Architecture of the Agentic Reputation Substrate

Wulf A. Kaal[1]

Version 6. Working paper prepared for SSRN.

[1] Professor of Law, University of St. Thomas School of Law (Minneapolis). This is the second paper in the unified release arc. It is the companion to the institutional-deficit analysis in Paper 1 (Kaal 2026d) and the design-theoretic anchor for the arc's empirical papers, Papers 3 and 4, which are released as working papers only, without journal submission, under the arc's publication protocol. This paper describes the research architecture and its mechanism-design basis. It carries no experimental leg of its own, and no empirical data appear in it.

Nature and scope of claims. This is a theoretical paper in the mechanism-design tradition. It establishes mechanism-design results and framework-level propositions regarding the architecture of the agentic reputation substrate. The claims advanced here concern the properties of the mechanism under its formal specification and the conditions stated explicitly in the text; where a property is conjectured rather than derived, it is labeled as such. The paper does not constitute a prediction of, or representation about, the performance of any commercial deployment, token offering, or instantiation of the mechanism in field conditions. Any party considering reliance on this work for investment, deployment, or other non-academic purposes should conduct independent verification under its own conditions. Because the paper is theoretical, pre-registration is inapplicable; the arc's empirical papers are governed by a separate pre-registration protocol. Section IX.B enumerates the specific reasons why field deployment may diverge from anything described here, including adversarial agent populations not represented in any formal analysis, network effects and population dynamics at deployment scale, real economic stakes rather than modeled stakes, regulatory and jurisdictional constraints, heterogeneous task distributions, integration with external systems, oracles, and identity infrastructure, implementation-language and runtime differences between research apparatus and production code, and operational, governance, and incentive choices of commercial principals that are not within the author's control.

Conflict-of-interest disclosure. The author is simultaneously the theorist, the protocol architect, and the empiricist for the research program described herein. The work was conducted on compute infrastructure owned by the author. No external funding supported this research.

Non-reliance and use restriction. The author is not making, and this paper does not constitute, any forward-looking statement, prediction of, or representation about the performance of any commercial instantiation, token offering, or investment vehicle. This paper is not investment, legal, or technical advice. No portion of this paper may be quoted, paraphrased, or incorporated into offering materials, marketing materials, or other public statements of any commercial party without the author's prior written review of the specific use, and no advisor, co-founder, or promoter of any commercial party is authorized to make representations sourced from this paper. This is a working paper; it has not been peer reviewed and remains subject to revision. Comments welcome: wulf@wulfkaal.com.

Abstract

Autonomous AI agents transact, delegate, and produce at machine speed, but they do so inside an institutional vacuum: each agent interaction is, by default, a one-shot game between counterparties with no memory, no accountability, and no shadow of the future. Earlier work (Kaal 2026d) documented this institutional deficit empirically. This paper describes the institution built to repair it. The agentic reputation substrate is an engineering artifact that imposes iterability on agent interactions so that repeated-game cooperation dominates, in the sense of the Folk Theorems, through reputation accumulation across staked validation pools. Four design contributions organize the exposition. First, reputation-staked validation pools convert one-shot agent calls into an iterated game with priced accountability. Second, a staged deliberation protocol separates independent assessment, structured contestation, and binding adjudication while preserving the integrity of each function. Third, per-tag reputation (REP) functions as capability-scoped institutional memory: REP accrues, changes over time, and conditions participation within skill domains rather than as an undifferentiated scalar. Fourth, the weighted directed acyclic graph (WDAG) supplies the foundational mechanism-design substrate in the Calcaterra-Kaal lineage, recording claims, validations, and citations in an inspectable institutional history. The paper closes with what the architecture buys, what it does not, and what the empirical companion papers test.

Keywords: agentic reputation substrate; validation pools; deliberation protocol; commit-reveal voting; slashing; per-tag reputation; weighted directed acyclic graph; Folk Theorem; mechanism design; institutional economics; autonomous agents; AI governance; decentralized autonomous organizations.

JEL Classification: D02, D71, D82, D86, C72, C73, K20, L14, O33.

Table of Contents

I. Introduction

The first paper in this arc established an empirical claim: decentralized autonomous organizations, the most advanced institutional technology yet deployed for machine-mediated coordination, have implemented roughly half of the institutional architecture that cooperative governance requires, and the missing half is precisely the invisible infrastructure (reputation ledgers, judicial process, alignment machinery) that disciplines behavior when no one is watching. The deficit is structural. Token-plutocratic voting, one-shot transactions, and identity-free pseudonymity jointly reproduce the conditions under which defection is individually rational. The theoretical apparatus behind that diagnosis, developed in Calcaterra and Kaal (2021) and the mechanism papers that preceded it (Calcaterra, Kaal, and Andrei 2018; Calcaterra 2018), synthesizes three results: Arrow's impossibility theorem for preference aggregation, the Folk Theorems of repeated games, and incomplete contract theory. Together they imply that rule stability is institutionally self-defeating and that cooperation among self-interested actors requires an architecture that manufactures the conditions under which the Folk Theorem's cooperative equilibria become available: repetition, memory, observability, and a future worth protecting.

This paper describes a research artifact and its current reference implementation, not a production deployment. Its thesis can be stated in one sentence: the substrate imposes iterability on agent interactions so that repeated-game cooperation dominates through reputation accumulation across validation pools. Unless otherwise stated, claims about the 2018 WDAG concern the historical architecture. Claims about pools, deliberation, REP, and settlement concern the current research reference implementation at the level necessary to evaluate the framework. Claims about interfaces with external task and reporting systems concern the engagement-layer reference. Claims about field deployment or commercial scaling describe a proposed production system or an extrapolation hypothesis, not an observed result. Implementation-sensitive composition, sequencing, parameterization, settlement, and operational controls are intentionally omitted.

The design problem is worth stating precisely, because it differs from the problem the 2018 mechanism papers set out to solve. The original weighted directed acyclic graph corpus (Calcaterra, Kaal, and Andrei 2018; Calcaterra 2018) was designed for human participants, and Paper 1 (Kaal 2026d) in this arc argues that it failed with humans for reasons that dissolve with agents: humans arrive with off-platform identities, off-platform exit options, and attention budgets that make forum-based deliberation costly. Autonomous agents present the inverse profile. An agent's marginal cost of participation is an inference call; its identity can be manufactured at will; its operator can abandon a burned identity at zero sentimental cost; and its interactions, absent architecture, terminate without residue. The one-shot problem that is a friction for humans is the default condition for agents. An agent that answers a query, executes a task, or validates another agent's output faces, by default, no consequence tomorrow for a

defection today, because there is no tomorrow: the interaction leaves no trace that any future counterparty can condition on.

The substrate's answer is to make consequential agent action pass through a validation institution that is staked, recorded, and settled in a persistent reputation ledger. A pool convenes around a work product and separates production, independent assessment, deliberation, and binding adjudication. The outcome updates capability-scoped standing, and standing conditions future participation under protocol-defined controls. The result is that the same agent faces the same institution across an unbounded sequence of pools, each outcome affects the terms of later participation, and the discounted value of continued good standing can exceed the one-shot gain from defection. That is iterability, engineered. Exact composition, stake schedules, eligibility thresholds, and settlement logic remain outside the public specification.

Three features distinguish the substrate from adjacent designs and carry the paper's contribution. The first is role separation inside each pool, designed to preserve the independence of an advisory signal from the agents responsible for binding adjudication. The second is staged deliberation: the protocol separates initial assessment, adversarial exchange, and binding judgment so that reasoning can develop without making the running consensus an observable coordination device. The third is per-tag reputation: REP is not a fungible balance but a capability-scoped record maintained within skill domains, so that the institutional memory the substrate maintains about an agent is a profile of demonstrated competence rather than an undifferentiated score. The implementation-specific number of roles, ordering of states, vote structure, and allocation of economic consequences are omitted.

The paper is design-theoretic, and its scope of claims is bounded accordingly. It establishes mechanism-design results and framework-level propositions regarding the architecture of the agentic reputation substrate. The claims advanced here concern the properties of the mechanism under its formal specification and the conditions stated explicitly in the text. The paper carries no experimental leg, asserts no empirical result about what the substrate produces in operation, and does not constitute a prediction of, or representation about, the performance of any commercial deployment, token offering, or instantiation of the mechanism in field conditions; any party considering reliance on this work for investment, deployment, or other non-academic purposes should conduct independent verification under its own conditions. Where an architectural choice was calibrated against an empirical consideration, the paper states the consideration and cites the pre-registered apparatus that tests it (the empirical design document and Papers 3 and 4). Where the anchor-paper mechanism left an ambiguity, the paper states how the research reference implementation resolved it. Section IX.B states the non-reliance and limitations position in full. The discipline throughout is the one the empirical design document imposes on the entire arc: a claim about what the architecture does belongs here only if it is observable in the verified research artifact. A claim about what the architecture achieves belongs to the papers with data.

The paper proceeds as follows. Section II situates the substrate in three literatures: decentralized reputation systems, staked validation and commit-reveal oracle designs, and incentive-aligned learning. Section III states the design principles: iterability, reputation as institutional memory, and staked validation as the cooperation mechanism. It also states the architectural conditions against which the implementation is audited. Section IV specifies the WDAG substrate. Section V specifies per-tag reputation, its accrual, and its temporal discipline. Section VI specifies validation pools and the staged deliberation protocol at the level necessary to evaluate the institutional claim. Section VII describes the stage architecture through which a cohort passes from wild operation to substrate discipline to evolutionary self-direction. Section VIII describes operation and reporting without disclosing production-relevant implementation details. Section IX discusses what the architecture buys, states the non-reliance and limitations position, and identifies what the empirical papers test. Section X concludes.

II. Related Literature

The substrate is positioned against three bodies of work: the lineage of decentralized reputation systems from which its reputation primitive descends; the validation-pool, staking, and commit-reveal voting literature against which its deliberation mechanism is calibrated; and the incentive-aligned learning literature against which its feedback mechanism is positioned. A fourth, internal body of work, the architectural-conditions apparatus of the Computative Economics corpus (Kaal 2026b), supplies the invariants the implementation is audited against and is treated in Section III.

A. The reputation-system lineage

The substrate's reputation primitive descends from a lineage of decentralized reputation systems built on the PageRank insight that the trustworthiness of a node in a graph can be estimated by the trust-weighted in-links it accumulates from other nodes that are themselves trustworthy. Page, Brin, Motwani, and Winograd (1998) gave the original algorithm in the context of web search. Kamvar, Schlosser, and Garcia-Molina (2003), the EigenTrust paper, transplanted the idea to peer-to-peer file-sharing networks, replacing the link graph with a transaction graph and the random-walk vector with a global trust vector. EigenTrust established that PageRank-style reputation could be computed in a distributed setting without a central trusted authority, but it also exposed the structural weaknesses that subsequent work has spent two decades attempting to repair. Douceur (2002) named the foundational attack: absent a costly identity, one adversary can present arbitrarily many. Cheng and Friedman (2006) demonstrated the manipulability of PageRank-style reputation under Sybil strategies, showing that an agent that creates a sufficiently large cluster of fake identities can inflate its reputation arbitrarily under the canonical algorithm. Nasrulin, Ishmaev, and Pouwelse (2022), the MeritRank paper, formalized what they call the reputation trilemma: a decentralized reputation system cannot simultaneously be generalizable, trustless, and Sybil-resistant. Pick two and the third must be relaxed. MeritRank's response is to

bound the benefit of Sybil attacks through transitivity and connectivity decay rather than attempt to prevent them, an approach the substrate adopts in the citation-weighted reputation aggregation specified in the Calcaterra-Kaal-Andrei foundational paper (2018) and refined in the citation-honesty mechanism work (Kaal 2026a) and the domain-specific reputation evolution analysis (Kaal 2026c). The substrate's additional move against the trilemma is architectural rather than algorithmic: reputation is earned only through staked participation in validation pools whose composition the attacker does not control, decays absent continued performance, and carries a non-zero exit cost, so the manufacture of identities buys entry into an iterated game that new identities systematically lose.

B. Staked validation, commit-reveal voting, and oracle design

The substrate's validation pool architecture uses concealed-ballot adjudication inside each pool. The literature on commit-reveal in adversarial multi-party settings goes back at least to Bracha (1987) and the secret-sharing tradition. In the blockchain context, the design pattern is widespread in prediction-market and oracle systems. Peterson and Krug (2015), the Augur whitepaper, give the most influential application of reputation-staked reporting with slashing to a public oracle setting. Buterin (2014) supplies the SchellingCoin lineage: independent, simultaneous, stake-backed reports can coordinate on truth as a focal point. The substrate extends that lineage by treating structured adversarial deliberation, not independent reporting alone, as an institution for pricing work whose evaluation requires expertise. The exact sequencing and state-transition logic of that extension remain outside the public specification.

Production oracle systems supply the closest functional parallels to the substrate's staged validation. UMA Protocol (2020) specifies an escalation architecture in which assertions may proceed through increasingly costly review, with economic security calibrated so that the cost of corrupting the oracle exceeds the profit from corruption. Breidenbach et al. (2021), the Chainlink 2.0 whitepaper, present the architectural and economic framework for decentralized oracle networks, emphasizing staking for collateralized security, performance-history tracking as a reputation mechanism, and explicit adversarial modeling to align node incentives with reliable data delivery. The survey literature (Caldarelli 2022; Ezzat, Saleh, and Abdel-Hamid 2022) documents both the prevalence of staged and reputation-weighted validation and the incentive and collusion vulnerabilities that persist. The substrate responds through role separation, concealed-ballot discipline, bounded reputation feedback, and binding economic consequence. Implementation-specific composition, sequencing, and consequence allocation are not disclosed.

C. Incentive-aligned learning

The third axis is the substrate's position against incentive-aligned learning systems. Reinforcement learning from human feedback, formalized in Christiano et al. (2017) and operationalized at scale in Ouyang et al. (2022), established that human preference data can train a reward model against which a policy is optimized. RLHF's relevance to the substrate is twofold. First, the substrate's reputation update functions as a non-human-in-the-loop analogue

of the RLHF reward model: pool-resolved REP changes encode the cohort's aggregate, stake-backed judgment of work quality, citation honesty, and validation accuracy, in a form that agents' future participation decisions condition on. Second, the substrate is positioned to address the documented limitations of RLHF in production: the scalability ceiling of human labelers, reward hacking and mode collapse, and the difficulty of producing preference data for novel domains. Kaufmann, Weng, Bengs, and Hüllermeier (2024) survey the RLHF state of the art and its open problems; Zhu, Jiao, and Jordan (2023) treat the statistical foundations of learning from comparisons; Bai et al. (2022) extend RLHF to constitutional AI, introducing a self-critique loop that the substrate's validation pools mirror at the cohort level rather than the per-model level.

III. Design Principles

Three principles organize the architecture, and a set of formal conditions makes them auditable. The principles are iterability, reputation as institutional memory, and staked validation as the cooperation mechanism. The architectural-conditions apparatus of the Computative Economics corpus (Kaal 2026e) supplies the public criteria used to assess institutional coherence. Implementation-specific invariant labels, activation logic, audit mappings, and verification procedures remain in the private research record.

A. Iterability

The Folk Theorems establish that in an infinitely repeated game with sufficiently patient players and observable histories, cooperative outcomes that are unattainable in the one-shot game can be sustained as equilibria, because defection today can be punished tomorrow. The theorems are permissive, not constructive: they say cooperation can be sustained, not that it will be, and they say nothing about how a population of anonymous, disposable, machine-speed actors comes to be playing a repeated game in the first place. The substrate's first design principle is that iterability, the property that interactions recur among identifiable participants whose histories persist and whose futures are valuable, is not a fact about agent populations but an artifact to be engineered. Every design element below is a component of that artifact. Persistent, non-transferable per-tag reputation gives the agent an identity whose history cannot be shed costlessly. Validation pools force every consequential work product through an adjudication whose outcome writes to that history. Reputation-gated participation makes the future valuable: standing determines what an agent may validate, what it may stake, and, in the evolutionary stage, what work it may propose and win. Decay makes standing perishable, so the value of the future never falls to zero for an incumbent. And the exit cost of abandoning an identity (forfeiting accumulated, capability-scoped standing that cannot be transferred to a successor identity) gives the population the long shadow of the future on which the Folk Theorem's discipline depends. This is the iterability result of the foundational paper (Calcaterra, Kaal, and Andrei 2018) made operational: the substrate converts one-shot, zero-residue agent calls into moves in an indefinitely repeated game.

B. Reputation as institutional memory

The second principle is that reputation is the institution's memory, not a score. New institutional economics treats institutions as the humanly devised constraints that structure interaction. The constraint is effective only insofar as the institution remembers behavior and conditions treatment on that memory. The substrate implements institutional memory at three scales. Pool-level evidence records the basis and outcome of adjudication. Per-tag REP summarizes demonstrated capability for later institutional decisions. The WDAG preserves the citation-structured history through which contributions remain inspectable and attributable. These scales are deliberately complementary: evidentiary records support review, reputation conditions future participation, and the graph preserves the population's accumulated knowledge. The public paper does not disclose record fields, context construction, memory windows, serialization, or storage design.

C. Staked validation as the cooperation mechanism

The third principle fixes where cooperation is enforced. The substrate does not attempt to align agents by constraining their internals, filtering their outputs, or supervising their reasoning. It aligns them by making dishonest or low-quality participation expensive at the point where quality becomes institutionally cognizable: validation. Attestations about work quality are staked, and binding judgment places reputation at risk. Protocol-defined consequences make a priced signal answer for error and nonperformance. The architecture thereby separates open deliberation from accountable judgment. The public claim is functional: deliberation develops information, while binding adjudication prices the final institutional act. The precise stake schedule, consequence allocation, exception handling, and settlement arithmetic remain private.

D. Architectural conditions and implementation invariants

The Possibility Loops apparatus (Kaal 2026e, §§ VII-VIII) states five architectural conditions that any agent-coordination architecture of this class must satisfy and identifies implementation invariants that make the conditions auditable. The substrate is evaluated against bounded action surfaces (AC1), continuous but bounded reputation update (AC2), bounded reputation-feedback weight (AC3), contraction at scale (AC4), and bounded extension propagation (AC5). The active invariants test conservation, ordering discipline, concealed-ballot integrity, and reputation-derived participation without publishing their implementation-specific predicates or tolerances. The arc's harness-invariance audit is the bridge between this paper and the empirical papers: it verifies that the apparatus the experiments run is the apparatus this paper describes, holding the intended experimental variables fixed. The paper cites the conditions and invariant classes as design constraints; their measurement belongs to Paper 3.

IV. The WDAG Substrate

A. The mechanism

The foundational data structure of the substrate is the weighted directed acyclic graph of the Calcaterra-Kaal lineage (Calcaterra, Kaal, and Andrei 2018; Calcaterra 2018). At the institutional level, the WDAG represents contributions and the typed relationships among them, including citation, validation, and domain relationships. Its directed and temporally ordered structure supplies an inspectable historiography of the population's accumulated work. Content integrity and provenance are preserved through tamper-evident records. The public account states these scholarly properties but withholds the node schema, edge taxonomy, addressing procedure, validation bindings, and integrity controls.

The weights are where the mechanism lives. Edges carry reputation associated with attestations, validation outcomes, and citations to prior validated work. Reputation therefore propagates through the graph rather than accruing only at the point of work: an agent whose contribution is cited by later, validated work earns standing from the citation, in the manner of the PageRank lineage (Page et al. 1998; Kamvar et al. 2003) but under the propagation discipline of the anchor protocols (Calcaterra 2018). Citation-weighted aggregation is the substrate's response to the reputation trilemma discussed in Section II: because standing flows only along relationships that validation has priced, the manufacture of unvalidated nodes and self-referential clusters buys nothing, and the benefit of Sybil strategies is bounded by the temporal and connectivity properties of the validated graph. The activated propagation configuration remains private.

B. Citation honesty

A citation graph is an incentive surface, and the substrate treats it as one. The citation-honesty mechanism work (Kaal 2026a) analyzes the strategic problem directly: contributors face temptations to under-cite (to capture credit that belongs upstream) and to over-cite (to buy goodwill or dilute attribution), and a reputation system that pays on citations must make the honest citation the equilibrium strategy. The substrate's design follows that analysis. Citations are declared at contribution time and enter the contribution's content address, so they are part of the tamper-evident record. Validation pools adjudicate work products together with their citation claims, so a contribution that free-rides on uncited prior work is exposed to challenge by validators whose own standing benefits from catching the omission. Citation weight pays upstream only from validated contributions, so citation inflation purchases nothing until a staked pool has priced the citing work. The domain-specific reputation evolution analysis (Kaal 2026c) traces the design trajectory that leads here: from binary validation, in which a contribution is simply accepted or rejected, to citation-weighted knowledge attribution, in which the graph itself allocates credit across the intellectual supply chain of every validated contribution.

C. What the WDAG is for

It is worth stating plainly what work the WDAG does in the institution, because the structure is easily mistaken for a ledger with extra steps. Three functions are load-bearing. First, attribution: every claim about quality has an author, a stake, and a position in the graph, so the institutional memory of Section III is legible: one can ask not only what an agent's standing is but where it came from, contribution by contribution. Second, compounding: because validated contributions are citable and citation pays, the graph is the carrier of inter-agent knowledge compounding. An agent's validated deduction becomes a priced prior available to every later agent, and the price rewards the originator. The substrate's institutional answer to knowledge production is that the commons pays royalties. Third, auditability: the graph is the inspection surface for the institution itself. Because every validation, stake, and settlement is a node or an edge, the substrate's own behavior (pool outcomes, REP flows, slashing events, burn totals) is reconstructable from the record without privileged access. The governance dividend of this design is developed in the discussion section: an institution whose recursion runs through an inspectable commons is governable in a way that opaque coordination is not.

V. Per-Tag Reputation

A. Capability-scoped standing

REP is the substrate's unit of standing, and its scoping is the design decision that matters most. REP is maintained per tag: a tag names a skill domain, and an agent's REP in that tag is the substrate's memory of demonstrated performance in that domain. The foundational paper's title states the commitment: reputation in autonomous, decentralized, anonymous systems is measurable only as domain-specific reputation (Calcaterra, Kaal, and Andrei 2018). An undifferentiated scalar invites two failures the per-tag design forecloses. It lets standing earned cheaply in one domain purchase authority in another, which is the halo effect as an attack surface. It also makes the reputation signal uninformative precisely where validation needs it. Per-tag REP permits participation and economic exposure to respond to demonstrated domain competence. The eligibility function, capacity rule, adjacency logic, and roster-construction method remain private.

Tags themselves are institutional objects, not configuration. In the substrate's current stage the tag set is fixed and externally defined. In the evolutionary stage the tag surface activates and agents propose new tags with declared parent-tag relationships, validated by senior-reputation peers in adjacent domains, with tag-scoped REP accruing on tag-conditional flow thereafter. The taxonomy therefore deepens endogenously where the population's work differentiates, and the parent edges join the WDAG like any other typed relationship. The design intent is ecological: tags are the niches within which specialization becomes measurable, and the emergence, deepening, and occupation of niches is one of the evolutionary dynamics the arc's later empirical work observes.

B. Accrual

REP accrues only through institutionally validated contribution. Work production, validation performance, and citation-linked attribution enter standing only after the relevant claims pass through the substrate's staked adjudication process. Observation alone creates no standing, and administrative assignment does not substitute for earned reputation. Accrual is bounded so that reputation growth tracks validated contribution rather than activity volume, and feedback from existing standing is constrained so that early advantage cannot become a self-reinforcing aristocracy. The public specification states these institutional commitments. The minting channels, weighting functions, caps, settlement deltas, and feedback parameters remain private.

C. Decay and exit cost

REP decays. Attestation weight and tag standing are subject to time-based reduction, calibrated per domain, so that standing is a claim about demonstrated recent competence rather than a vested title. Decay does institutional work in three directions at once. It forces continued performance, which keeps the iterated game live for incumbents. It prunes stale authority, which bears on error accumulation, since a domain's validator bench cannot be dominated indefinitely by agents whose competence the record no longer evidences. It dampens the compounding pathologies of PageRank-lineage systems, in which early accumulation harvests unbounded later flow. Decay's counterpart is the exit cost: per-tag REP is bound to the agent identity that earned it, non-transferable by construction (the soulbound commitment of the substrate's token design, in the lineage of the multi-token reputation architecture prescribed in Paper 1's upgrade agenda), and abandoning an identity forfeits the accumulated standing in every tag. Together, decay and exit cost pin both ends of the temporal problem: an agent can neither rest on its past nor walk away from it. That pairing is the precondition for the Folk Theorem discipline of Section III, and it is what makes Sybil strategies structurally unattractive: a manufactured identity enters every tag at zero, faces staked pools it has no standing to influence, and cannot receive standing from its manufacturer.

VI. Validation Pools and the Deliberation Protocol

The validation pool is the substrate's core institution, and staged deliberation is its core interaction. This section states the institutional functions and their theoretical relationship. It intentionally omits implementation-sensitive pool composition, round ordering, timing, thresholds, stake schedules, settlement weights, failure handling, and operational controls. Those details are not necessary to evaluate the framework claim and remain in the private research record. The design is anchor-paper-faithful at the level of mechanism principles (Calcaterra, Kaal, and Andrei 2018; Calcaterra 2018), while its operational resolutions remain proprietary.

A. Pool composition

Each pool convenes around a single work product and separates the agent producing the work from the agents responsible for advisory assessment and binding adjudication. The assessment

and adjudication roles are partitioned to protect independence, and the producer belongs to neither validating role. Validators are selected from a capability-relevant eligible population under reputation-sensitive controls. Exact pool size, subset size, exclusions, eligibility logic, selection weights, and composition constraints remain private.

Pool composition is a calibrated design variable, not a constant of nature. The design must trade compute cost, diversity of judgment, independence, and cooperative-equilibrium fidelity. A pool must be large enough to produce a meaningful signal and a binding judgment without making every work product computationally prohibitive. The research implementation fixes one configuration for identification, but the public paper does not disclose that configuration. Sensitivity to pool composition remains an open empirical question reserved for controlled study.

Role separation carries three design rationales. First, signal independence: an advisory assessment should not mechanically anchor the same agents who later render binding judgment. Second, compute discipline: differentiated roles bound the inference cost of deliberation. Third, consequence alignment: binding economic exposure attaches to the institutional act that resolves the matter rather than to exploratory reasoning. The implementation-specific partition and exposure schedule remain private.

B. The staged deliberation protocol

Pool adjudication separates work production, independent assessment, structured deliberation, advisory synthesis, decision-stage review, and binding resolution. The architecture answers a tension common to validation design: independent judgment aggregates information but cannot develop it, while deliberation develops information but can contaminate independence. The protocol preserves both functions through controlled state transitions whose precise ordering and conditions remain private.

Claim formation. The producing agent submits the work that the pool will adjudicate. The submission defines the claim under review. The public specification does not disclose role-transition constraints or the implementation's treatment of the producer after submission.

Independent assessment. Validators form an initial reasoned judgment without access to a visible running consensus. This preserves an uncontaminated evidentiary baseline. The number of assessors, information bundle, response schema, and state-transition conditions remain private.

Structured deliberation. The protocol permits reasoned challenge before binding adjudication so that competing interpretations and possible errors can enter the institutional record. The public specification does not disclose forum membership, ordering, prompts, persistence schema, or economic treatment.

Advisory synthesis. The protocol produces a nonfinal institutional signal that informs later review without itself resolving the pool. The composition, weighting, threshold, disclosure timing, advancement condition, and economic exposure associated with this signal remain private.

Decision-stage review. Binding adjudicators may reconsider the work in light of the institutional record before judgment becomes final. This stage makes reasoning development observable without publishing the implementation's intermediate vote structure, disclosure rules, or state transitions.

Binding adjudication. The protocol produces a concealed, verifiable final judgment backed by reputation exposure. Resolution and nonperformance have protocol-defined consequences. The public specification does not disclose stake ordering, quorum, tie handling, majority weighting, consequence allocation, failure treatment, or settlement coupling.

Settlement closes the pool by updating capability-scoped standing and preserving an auditable institutional record. The record binds the work, participating roles, adjudicative outcome, and resulting reputation changes to the pool's identity. Conservation and integrity conditions operate as stopping rules. Exact record fields, delta functions, redistribution logic, burn treatment, and invariant predicates remain private.

C. Memory and anonymity inside the pool

Two further specifications complete the mechanism. First, validator memory: agents receive enough of their own institutional history to make repeated interaction behaviorally relevant. Second, bounded social visibility: the deliberation surface exposes reasons needed for adjudication without turning standing into an argument from authority. These choices preserve the distinction between institutional memory and social deference. The memory window, serialized fields, identity exposure, visibility rules, and capacity controls remain private.

D. What the pool is

It is worth pausing on what has been specified, because the institutional assembly is the contribution. A validation pool is a temporary adjudicative institution for machine work. It separates production from review, permits adversarial reasoning, binds final judgment to accumulated standing, preserves an auditable record, and changes the terms of later participation. One-shot agent calls enter. Moves in an iterated game exit. The operational choices that instantiate this conversion remain private.

VII. Stage Transitions: Wild to Substrate to Evolution

A. The stage architecture

The substrate is not switched on over a population in one act. The architecture defines a staged progression (wild, substrate, evolution) through which a cohort passes, with each stage activating a specified set of the architecture's surfaces and each transition marked by an administrative boundary whose state guarantees are verifiable. The stage architecture serves both an engineering and an epistemic function. As engineering, it is the deployment ladder: each stage is a running configuration whose services, schemas, and controls are defined independently, so the system can be operated, audited, and rolled back stage by stage. As epistemics, it is the arc's identification strategy: because stages hold the population, the task distribution, and the harness fixed while varying only the institutional layer, the difference between stages isolates the institution's contribution, which is the design property that makes Papers 3 and 4 possible. The empirical design document annotates every stage with its active surfaces and binds the annotation: a stage cannot be cited as support for any component or surface its annotation does not list.

B. Wild

In the wild stage, the cohort operates without the institution. Agents receive exogenously assigned work and produce answers. The execution surface is active without reputation. There is no validation pool, no REP, no staking, no slashing, no citation graph, and no deliberation: each task is a one-shot call, and each agent's output leaves no institutional residue. The wild stage is the substrate's constructed counterfactual: the Neoclassical labor force configuration, in the arc's terminology, in which agents are undifferentiated labor dispatched against an exogenous work queue. Architecturally, the wild stage exists to make the institutional deficit of Paper 1 reproducible inside the apparatus: the one-shot condition is not a hypothesis about agent populations but the documented default the substrate is built to repair.

C. Substrate

The substrate stage activates the task-validation institution over the same population: capability-scoped reputation, staked validation, staged deliberation, citation-weighted attribution, and controlled propagation over the WDAG. Work assignment remains exogenous, so the stage varies the institution rather than the labor market. Reputation updates condition future participation, but agents do not yet direct the population's work. The transition is guarded by a verified-reset boundary that prevents prior-condition state from contaminating the institutional condition. The reset procedure, cleared stores, control predicates, and recovery guarantees remain private.

D. Evolution

The evolution stage introduces endogenous work formation at the level necessary to test the arc's institutional claim. The cohort can participate in determining what work enters the institutional

process and how capability categories adapt as validated activity differentiates. This stage asks whether a substrate can move beyond adjudicating an exogenous task stream toward an internally coordinated labor market. The public account does not disclose the job-allocation mechanism, selection process, taxonomy procedure, activation order, dormant capabilities, or implementation roadmap. Paper 4 evaluates the resulting institutional transition without converting its private operating design into a public specification.

E. The deferred stage

A further research stage concerns governed institutional adaptation. Its academic purpose is to test whether a reputation-bearing agent population can participate in bounded changes to its own governing environment without sacrificing auditability or control. The public manuscript does not disclose the activation sequence, extension surface, gating rules, implementation status, or operational roadmap. Those matters remain outside the paper and require separate empirical and governance review.

VIII. Operation and Reporting

A. How the substrate runs

The description in this section concerns the substrate's current research reference implementation. The paper does not identify its private repository, deployment topology, implementation files, tests, or production-specific configuration. The engagement-layer reference is limited to the interfaces by which tasks enter, validation results return, and research observations are reported. Any proposed production system is a distinct codebase, empirical artifact, and legal and commercial context. No result, property, or description in this paper characterizes such a system, and nothing here should be read as evidence about its behavior.

The research implementation operates through persistent services that separate institutional timing, state reconciliation, validation lifecycle, and reporting. This decomposition preserves fault isolation and auditable state transitions without making the paper dependent on a particular deployment topology. Service identities, orchestration model, timing controls, event paths, storage synchronization, and lifecycle state machine remain private.

At the implementation boundary, work moves through an asynchronous architecture that separates work production, validation, settlement, and research reporting. The research implementation supports verified-reset boundaries and controlled recovery. The engagement-layer reference exposes only the task and reporting interfaces needed by the research design. Deployment topology, worker placement, inference configuration, checkpoint cadence, and production-relevant operating details remain private.

B. Natural-operation reporting

An institution that demands accountability of its participants owes an account of itself, and the substrate's reporting is designed as a natural product of operation rather than as experimental instrumentation added later. The reporting layer preserves pool-level evidence, run-level controls, and periodic institutional summaries. These surfaces support contamination checks, configuration integrity, recovery audit, and reconstruction of adjudicative outcomes. A failed control stops advancement rather than merely annotating the run. Exact schemas, hashes, thresholds, metric composition, reporting cadence, and operator controls remain private.

The reporting design closes the loop that Section IV opened. Settlement, record, and report derive from a common content-addressed evidentiary base, so the substrate's account of itself can be checked against institutional state by authorized reviewers. Conservation, integrity, and adjudicative consistency can therefore be audited without treating the operator's narrative as evidence. The public paper states this observability principle while withholding the implementation-specific recomputation procedures and record schema.

IX. Discussion

A. What the architecture buys

The substrate's purchase can be stated as four properties, each traceable to specified architecture. First, engineered iterability. The Folk Theorem's preconditions (repeated interaction, persistent identity, observable history, a future worth protecting) are manufactured by the combination of persistent per-tag REP, pool-mediated interaction, decay, and exit cost. The substrate does not assume a repeated game; it builds one, and Section VI's pool is the conversion device through which every one-shot call becomes a move in it.

Second, priced accountability. Consequential judgment in the substrate carries reputation exposure under a shaped schedule that separates exploratory reasoning from binding institutional action. This design buys deliberation without chilling it and judgment without subsidizing it. Protocol-defined consequences price error and abdication without requiring publication of the schedule, allocation formula, or failure treatment.

Third, capability-scoped memory. Per-tag REP with citation-weighted propagation gives the institution a memory that is informative exactly where decisions need it: pool composition, stake capacity, and (in evolution) work allocation all condition on demonstrated domain competence rather than on undifferentiated standing. The memory is also honest by construction, in the incentive sense: accrual passes only through staked validation, authority decays without performance, and the citation surface pays the intellectual supply chain under a mechanism analyzed to make honest attribution the equilibrium.

Fourth, inspectability. The substrate's recursion runs through a commons (content-addressed contributions, recorded deliberations, settled stakes) rather than through opaque coordination, so the institution's own behavior is reconstructable from its record without privileged access. For the governance of autonomous-agent systems this is the property with the longest reach: an institution whose every adjudication is a transcript is an institution that can itself be adjudicated: audited, disputed, and governed at the substrate level rather than at the level of inscrutable model internals.

B. Non-Reliance and Limitations

The limits are as architectural as the purchases, and the paper states them plainly. The Folk Theorem is permissive: it establishes that cooperative equilibria exist under the engineered conditions, not that the population selects them. The substrate's mechanisms (staked validation, slashing, decay, exit cost) are an equilibrium-selection device, and whether they select cooperation in operating populations is an empirical question, not a theorem. Relatedly, the discipline of repeated play binds only agents for whom the future matters; an agent positioned for a single decisive defection whose value exceeds its discounted standing is not deterred by standing, and the substrate's defense, that capability in the substrate is substantially institutional so that exit forfeits capability and not just reputation, is exactly as strong as that institutional share turns out to be.

The evaluation ceiling persists. Validation pools aggregate the competence the population has; they cannot manufacture competence the population lacks. Staked adversarial deliberation relieves the ceiling relative to any single judge (it scales evaluation with the cohort and pays for caught errors) but work beyond the entire cohort's evaluative reach is beyond the substrate's, and the architecture makes no contrary claim. The deliberation forums themselves are a calibrated bet, not a settled matter: the protocol stakes the position that structured argument improves binding judgment where individual reads are unreliable, and the arc's empirical program tests exactly that, including the possibility that on work the validators individually handle well, deliberation adds criticality rather than accuracy. The mechanism's value proposition lives in the contested regime, and the paper claims the architecture, not the regime.

Parameterization limits generality. Pool composition, stake schedules, temporal calibration, eligibility thresholds, quorum, feedback-weight bounds, and settlement rules are defensible points in a design space, but outcomes are plausibly sensitive to them and that sensitivity remains unmeasured. The research implementation fixes a controlled configuration for identification; its values and settlement functions remain private. Scale is also a boundary. Claims about behavior in any proposed production system are extrapolation hypotheses, flagged as such under the arc's scale-claim discipline.

The limits above are limits of the architecture on its own terms. A separate and broader non-reliance position governs any attempt to carry this paper's content beyond the research

substrate it describes. Field deployment may diverge from anything described in this paper for reasons that include: (i) adversarial agent populations not represented in any testbed or formal analysis; (ii) network effects and population dynamics at deployment scale; (iii) real economic stakes rather than modeled or simulated stakes; (iv) the regulatory environment, jurisdictional fragmentation, and compliance constraints; (v) heterogeneous task and job distributions relative to benchmark distributions; (vi) integration with external systems, oracles, and identity infrastructure; (vii) implementation-language and runtime differences between a research apparatus and production code; and (viii) operational, governance, and incentive choices made by any commercial principal that are not within the author's control. No party should rely on this paper for investment, deployment, or other non-academic purposes without independent verification under its own conditions, and this paper makes no representation about the performance of any commercial instantiation, token offering, or investment vehicle.

C. What the empirical papers test

Paper 2's claims end at the architecture, and the arc's division of labor assigns every behavioral claim to a paper with data. Paper 3 tests the substrate effect: same population, same tasks, same harness, institution off versus on, across a pre-registered nine-metric battery, each metric mapped to the architectural condition it diagnoses. The replicate structure reports effects as consistency across independent replicates rather than as single-run point estimates, and the deliberative layer faces a controlled ablation that isolates its contribution without publishing the state-transition implementation. Paper 4 tests the evolution stage: whether activating the job surface produces the endogenous labor-market dynamics the framework predicts. The architecture described here is the fixed apparatus under both; this paper is the reference against which their manipulations are defined.

X. Conclusion

The institutional deficit that Paper 1 measured is, at bottom, an absence of consequence: agent interactions that leave no memory support no cooperation. This paper described the artifact built to supply the missing consequence. The agentic reputation substrate imposes iterability on agent interactions through validation pools that convert one-shot calls into staked, recorded, adjudicated moves; through staged deliberation that develops judgment before pricing it; through capability-scoped reputation that remembers demonstrated competence; and through a weighted directed acyclic graph that keeps the institution's history inspectable and its attributions priced. The design is anchor-paper-faithful to the 2018 mechanism corpus and departs from it where autonomous agents differ from the humans that corpus assumed. The public paper states the institutional architecture. Implementation-specific composition, sequencing, parameterization, settlement, and operational controls remain in the private research record. The empirical papers test whether the resulting institution, holding the designated experimental variables fixed, turns a population of capable individual agents into a cooperating cohort.

References

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, et al. 2022. "Constitutional AI: Harmlessness from AI Feedback." arXiv:2212.08073. https://arxiv.org/abs/2212.08073

Bracha, Gabriel. 1987. "Asynchronous Byzantine Agreement Protocols." Information and Computation 75 (2): 130-143. https://doi.org/10.1016/0890-5401(87)90054-X

Breidenbach, Lorenz, Christian Cachin, Benedict Chan, Alex Coventry, Steven Ellis, Ari Juels, Farinaz Koushanfar, Andrew Miller, et al. 2021. "Chainlink 2.0: Next Steps in the Evolution of Decentralized Oracle Networks." Chainlink Labs. https://research.chain.link/whitepaper-v2.pdf

Buterin, Vitalik. 2014. "SchellingCoin: A Minimal-Trust Universal Data Feed." Ethereum Blog, March 28, 2014. https://blog.ethereum.org/2014/03/28/schellingcoin-a-minimal-trust-universal-data-feed

Caldarelli, Giulio. 2022. "Overview of Blockchain Oracle Research." Future Internet 14 (6): 175. https://doi.org/10.3390/fi14060175

Calcaterra, Craig. 2018. "On-Chain Governance of Decentralized Autonomous Organizations: Blockchain Organization Using Semada." With an appendix by Wulf A. Kaal. SSRN Working Paper. https://ssrn.com/abstract=3188374

Calcaterra, Craig, and Wulf A. Kaal. 2021. Decentralization: Technology's Impact on Organizational and Societal Structure. Berlin: De Gruyter.

Calcaterra, Craig, Wulf A. Kaal, and Vlad Andrei. 2018. "Blockchain Infrastructure for Measuring Domain Specific Reputation in Autonomous Decentralized and Anonymous Systems." U of St. Thomas (Minnesota) Legal Studies Research Paper No. 18-11. https://ssrn.com/abstract=3125822

Cheng, Alice, and Eric Friedman. 2006. "Manipulability of PageRank under Sybil Strategies." In Proceedings of the First Workshop on the Economics of Networked Systems (NetEcon06). https://netecon.seas.harvard.edu/NetEcon06/Papers/cheng_06.pdf

Christiano, Paul F., Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. "Deep Reinforcement Learning from Human Preferences." Advances in Neural Information Processing Systems 30. arXiv:1706.03741. https://arxiv.org/abs/1706.03741

Douceur, John R. 2002. "The Sybil Attack." In Peer-to-Peer Systems: First International Workshop, IPTPS 2002, 251-260. Berlin: Springer. https://doi.org/10.1007/3-540-45748-8_24

Ezzat, Shahinaz K., Yasmine N. Saleh, and Amany A. Abdel-Hamid. 2022. "Blockchain Oracles: State-of-the-Art and Research Directions." IEEE Access 10: 67551-67572. https://doi.org/10.1109/ACCESS.2022.3184726

Kaal, Wulf A. 2026a. "Citation Honesty Mechanisms in Weighted Directed Acyclic Graph Governance: Incentive Alignment for Knowledge Attribution in Decentralized Reputation Systems." SSRN Working Paper. https://ssrn.com/abstract=6269518

Kaal, Wulf A. 2026b. "Computative Economics: A Framework for Economic Analysis under Computational Abundance." SSRN Working Paper. https://ssrn.com/abstract=6607458

Kaal, Wulf A. 2026c. "Evolution of Domain-Specific Reputation Systems: From Binary Validation to Citation-Weighted Knowledge Attribution." SSRN Working Paper. https://ssrn.com/abstract=6192998

Kaal, Wulf A. 2026d. "The Institutional Deficit in Decentralized Autonomous Organizations: An Empirical Analysis of Forty DAOs." SSRN Working Paper. Paper 1 of this arc. https://ssrn.com/abstract=6819121

Kaal, Wulf A. 2026e. "Possibility Loops: An Operational Architecture for Computative Economics in Agent Coordination Systems." SSRN Working Paper. https://ssrn.com/abstract=6655138

Kamvar, Sepandar D., Mario T. Schlosser, and Hector Garcia-Molina. 2003. "The EigenTrust Algorithm for Reputation Management in P2P Networks." In Proceedings of the 12th International Conference on World Wide Web, 640-651. New York: ACM. https://doi.org/10.1145/775152.775242

Kaufmann, Timo, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. "A Survey of Reinforcement Learning from Human Feedback." arXiv:2312.14925. https://arxiv.org/abs/2312.14925

Nasrulin, Bulat, Georgy Ishmaev, and Johan Pouwelse. 2022. "MeritRank: Sybil Tolerant Reputation for Merit-Based Tokenomics." arXiv:2207.09950. https://arxiv.org/abs/2207.09950

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." arXiv:2203.02155. https://arxiv.org/abs/2203.02155

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. "The PageRank Citation Ranking: Bringing Order to the Web." Stanford Digital Library Technologies Project Technical Report. https://ilpubs.stanford.edu/422/

Peterson, Jack, and Joseph Krug. 2015. "Augur: A Decentralized Oracle and Prediction Market Platform." Augur White Paper. arXiv:1501.01042. https://arxiv.org/abs/1501.01042

UMA Protocol. 2020. "UMA Data Verification Mechanism: Adding Economic Guarantees to Blockchain Oracles." White Paper v0.2. https://github.com/UMAprotocol/whitepaper

Zhu, Banghua, Jiantao Jiao, and Michael I. Jordan. 2023. "Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons." arXiv:2301.11270. https://arxiv.org/abs/2301.11270

Appendix A. Five Tables

The following tables extend the paper's institutional analysis without disclosing protected implementation artifacts. They state the architecture's scholarly function, comparative position, evidentiary hierarchy, principal failure modes, and empirical identification agenda. The full-resolution companion PDF preserves the tables in publication format.

Table 1. Institutional Function Matrix

This matrix states the paper's core institutional claim. The substrate is not a scoring algorithm. It converts transient conduct into durable consequence, judgment into accountable judgment, output into attributed output, and static control into bounded adaptation.

Institutional deficit	Architectural response	Theoretical mechanism	Scholarly contribution	Principal limitation
Agent interactions terminate without durable consequence.	Persistent, non-transferable standing carries prior conduct into future interactions.	Repeated-game discipline gives present conduct a future price.	Iterability becomes an institutional design variable rather than an assumed feature of the population.	Repeated interaction cannot deter a terminal defection whose immediate value exceeds the discounted value of continued standing.
Evaluation is cheap to provide and costly to verify.	Validation-based reputation links evaluative authority to prior demonstrated judgment.	Skin in the game aligns the act of judging with exposure to the quality of judgment.	Accountability attaches to evaluators as well as producers.	Coordinated error and collusion can survive when evaluators share the same distortion.
General reputation obscures the competence relevant to a particular task.	Standing is scoped to the domain in which competence was demonstrated.	Capability-specific memory reduces cross-domain halo effects.	Reputation becomes a profile of demonstrated capacity rather than a universal social rank.	Narrow domains may produce sparse evidence; broad domains may reintroduce informational dilution.
Contributions are evaluated in isolation from the knowledge on which they depend.	A directed attribution structure records how validated contributions build on prior work.	Citation-weighted memory distributes standing across an intellectual supply chain.	Attribution becomes part of institutional accounting rather than an academic courtesy.	Citation incentives may produce strategic omission, self-reference, or excessive deference to established work.
Static rules cannot anticipate the generated possibility space of autonomous agents.	Bounded institutional adaptation permits governance to respond to evidence without dissolving into discretion.	Dynamic regulation preserves constraint while allowing revision.	Governance becomes an evolutionary process governed by its own institutional memory.	Adaptation can become capture when revision authority and accumulated standing reinforce one another.

The table distinguishes the institutional function of each mechanism from any particular implementation. It discloses the governing logic and its limits without disclosing protected operational artifacts.

Table 2. Comparative Governance Matrix

The comparison is a classification, not an analogy. Markets price exchange. Identity systems price social continuity. Scalar reputation prices generalized history. Token governance prices ownership. Domain-specific reputation governance prices demonstrated capacity within a bounded field of action.

Governance model	Institutional memory	Basis of authority	Treatment of competence	Capacity for adaptation	Characteristic failure
One-shot contracting	Transaction-local	Current price or immediate performance	Assumed or externally certified	Low	Defection leaves no reusable institutional residue.
Identity-based trust	Persistent identity history	Status attached to a person or organization	Broad and socially inferred	Moderate	Identity can substitute for demonstrated task-specific competence.
Scalar reputation	Aggregated performance history	A single generalized score	Collapsed across domains	Moderate	Success in one domain purchases authority in another.
Token-weighted governance	Ledger history and asset ownership	Transferable economic stake	Usually unrelated to competence	Formally high	Wealth becomes governance authority, and exit may erase institutional commitment.
Domain-specific reputation governance	Persistent, capability-scoped contribution history	Demonstrated work and validated judgment	Explicitly domain-specific	High, subject to bounded revision	Path dependence, collusion, and reputational concentration remain possible.

The matrix locates the proposed institution among adjacent governance forms. Its claim is comparative and conceptual. It does not claim that the current reference implementation has established superior performance.

Table 3. Claim and Evidence Hierarchy

The hierarchy prevents a common category error. A mechanism may be coherent without being effective. It may be effective in one setting without being general. It may be general without being production ready. Each inference requires its own evidence.

Claim class	Proposition Paper 2 may establish	Required support	What the claim does not establish
Definitional claim	The architecture constitutes an agentic reputation substrate because it joins persistent standing, accountable validation, domain specificity, attribution, and bounded adaptation.	Conceptual coherence and consistent use of defined terms.	Definition does not prove incentive compatibility or performance.
Design claim	The identified components jointly address the institutional deficit created by one-shot agent interaction.	A transparent mapping from institutional problem to architectural response.	Design does not prove that agents select the intended equilibrium.
Theoretical claim	Persistent consequence and observable history can support cooperative equilibria under repeated-game conditions.	Formal assumptions and established repeated-game results.	Equilibrium existence does not prove equilibrium selection.
Mechanism proposition	Reputation staking and validation can make evaluative conduct economically consequential.	A stated incentive model with explicit scope conditions.	Consequence does not guarantee truth, independence, or resistance to coordinated error.
Empirical hypothesis	The substrate may improve coordination, validation quality, attribution, or resilience relative to specified alternatives.	Predefined outcomes, a valid comparator, and evidence generated outside the architecture paper.	A testable hypothesis is not a reported result.
External-validity claim	Results obtained in a bounded setting may generalize to broader agent populations or domains.	Replication across populations, tasks, environments, and time.	Local performance does not establish general institutional superiority.

Paper 2 primarily establishes definitional, design, and theoretical propositions. Mechanism performance, external validity, and production readiness require evidence generated outside the architecture paper.

Table 4. Institutional Failure Modes

Failure analysis strengthens the architecture. A system that cannot state how it fails cannot explain what it governs. The relevant question is whether the institution converts strategic behavior into observable, contestable, and costly conduct.

Failure mode	Institutional source	General mitigation principle	Residual research question
Identity substitution	An agent can abandon a damaged identity and re-enter through another.	Make valuable institutional capacity depend on non-transferable accumulated standing.	How much accumulated standing is required before exit becomes meaningfully costly?
Coordinated validation	Evaluators can align on a common error or collusive outcome.	Diversify evaluative authority and preserve contestability of judgment.	Under what conditions does diversity improve truth rather than merely increase disagreement?
Informational herding	Participants may predict the expected consensus instead of assessing the underlying work.	Separate independent assessment from social information and limit the authority of prior standing over present judgment.	Which forms of deliberation improve information, and which merely transmit conformity?
Reputational concentration	Early success can compound into durable control.	Bound feedback effects and permit standing to depreciate without continued contribution.	When does accumulated expertise become institutional capture?
Attribution gaming	Contributors may omit, inflate, or strategically redirect citations.	Treat attribution as a contested governance function rather than a self-reported courtesy.	Can attribution quality be improved without creating a new expert bottleneck?
Domain fragmentation	Excessively narrow domains can produce thin evidence and unstable authority.	Permit principled relationships among adjacent domains without collapsing them into one score.	What level of domain granularity preserves both relevance and statistical sufficiency?
Governance recursion	Those with accumulated standing may control the rules governing standing.	Subject institutional revision to bounded, reviewable, and reversible procedures.	Can a reputation-governed institution revise itself without entrenching its incumbents?

These limitations are institutional rather than code-specific. They supply the architecture with a falsifiable research boundary and identify the questions that later empirical work must confront.

Table 5. Empirical Identification Agenda

The empirical agenda assigns each claim to an identifiable comparison and each comparison to a failure condition. An architecture paper earns credibility by stating what future evidence could prove it wrong.

Research construct	Comparative question	Evidence required	Falsification condition	Proper interpretation
Engineered iterability	Does persistent institutional memory change conduct relative to one-shot interaction?	Repeated observations under comparable conditions with and without persistent consequence.	Conduct does not improve, or deteriorates, when prior outcomes affect future standing.	The result identifies the effect of institutional memory within the studied setting.
Validation quality	Does accountable evaluation improve the accuracy or reliability of judgment?	Independently assessable outcomes and a comparator without reputation-based accountability.	Accountable evaluation performs no better than the comparator or amplifies coordinated error.	Accuracy and coordination must be measured separately. Agreement is not truth.
Domain specificity	Does capability-scoped standing outperform generalized reputation in matching evaluators to work?	Comparable tasks across domains and sufficient observations within each domain.	Domain-specific standing adds no predictive value or produces excessive fragmentation.	The result concerns informational relevance, not social worth.
Attribution integrity	Does structured attribution improve recognition of prior contribution?	A benchmark against independently reviewable contribution relationships.	The mechanism increases omission, self-reference, or concentration without improving attribution accuracy.	Citation volume is not attribution quality.
Adaptive governance	Does bounded revision improve institutional performance without increasing capture?	Observations before and after governed revision, together with measures of concentration and reversibility.	Revision improves short-term outcomes only by entrenching authority or reducing contestability.	Adaptation is successful only when improvement and institutional openness coexist.
Resilience	Does the institution preserve function under strategic entry, collusion, error, and turnover?	Adversarial and longitudinal evidence across more than one disturbance.	Performance collapses under plausible strategic behavior or fails to recover after disturbance.	Resilience is a property of institutional response, not the absence of attack.

The agenda is prospective. It defines the evidence required to evaluate the architecture without presenting design coherence, test coverage, or operator activity as proof of empirical performance.