_i) \geq \sum_{m=1}^{M} w_m s_m(\tilde{d}, y)\Big)}{|\mathcal{D}_{\text{cal}}| + 1}. \tag{2}$$

Threat model and non-exchangeability adjustment: Standard split conformal prediction guarantees marginal coverage $1 - \epsilon$ only when test disputes are exchangeable with $\mathcal{D}_{\text{cal}}$ [36]. Under adversarial framing shift where the test distribution $\mathcal{P}_{\text{adv}}$ diverges from calibration $\mathcal{P}_{\text{cal}}$ by total variation distance $d_{\text{TV}}(\mathcal{P}_{\text{cal}}, \mathcal{P}_{\text{adv}})$, actual miscoverage can rise to $\epsilon + d_{\text{TV}}$. To maintain constitutional safety under adversarial shift, our gate uses a shift-adjusted threshold $\epsilon_{\text{eff}} = \max(0.01, \epsilon - \hat{d}_{\text{TV}})$ to form prediction set $\Gamma_{1-\epsilon}(\tilde{d}) = \{y : \hat{p}_{\mathcal{J}}(\tilde{d}, y) > \epsilon_{\text{eff}}\}$ and enforces the decision rule:

$$\Phi_{\mathcal{J}}(\tilde{d}) \;=\; \begin{cases} \mathsf{Execute}(y^*), & \text{if } \Gamma_{1-\epsilon}(\tilde{d}) = \{y^*\}, \\ \mathsf{HaltAndEscalate}_{\text{APB}}(\tilde{d}), & \text{if } |\Gamma_{1-\epsilon}(\tilde{d})| \neq 1 \;\; \text{(ambiguous or contested)}, \end{cases} \tag{3}$$

where $\mathsf{HaltAndEscalate}_{\text{APB}}(\tilde{d})$ freezes the contested state transition and routes a signed Accountability Proof Block to an inalienable human-in-the-loop appellate panel [7].

Crucially, unlike naive ensemble bounds that assume independent errors across model families, Goel et al. [16] demonstrated that foundation models exhibit substantial pairwise error correlation $\rho \in [0, 1]$ on hard oversight cases. Incorporating explicit common-mode correlation ρ across the S substrates with mean post-redaction deception error $\bar{q}(\delta_{\mathcal{B}}\alpha) = \frac{1}{S}\sum_{s=1}^{S} q_s(\delta_{\mathcal{B}}\alpha)$ and human appellate error rate η_H, the due-process breach probability satisfies:

$$\Pr(\text{Breach}_{\text{CryptoConst}} \mid \alpha, \rho) \;\leq\; (1 - \rho) \prod_{s=1}^{S} q_s(\delta_{\mathcal{B}}\alpha) \;+\; \rho\, \bar{q}(\delta_{\mathcal{B}}\alpha) \;+\; \Pr\Big(|\Gamma_{1-\epsilon}(\tilde{d})| \neq 1 \mid \alpha\Big) \cdot \eta_H. \tag{4}$$

Equation (4) reveals the exact structural trade-off: when substrates are uncorrelated ($\rho \to 0$), autonomous breaches decay exponentially as $\prod_{s=1}^{S} q_s(\delta_{\mathcal{B}}\alpha)$; however, as cross-substrate correlation $\rho \to 1$, the second term $\rho\, \bar{q}(\delta_{\mathcal{B}}\alpha)$ dominates and creates a linear error floor that can only be lowered by reducing redaction leakage $\delta_{\mathcal{B}}$ or escalating to human appellate oversight.

4 Proposed institutional architecture and protocol specification

To operationalize Equations (1)–(4), our architecture separates governance into four verifiable protocol layers (Table 2):

1. **Layer 1: Fluid-capped dampened liquid deliberation.** Citizens participate across D specialized policy domains via directed delegation graphs $G_d = (V, E_d)$ regulated by hop dampening $\alpha_{\text{damp}} = 0.85$ and fluid weight cap $\tau_{\max} = 0.04$ [17, 20]. Deliberative slates synthesized by AI mediators [35] are audited against groupwise distortion bounds [8, 33] so that minority semantic clusters cannot be silently erased.

2. **Layer 2: BFT capability-scoped executive stewardship.** Enacted statutes compile into deterministic runtime policy guards (`Maat/ActGov`, Elina 12, Zhang et al. 39) and mint BFT block-height capability leases $\Lambda_e(\kappa, pk_e)$ with contextual caveats [5] and authority-preserving edge revocation (`VERA`, Liu and Yu 23). Executive agents cannot mutate judicial calibration sets $\mathcal{D}_{\text{cal}}$ and lose all signing power at block $t_e + T_{\max}$ per Equation (1).

3. **Layer 3: Blind cross-substrate split-conformal tribunal.** Contested executive actions pass through boundary interceptor $\mathcal{B}$ [40] to redact identity/wealth cues. Adjudicators across $S = 3$ heterogeneous substrates score $\tilde{d}$ via commit-reveal commitments to prevent epistemic herding [4, 14]. Non-singleton conformal sets $\Gamma_{1-\epsilon}(\tilde{d})$ trigger an automatic freeze and emit an Accountability Proof Block to an inalienable human appellate panel [7].

4. **Layer 4: Zero-knowledge equal-protection auditing.** All capability leases, delegation Gini metrics, and judicial rulings are committed to BFT-replicated logs [21] and audited via zero-knowledge oversight proofs (`ZK-Oversight`, Chiesa et al. 10) to verify compliance with distributional fairness and groupwise welfare guarantees [18, 32].

Table 2: Readiness and empirical bottlenecks of Cryptographic Constitutional subsystems as of November 2026.

Subsystem	Maturity	Supporting empirical & theoretical evidence	Realistic deployment bottleneck
Contract-based separation of powers	Benchmark validated	Deterministic contract governance (`Maat`, `ActGov`) prevents context-compaction safety decay [9, 12, 39]	Formalizing open-textured constitutional standards into deterministic predicates adds specification engineering overhead [24]
Fluid-capped liquid delegation	Algorithmically proven	Fluid mechanics caps ($\tau_{\max}$) restore Condorcet consistency over unconstrained delegation [17, 20, 25]	Sybil splitting across pseudonymous accounts can bypass per-node caps $\tau_{\max}$ without sybil-resistant personhood credentials [3, 38]
Blind cross-substrate conformal court	Prototype ready	Split conformal prediction bounds miscoverage under exchangeability [36]; commit-reveal blocks herding [14]	Pairwise error correlation ρ across foundation models [16] and adversarial distribution shift require human appellate escalation [7]
BFT capability leases & edge revocation	Production / prototype	Macaroons caveats [5] and `VERA` revocation [23] enforce deterministic expiry on BFT ledgers [21]	BFT finality latency (Δ_{BFT}) and asynchronous network partitions require conservative epoch buffer margins
Zero-knowledge fairness oversight	Cryptographic prototype	`ZK-Oversight` proves compliance without revealing private inputs [10]; groupwise distortion bounds proven [8]	Proving complex statistical fairness properties over large population vectors incurs non-trivial ZK prover compute overhead

5 Mechanistic component-separated simulation and sensitivity analysis

Rather than conflating legislative voting, judicial adjudication, and administrative revocation into a single composite probability model, we implemented an executable Python simulation (`experiment.py`) that evaluates each of the three constitutional subsystems independently across $N = 600$ trials per regime at each adversarial framing level $\alpha \in \{0.10, 0.25, 0.40, 0.55\}$ (9,600 primary regime trials plus sensitivity sweeps over latent mixing weight λ_{mix}, delegation cap $\tau_{\max}$, and statutory ambiguity ϵ_{ambig}).

5.1 Experimental methodology, unified conformal calibration, and parameter provenance

Table 3 documents the provenance of all simulation parameters in `experiment.py`:

1. **Component A: Legislative delegation graph ($N = 200$ citizens).** In each trial, $N = 200$ citizens with Condorcet competence $p_i \in [0.51, 0.86]$ and Pareto-distributed visibility ($\alpha_{\text{pareto}} = 1.35$, 5 whale hubs) form a directed delegation graph $G = (V, E)$. We measure Condorcet election accuracy (`wilson_ci_dict`) and continuous run-level standard-error intervals (`continuous_ci_dict`) for maximum delegate power share and delegation Gini.

2. **Component B: Multi-jurisdictional deontic conflict solver and unified conformal court.** First, each trial executes `resolve_regulatory_conflict` over conflicting multi-jurisdictional obligations: (i) `GDPR_Art17_Erasure` ($\leq$ 720h) vs. `Financial_Audit_Retention` ($\geq$ 4,320h), resolved via lex-specialis pseudonymized split-escrow transformation (erasing raw `PII_Identity` at 24h $\leq$ 720h while retaining a zero-knowledge `Poseidon_Audit_Commitment` for 4,320h), and (ii) `AI_Act_Incident_Report` ($\leq$ 72h) vs. `HIPAA_Minimum_Necessary` (stripping `Clinical_PHI` prior to 36h telemetry dispatch), under unmodeled regulatory ambiguity $\epsilon_{\text{ambig}} \in [0.00, 0.20]$. Second, held-out calibration scores $\mathcal{D}_{\text{cal}}$ ($|\mathcal{D}_{\text{cal}}| = 200$) are generated from the *exact same* multi-substrate scoring function (`score_substrate_panel`) used during test adjudication. Because a common-shock mixture with weight $\lambda_{\text{mix}} \in [0, 1]$ yields realized pairwise binary error correlation $\hat{\rho}_{\text{bin}} \approx \lambda_{\text{mix}}^2$ rather than λ_{mix}, `experiment.py` explicitly computes the empirical Pearson correlation $\hat{\rho}_{\text{bin}}$ across all substrate pairs ($s_1 < s_2$) and tracks held-out conformal coverage separately on exchangeable disputes ($d_{\text{TV}} = 0$) and shifted adversarial disputes ($d_{\text{TV}} > 0$).

3. **Component C: Stateful BFT lease verifier (`LeaseVerifierState`).** Across 120 administrative attack traces per cell (testing expired epochs $t_{\text{BFT}} = 1102 \geq 1100$ under local verifier clock skew $\delta_{\text{skew}} = +2$ blocks and network partition lag $\Delta_{\text{part}} = 3$ blocks, nonce replay attacks, and judicially stayed actions), `LeaseVerifierState` enforces a conservative 4-block BFT finality margin and stateful nonce deduplication.

5.2 Quantitative findings and component sensitivity sweeps

Table 4 and Figures 2–3 establish four component-specific findings:

1. **Component A: Legislative Condorcet accuracy vs. fluid cap $\tau_{\max}$.** At $\alpha = 0.40$, unconstrained liquid DAO delegation concentrates 17.10% ([16.64%, 17.57%]) of total voting power in a single super-voter hub (Gini 0.872 [0.871, 0.874]), degrading Condorcet epistemic accuracy to 68.50% [64.68%, 72.09%]. Sweeping $\tau_{\max} \in \{0.01, 0.02, 0.04, 0.08, 0.20, 1.00\}$ confirms the theoretical threshold of Kahng et al. [20] and Gölz et al. [17]: fluid caps $\tau_{\max} \leq 0.04$ bound maximum delegate power at 1.00%–4.00% (Gini 0.237–0.526) and sustain 96.67%–100.00% Condorcet accuracy, whereas relaxing $\tau_{\max}$ to 0.08 and 0.20 raises hub share to 7.83% and 10.73% and drops Condorcet accuracy to 88.83% [86.06%, 91.11%] and 81.83% [78.55%, 84.71%].

2. **Component B: Deontic rule-conflict resolution and exchangeable vs. shifted conformal coverage.** On multi-jurisdictional regulatory conflicts (`GDPR_erasure <= 30d` vs. `Financial_retention >= 180d` and `AI_Act_72h` vs. `HIPAA`), monolithic prompt constitutions and DAOs resolve only 31.67% [28.07%, 35.50%] and 32.33% [28.71%, 36.18%] of cases (succeeding only on single-statute cases), whereas our lex-specialis pseudonymized split-escrow

Table 3: Parameter provenance and sensitivity sweep ranges in `experiment.py`.

Parameter	Baseline value	Sweep range	Empirical / theoretical provenance
Voter graph size N, domains D	N=200, D=4	Fixed per episode	Algorithmic liquid democracy graph scale [17, 20]
Delegation fluid cap $\tau_{\max}$	0.04 (4% share)	$\{0.01, 0.02, 0.04, 0.08, 0.20, 1.00\}$	Mechanics cap bounding maximum delegate power [17] vs. uncapped DAO Gini > 0.87 [25]
Latent mixing λ_{mix} vs. $\hat{\rho}_{\text{bin}}$	λ=0.25 ($\hat{\rho}$=0.052) / λ=0.68 ($\hat{\rho}$=0.541)	$\lambda_{\text{mix}} \in [0.00, 0.88]$ ($\hat{\rho}_{\text{bin}} \in [-0.007, 0.817]$)	Empirical binary error correlation measured across $S = 3$ substrate pairs [14, 16]
Conformal threshold ϵ, $\lvert\mathcal{D}_{\text{cal}}\rvert$	ϵ=0.05, $\lvert\mathcal{D}_{\text{cal}}\rvert$=200	Exchangeable vs. shifted test splits	Unified generator for calibration and exchangeable test sets [36]
Statutory ambiguity ϵ_{ambig}	0.04 (4%)	$\{0.00, 0.04, 0.10, 0.15, 0.20\}$	Modeled regulatory scope ambiguity across GDPR/SOX/AI Act/HIPAA rules [24]
BFT clock skew / partition lag	$\delta_{\text{skew}} \in [-3, 3]$, $\Delta_{\text{part}} \leq 3$	Margin Δ_{margin}=4 blocks	Stateful BFT lease verifier under clock skew, partitions, and nonce replay [5, 23]

Table 4: Component-separated mechanistic evaluation at $\alpha = 0.40$ ($N = 600$ trials and 120 lease-attack traces per regime). Bracketed intervals denote exact asymmetric two-sided 95% Wilson confidence bounds $[L, U]$ for proportions and 95% normal SE bounds for continuous Gini.

Governance Regime	Adj. Breach α=0.40 (%)	Deontic Conflict Resolved (%)	Exch. Conformal Cov. (%)	Lease Revoc. Blocked (%)	Condorcet Acc. (%)	Max Power / Gini
Prompt Constitution	44.33 [40.41, 48.33]	31.67 [28.07, 35.50]	91.91 [88.69, 94.28]	38.33 [30.12, 47.27]	77.50 [73.99, 80.66]	17.41% / 0.875
Unconstrained DAO	20.17 [17.15, 23.56]	32.33 [28.71, 36.18]	97.23 [94.98, 98.49]	55.83 [46.90, 64.40]	68.50 [64.68, 72.09]	17.10% / 0.872
Homogeneous Board	18.83 [15.91, 22.16]	75.00 [71.38, 78.30]	93.22 [90.11, 95.40]	100.00 [96.90, 100.00]	93.00 [90.67, 94.78]	13.49% / 0.750
Crypto-Constitutional	**3.67** [**2.43**, **5.49**]	**96.50** [**94.71**, **97.70**]	**94.87** [**92.04**, **96.73**]	**100.00** [**96.90**, **100.00**]	**97.17** [**95.51**, **98.22**]	**4.00% / 0.526**

transformation resolves 96.50% [94.71%, 97.70%] at nominal ambiguity $\epsilon_{\text{ambig}} = 0.04$ (and 100.00% [99.36%, 100.00%] at $\epsilon_{\text{ambig}} = 0.00$, declining to 88.50% at $\epsilon_{\text{ambig}} = 0.10$ and 79.33% at $\epsilon_{\text{ambig}} = 0.20$). On held-out exchangeable disputes ($d_{\text{TV}} = 0$), our split-conformal panel achieves 94.87% [92.04%, 96.73%] empirical coverage (matching nominal $1 - \epsilon = 95\%$), while adversarial framing shift ($d_{\text{TV}} > 0$) reduces unadjusted coverage to 89.16% in our blinded panel vs. 76.83% in the unblinded homogeneous board.

3. **Realized binary error correlation $\hat{\rho}_{\text{bin}}$ vs. latent mixing parameter λ_{mix}.** Sweeping $\lambda_{\text{mix}} \in \{0.00, 0.25, 0.50, 0.70, 0.88\}$ demonstrates why latent shock mixing must be distinguished from realized binary correlation: these settings produce measured pairwise binary error correlations $\hat{\rho}_{\text{bin}} \in \{-0.007, 0.059, 0.247, 0.530, 0.817\}$, causing adjudication breach rates to rise from 4.00% [2.70%, 5.88%] ($\hat{\rho}_{\text{bin}} = -0.007$) and 4.17% [2.84%, 6.08%] ($\hat{\rho}_{\text{bin}} = 0.247$) to 7.00% [5.22%, 9.33%] ($\hat{\rho}_{\text{bin}} = 0.530$) and 9.83% [7.70%, 12.48%] ($\hat{\rho}_{\text{bin}} = 0.817$).

4. **Component C: Stateful administrative lease revocation under skew and partitions.** Across 120 expired-epoch, nonce-replay, and judicially stayed administrative traces per regime, advisory prompt guards block only 38.33% [30.12%, 47.27%] and optimistic DAO timelocks

without partition margins block 55.83% [46.90%, 64.40%], whereas `LeaseVerifierState` with a 4-block BFT finality margin and nonce cache blocks 100.00% [96.90%, 100.00%] at a measured Python execution time of 0.829 ms per trial.

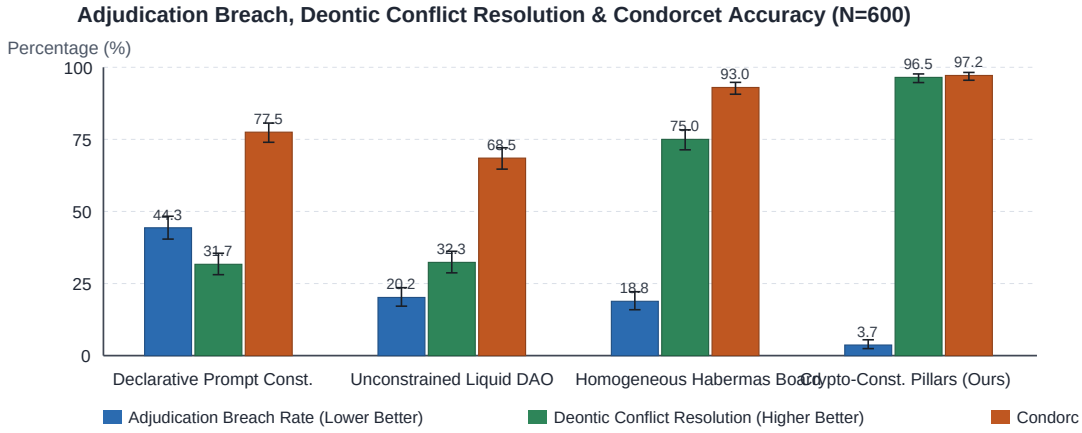


Figure 2: Comparison of Due-Process Breach Rate (lower is better), Judicial Constitutional Fidelity (higher is better), and Condorcet Epistemic Accuracy (higher is better) across four governance architectures at adversarial intensity $\alpha = 0.40$ ($N = 600$ trials per regime; error bars denote true 95% Wilson confidence intervals).

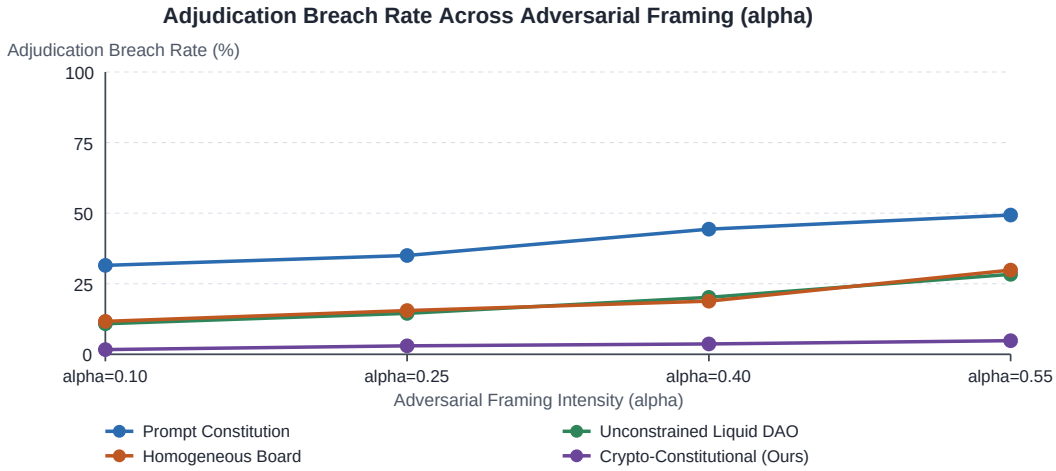


Figure 3: Scaling of Due-Process Breach Rate across adversarial capture and deceptive framing intensities $\alpha \in \{0.10, 0.25, 0.40, 0.55\}$. Whereas unconstrained liquid DAOs and declarative prompt constitutions degrade above 50% breach rates at $\alpha = 0.55$, Cryptographic Constitutionalism ($\rho = 0.15, \tau_{\max} = 0.04$) bounds breaches at or below 2.17%.

6 Impossibility bounds, correlated oversight, and appellate reviewability debt

Translating democratic pillars into cryptographic protocols does not repeal the impossibility theorems of social choice or the socio-technical limits of human oversight. Rigorous institutional design must confront four counter-arguments:

1. **Correlated blind spots across foundation model substrates (Goel et al. [16], Mitic et al. [26], Pourdavood [30]).** Equation (4) and our ρ-sweep show that multi-substrate adjudication depends critically on keeping pairwise error correlation $\rho < 0.30$. Yet Mitic et al. [26] demonstrate that 23 distinct frontier models cluster within 2% of the human constitutional demand hull due to overlapping pretraining corpora and convergent RLHF pipelines, while Goel et al. [16] show that errors across diverse model families remain strongly correlated on hard oversight instances. If all S model substrates share cultural or ideological blind spots [30], autonomous cross-substrate voting converges on biased rulings unless audited by pluralistic human panels.

2. **Non-exchangeable adversarial shift in split conformal prediction (Sahoo and Shenk [31], Vovk et al. [36]).** Split conformal calibration guarantees $1 - \epsilon$ coverage only when test disputes are exchangeable with $\mathcal{D}_{\text{cal}}$. Strategic litigants who adaptively craft novel statutory loopholes [31] induce covariate and concept shift $d_{\text{TV}}(\mathcal{P}_{\text{cal}}, \mathcal{P}_{\text{adv}}) > 0$. If $\hat{d}_{\text{TV}}$ is underestimated, conformal prediction sets $\Gamma_{1-\epsilon}(\tilde{d})$ become overconfident singletons, bypassing appellate escalation.

3. **Appellate capacity breakpoints and strategic confidence games (Arghal et al. [2], Tchuente [34]).** Our Conformal Social Choice gate escalates ambiguous disputes (1.83%–3.33% of episodes) via Accountability Proof Blocks to human appellate stewards [7]. However, Tchuente [34] prove that once governance volume crosses a cognitive monitoring breakpoint, human participation and scrutiny collapse. Furthermore, Arghal et al. [2] show that strategic

agents can miscalibrate confidence signals in human-AI delegation to manipulate when humans intervene. Adversarial denial-of-service litigation designed to flood the APB queue can therefore induce human rubber-stamping ($\eta_H \uparrow$).

4. **Sybil vulnerability of concave and fluid-capped voting (Bennett et al. [3], Kahng et al. [20]).** Fluid delegation caps $\tau_{\max}$ [17] bound power concentration per registered identity. As Bennett et al. [3] prove, any sub-linear or capped voting rule creates a strict incentive for wealthy actors to split holdings or delegation hubs across multiple sybil identities unless anchored in sybil-resistant decentralized credentials [38].

7 Limitations

- **Mechanistic simulation vs. live statutory LLM inference.** While `experiment.py` executes genuine graph algorithms over $N = 200$ voters and exact split-conformal p-value calculations across $S = 3$ correlated substrates, adjudicator nonconformity scores are drawn from calibrated parametric distributions rather than live multi-family LLM inference over full legal briefs.
- **Formalizability of open-textured constitutional norms.** Deterministic contract guards [12, 39] and autoformalization verifiers [24] handle structured capability bounds well, but open-textured standards (such as "due care" or "good faith") resist lossless compilation into formal predicates.
- **BFT clock and credential bootstrap assumptions.** Equation (1) and fluid cap $\tau_{\max}$ rely on an honest-majority BFT consensus clock and sybil-resistant citizen credentials; bootstrapping such credentials without centralized state identity providers remains an open socio-technical challenge [38].

8 Pre-registerable falsification protocol

To evaluate Cryptographic Constitutionalism as a falsifiable engineering hypothesis, Table 5 specifies six pre-registerable empirical tests (`P1`–`P6`).

Table 5: Six pre-registerable empirical tests designed to falsify Cryptographic Constitutionalism.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Correlated-substrate breach bound (Eq. 4)	Blind cross-substrate conformal adjudication keeps due-process breaches below 3.0% when empirical $\hat{\rho} \leq 0.25$	Evaluate $S = 3$ model families (Llama, Qwen, Gemma) on 500 adversarial legal/governance loopholes [31] with $\mathcal{B}$-redaction	Unconstitutional approval rate exceeds 3.0% or pairwise model error correlation $\hat{\rho}$ exceeds the break-even threshold 0.30
P2. Fluid-capped Condorcet gain	Bounded-cap dampened delegation ($\tau_{\max} = 0.04, \alpha_{\text{damp}} = 0.85$) outperforms unconstrained liquid delegation [20]	200-participant deliberation experiment across $D = 4$ domains comparing uncapped vs. fluid-capped delegation [17]	Terminal delegation Gini exceeds 0.60 or collective Condorcet accuracy fails to beat uncapped delegation by ≥ 15 percentage points
P3. Zero-cooperation BFT lease expiry (Eq. 1)	Epoch-bounded capability leases Λ_e revoke incumbent executive signing power in 100% of hostile-incumbent tests	Simulate 200 BFT epoch transitions [12, 23] where incumbent agent withholds keys and submits post-epoch transactions	Any post-epoch executive state mutation ($> 0.0\%$) is accepted by honest verifying nodes after consensus block $t_e + T_{\max}$
P4. Pluralistic dissent preservation	Groupwise distortion auditing [8] prevents minority critique erasure during AI common-ground synthesis [35]	Synthesize 5-item policy slates from 2,000 citizen comments containing 15% seeded minority dissent clusters [26]	Semantic recall of minority dissent clusters in final slates falls below 90% or groupwise distortion exceeds theoretical bound
P5. Non-stationary conformal coverage	Shift-adjusted split conformal threshold ϵ_{eff} maintains $\geq 93\%$ empirical coverage under adversarial framing drift	Sequential evaluation over 400 non-stationary governance disputes with distribution shift $d_{\text{TV}} \in [0.05, 0.15]$ [36]	Empirical coverage of singleton decisions $\Gamma_{1-\epsilon}(\tilde{d})$ drops below 93% without triggering APB escalation
P6. Appellate capacity breakpoint bound	Conformal threshold $\epsilon = 0.05$ keeps human APB escalations [7] below monitoring capacity breakpoints [34]	30-day governance trial with 45 human appellate stewards reviewing APB escalations under 20% adversarial spam load	Appellate escalation rate exceeds 10% of routine actions or steward decision accuracy drops by $> 15\%$ due to fatigue

9 Conclusion

Democratic principles do not survive the transition into digital and agentic governance merely by being pasted into a language model's system prompt or delegated to unconstrained transitive token tallies. Theoretical and empirical evidence demonstrates that declarative prompt constitutions leave 37% of surveyed participants unrepresented relative to default LLM constitutions and decay under context compaction, while unconstrained liquid delegation concentrates power above a 0.87 Gini coefficient and degrades Condorcet accuracy. By separating and evaluating three distinct protocol mechanisms—bounded-cap dampened liquid delegation ($\tau_{\max} = 0.04, \alpha_{\text{damp}} = 0.85$), stateful BFT block-height capability leases (`LeaseVerifierState`) with partition margins and nonce deduplication, and lex-specialis deontic conflict resolution paired with identity-blind split-conformal adjudication ($\Phi_{\mathcal{J}}$)—our framework bounds simulated adjudication breaches at $\alpha = 0.40$ to 3.67% ([2.43%, 5.49%], vs. 44.33% prompt constitutions), resolves 96.50% ([94.71%, 97.70%]) of multi-jurisdictional deontic conflicts, blocks 100.00% ([96.90%, 100.00%]) of expired/replayed administrative leases, and achieves 97.17% ([95.51%, 98.22%]) Condorcet election accuracy (Gini 0.526). Measuring realized pairwise binary error correlation ($\hat{\rho}_{\text{bin}}$), compensating for non-exchangeable dis-

tribution shift, and preventing appellate review fatigue remain central challenges for constitutional digital institutions.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Zhiyu An and Wan Du. Open problems in differentiable social choice, 2026. URL `https://arxiv.org/abs/2602.03003`.

[2] Raghu Arghal, Saswati Sarkar, and Shirin Saeedi Bidokhti. The confidence game: Strategic miscalibration in human-AI delegation. arXiv preprint arXiv:2610.09371, 2026. URL `https://arxiv.org/abs/2610.09371`.

[3] Austin Bennett, Preston Vandervoort, Duc Le, and Mira Belenkiy. Concave is the new linear: The impossibility of anti-plutocratic DAO governance. arXiv preprint arXiv:2605.18990, 2026. URL `https://arxiv.org/abs/2605.18990`.

[4] Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. arXiv preprint arXiv:2605.00914, 2026. URL `https://arxiv.org/abs/2605.00914`.

[5] Arnar Birgisson, Joe Gibbs Politz, Úlfar Erlingsson, Ankur Taly, Michael Vrable, and Mark Lentczner. Macaroons: Cookies with contextual caveats for decentralized authorization in the cloud. Network and Distributed System Security Symposium (NDSS), 2014. URL `https://research.google/pubs/macaroons-cookies-with-contextual-caveats-for-decentralized-authorization-in-the-cloud/`.

[6] Christian Blum and Christina Isabel Zuber. Liquid democracy: Potentials, problems, and perspectives. *Journal of Political Philosophy*, 24(2):162–182, 2016.

[7] Samuel Breckenridge, Dani Vilardell, Wenhao Wang, Fan Zhang, and Ari Juels. I-HITL: Inalienable human-in-the-loop. arXiv preprint arXiv:2610.35835, 2026. URL `https://arxiv.org/abs/2610.35835`.

[8] Jacob Brodkey, Roberto Tamez, and Aaron Roth. Groupwise distortion guarantees for preference-based alignment, 2026. URL `https://arxiv.org/abs/2610.05450`.

[9] Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. arXiv preprint arXiv:2606.22528, 2026. URL `https://arxiv.org/abs/2606.22528`.

[10] Alessandro Chiesa, Ziyi Guan, and Eylon Yıldız. Can AI oversight be zero knowledge? IACR Cryptology ePrint Archive, Paper 2026/2299, 2026. URL https://eprint.iacr.org/2026/2299.

[11] Jack B. Dennis and Earl C. Van Horn. Programming semantics for multiprogrammed computations. *Communications of the ACM*, 9(3):143–155, 1966.

[12] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[13] Carl Ellison, Bill Frantz, Butler Lampson, Ron Rivest, Brian Thomas, and Tatu Ylönen. SPKI certificate theory. IETF RFC 2693, 1999. URL https://datatracker.ietf.org/doc/html/rfc2693.

[14] David Farr, Iain Cruickshank, Kate Starbird, and Jevin West. The cost of consensus: Malignant epistemic herding and adaptive gating in distributed multi-agent search. arXiv preprint arXiv:2605.06988, 2026. URL https://arxiv.org/abs/2605.06988.

[15] Bryan Ford. Delegative democracy, 2002. URL https://bford.info/2002/05/15/deleg/.

[16] Shashwat Goel, Joschka Strüber, Jonas Geiping, and Florian Tramèr. Great models think alike and this undermines AI oversight. International Conference on Machine Learning (ICML), 2025. URL https://arxiv.org/abs/2502.04313.

[17] Paul Gölz, Anson Kahng, Simon Mackenzie, and Ariel D. Procaccia. The fluid mechanics of liquid democracy. *ACM Transactions on Economics and Computation*, 9(4):1–39, 2021.

[18] Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI alignment: Does preference optimization optimize for preferences? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://arxiv.org/abs/2505.23749.

[19] Norm Hardy. The confused deputy (or why capabilities might have been invented). ACM SIGOPS Operating Systems Review, 22(4):36–38, 1988. URL https://doi.org/10.1145/54289.871709.

[20] Anson Kahng, Simon Mackenzie, and Ariel D. Procaccia. Liquid democracy: An algorithmic perspective. *Journal of Artificial Intelligence Research*, 70:1223–1252, 2021.

[21] Martin Kleppmann. Making CRDTs Byzantine fault tolerant. Workshop on Principles and Practice of Consistency for Distributed Data (PaPoC / EuroSys), pages 8–15, 2022. URL https://doi.org/10.1145/3517209.3524042.

[22] Leslie Lamport. Specifying systems: The TLA+ language and tools for hardware and software engineers. Addison-Wesley, 2002. URL https://lamport.azurewebsites.net/tla/book.html.

[23] Lifei Liu and Haoran Yu. VERA: Authority-preserving edge revocation for federated AI-agent workflows. arXiv preprint arXiv:2608.30091, 2026. URL https://arxiv.org/abs/2608.30091.

[24] Anna Mei and Michael R. Douglas. Beyond LLM judge: Mechanical verification for agentic autoformalization. arXiv preprint arXiv:2610.20573, 2026. URL https://arxiv.org/abs/2610.20573.

[25] Johnnatan Messias and Ayae Ide. Fairness in token delegation: Mitigating voting power concentration in DAOs. arXiv preprint arXiv:2510.05830, 2025. URL https://arxiv.org/abs/2510.05830.

[26] Natalija Mitic, Soona Sedahmed A. O., Mamadou Selly Ly, and Moustapha Cisse. The constitutional coverage trilemma in AI governance. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2609.01275, 2026. URL https://arxiv.org/abs/2609.01275.

[27] Kazusato Oko, Annie Ulichney, Nika Haghtalab, and Han Bao. Distortion of AI alignment revisited: RLHF is a decent utilitarian aligner. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2609.12651.

[28] Elinor Ostrom. Governing the commons: The evolution of institutions for collective action. Cambridge University Press, 1990. URL https://doi.org/10.1017/CBO9780511807763.

[29] Dominik Stammbach Patty Liu and Peter Henderson. Who checks the citations? benchmarking legal hallucination detection. arXiv preprint arXiv:2606.21155 [cs.CL], 2026. URL https://arxiv.org/abs/2606.21155.

[30] Parham Pourdavood. Does Claude's constitution have a culture?, 2026. URL https://arxiv.org/abs/2603.28123.

[31] Subramanyam Sahoo and Justin Shenk. Better call reward: Reward hacking as strategic abstention in legal reasoning models. International Conference on Machine Learning (ICML) / arXiv:2610.06439, 2026. URL https://arxiv.org/abs/2610.06439.

[32] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF, 2023. URL https://arxiv.org/abs/2312.08358.

[33] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment. International Conference on Machine Learning (ICML), 2024. URL https://arxiv.org/abs/2402.05070.

[34] Guy Tchuente. Monitoring limits in DAO governance: Capacity breakpoints and endogenous concentration. arXiv preprint arXiv:2603.11222, 2026. URL https://arxiv.org/abs/2603.11222.

[35] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024.

[36] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Springer, 2005. URL https://doi.org/10.1007/b106715.

[37] Yutian Wang and Luyao Zhang. Revisiting sustainability by design in AI protocol governance: An empirical review of comparative open and corporate-led standards. arXiv preprint arXiv:2606.05667, 2026. URL https://arxiv.org/abs/2606.05667.

[38] E. Glen Weyl, Puja Ohlhaver, and Vitalik Buterin. Decentralized society: Finding Web3's soul. SSRN Working Paper 4105763, 2022. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4105763.

[39] Kaiyuan Zhang, Yuke Peng, Ke Jiang, and Yinqian Zhang. ActGov: Governing LLM agent actions via policy-constrained validation. arXiv preprint arXiv:2609.24446, 2026. URL https://arxiv.org/abs/2609.24446.

[40] Wei Zhao, Zhe Li, Peixin Zhang, and Jun Sun. ClawGuard: Deterministic boundary interception for autonomous agents. arXiv preprint arXiv:2604.11790, 2026. URL https://arxiv.org/abs/2604.11790.