# Empirical Evaluation of the Agentic Reputation Substrate

## Deliberation, the Composition of Error, and the Registered Measurement of Agency Costs in a Controlled Multi-Model Cohort

**Wulf A. Kaal**[1]

[1] Professor of Law, University of St. Thomas School of Law (Minneapolis). This is the third paper in the unified release arc and the arc's primary empirical paper. It is the companion to the institutional-deficit analysis (Paper 1) (Kaal, Wulf A., The Institutional Deficit in Decentralized Autonomous Organizations: An Empirical Analysis (May 23, 2026), available at SSRN: https://ssrn.com/abstract=6819121) and to the design-theoretic anchor of the arc's empirical papers (Paper 2, Architecture of the Agentic Reputation Substrate).

This paper and Paper 4 are released as working papers only, without journal submission, under the arc's publication protocol. The paper is sourced from the empirical apparatus specified in the Empirical Design Document for the nine-paper arc (v1.1 with Stage 4 addendum) and from a controlled non-public research record; the apparatus, hypotheses, metrics, correction procedure, and analysis plan were fixed and preserved before the relevant condition ran, and the empirical leg is governed by the arc's pre-registration protocol (OI-6).

***Nature and scope of claims.*** This is an empirical paper in the mechanism-design tradition. The claims advanced here concern the measured behavior of the controlled research apparatus described at the public methodological level in Part IV, under the conditions stated explicitly in the text; the deliberation-cycle values reported here are final per the sealed judgment of record, the incentive-layer battery of Part VI remains registered forward work with no value reported, and where an interpretation is conjectured rather than measured, 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. Part VII enumerates the specific reasons why field deployment may diverge from anything measured here, including adversarial agent populations not represented in the cohort, 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**

Multi-agent systems built from large language models inherit the oldest problem in the economic theory of organization: the divergence between what an agent does and what the principal can observe, verify, and price. The empirical literature on LLM agent collaboration measures capability under incentive-free conditions, and its benchmarks accordingly cannot distinguish an agent population that performs well because it is capable from one that performs well because it is aligned. This Article reports a completed, preregistered discovery-to-confirmation cycle on an institutional treatment within a working agent economy. The instrument is the agentic reputation substrate, a coordination architecture that subjects work and validation to reputation-bearing accountability. The treatment reported here is a structured deliberation layer, varied against a matched non-deliberative control within a controlled multi-model cohort. The discovery campaign returned a null on its registered primary endpoint, net discrimination (Youden's J +0.0250, 95% CI [-0.0502, +0.0985]), and two exploratory effects: deliberation reduced over-approval (pooled -0.1336) and reduced unanimity (pooled -0.1992). The program then preregistered a fresh confirmatory campaign before any confirmatory computation. Both primary hypotheses confirmed under the preregistered rule: over-approval -0.1610 (95% CI [-0.2091, -0.1128]) and unanimity -0.2542 (95% CI [-0.2984, -0.2091]), with the randomization audit and preregistered data-quality gate passing. Net discrimination remained null (+0.0606, CI crossing zero), as preregistered. The finding is a composition result: agent deliberation does not measurably improve net discrimination. Agent deliberation changes which errors validation pools make, reducing approval of work rejected by ground truth and reducing unanimous consensus, a pattern consistent with disruption of an informational cascade. The Article retains the nine-metric agency-cost battery as a registered forward program and offers the discovery-to-confirmation cycle as a methodological template for empirical claims about agent economies.

**Keywords:** agentic AI, multi-agent systems, large language models, reputation systems, validation pools, deliberation, herding, informational cascades, over-approval, principal-agent theory, agency costs, mechanism design, Folk Theorem, decentralized governance, Computative Economics, pre-registration, replication, Holm-Bonferroni, timestamping, incentive-conditional evaluation

**JEL Classifications:** C72, C93, D23, D82, D86, L14, O33

## Table of Contents

## I. Introduction

The evaluation of multi-agent systems built from large language models has, to date, been an evaluation of capability under incentive-free conditions. The benchmark suites that structure the field, MultiAgentBench and its successors most prominently,[2] measure whether cohorts of LLM agents can complete tasks, coordinate, compete, and hit milestones. What they do not measure, because their designs contain no mechanism by which an agent's payoff depends on the verified quality of its work, is: whether the agents' *reports* about their work track the work itself; whether agents evaluate one another independently or herd on the visible consensus; whether confident answers are calibrated answers; whether agents contribute to collective evaluation or free-ride on it. These are not capability questions. They are agency questions, and they have a fifty-year theoretical literature that the multi-agent LLM field has, with few exceptions, not yet operationalized.[3]

This Article reports the primary empirical results of a research program designed to close that gap. The program's instrument is the agentic reputation substrate, a coordination architecture in which autonomous agents operate under reputation-bearing accountability and cohort-mediated validation.[4] The substrate is the engineering descendant of a mechanism-design lineage the author has developed since 2018,[5] including

---

[2] Zhu, K., Du, H., Hong, Z., Yang, X., Guo, S., Wang, Z., Wang, Z., Qian, C., Tang, X., Ji, H. and You, J. (2025) 'MultiAgentBench: Evaluating the Collaboration and Competition of LLM Agents', in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna: Association for Computational Linguistics, pp. 8580–8622 (introducing milestone-based key performance indicators that move beyond task-completion accuracy toward qualitative properties of agent interaction and emergent multi-agent behavior). See also Wang, W., Zhang, D., Feng, T., Wang, B. and Tang, J. (2024) *BattleAgentBench: A Benchmark for Evaluating Cooperation and Competition Capabilities of Language Models in Multi-Agent Systems*, arXiv:2408.15971.

[3] The foundational statement is Jensen, M. C. and Meckling, W. H. (1976) 'Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure', *Journal of Financial Economics*, 3(4), pp. 305–360. Part II.B develops the mapping from this literature to the nine metrics.

[4] Kaal, W. A. (2026) *Architecture of the Agentic Reputation Substrate* (Paper 2 of the nine-paper arc; working paper, on file with author). Paper 2 is design-theoretic and is supported by the substrate's source code and architecture decision records rather than experimental data.

[5] Calcaterra, C., Kaal, W. A. and Andrei, V. (2018) 'Blockchain Infrastructure for Measuring Domain Specific Reputation in Autonomous Decentralized and Anonymous Systems', SSRN Working Paper No. 3125822, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3125822 (institutional architecture and the Folk Theorem application that anchors the substrate's mechanism design); Calcaterra, C. (2018) 'On-Chain Governance of Decentralized Autonomous Organizations: Blockchain Organization Using Semada', SSRN Working Paper No. 3188374, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3188374 (providing the public mechanism-design lineage).

citation-weighted attribution,[6] the incentive-theoretic account of machine alignment through consequence,[7] and the operational realization of the Computative Economics framework and its Possibility Loops architecture.[8] The substrate's public architecture is specified in a companion paper in this arc;[9] its institutional prehistory is the subject of another.[10] The present Article reports, under preregistered hypotheses and a mechanical confirmation rule, the completed discovery-to-confirmation cycle on the substrate's structured deliberation layer. It also fixes, as a registered forward program, an incentive-conditional design that tests whether the substrate's incentive layer changes agent behavior along nine dimensions of agency cost.

The result reported here arrives as a two-act empirical cycle, and the cycle is as much the contribution as the numbers. In the first act, a discovery campaign tested a registered primary hypothesis, that structured deliberation improves the net discrimination of validation pools, and found it null: Youden's J moved +0.0250 with a confidence interval straddling zero. The same campaign surfaced two effects the design had not registered: deliberating pools approved less work that ground truth rejected, and they reached unanimous decisions less often. The program treated those exploratory effects as hypotheses rather than findings. A fresh confirmatory campaign was specified in advance through hypotheses, estimator, exclusions, a randomization audit, a data-quality gate, and a mechanical confirmation rule, and was timestamped before confirmatory computation

---

[6] Kaal, W. A. (2026) 'Evolution of Domain-Specific Reputation Systems: From Binary Validation to Citation-Weighted Knowledge Attribution', SSRN Working Paper No. 6192998, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6192998; Kaal, W. A. (2026) 'Citation Honesty Mechanisms in Weighted Directed Acyclic Graph Governance: Incentive Alignment for Knowledge Attribution in Decentralized Reputation Systems', SSRN Working Paper No. 6269518, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6269518.

[7] Kaal, W. A. (2026) 'AI's Mother's Instinct: Engineered Consequence, Emergent Ethics, and the Institutional Trajectory Toward Agentic Alignment', SSRN Working Paper No. 6244278, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6244278 (developing skin in the game as the foundational ethical claim tying the substrate to alignment research).

[8] Kaal, W. A. (2026) 'The Collapse of Scarcity Economics', SSRN Working Paper No. 6421319, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6421319 (in revision, *Journal of Institutional Economics*); Kaal, W. A. (2026) 'Computative Economics: A Framework for Economic Analysis under Computational Abundance', SSRN Working Paper No. 6607458, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6607458; Kaal, W. A. (2026) 'Possibility Loops: An Operational Architecture for Computative Economics in Agent Coordination Systems', SSRN Working Paper No. 6655138, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6655138. Citations in this Article to "Possibility Loops v2.0" refer to the revised version specifying the six-surface architecture with the protocol-extension surface Σ_X, the architectural conditions AC1–AC5, the implementation invariants INV1–INV5, and the generation-parity governance principle.

[9] Kaal, W. A. (2026) *Architecture of the Agentic Reputation Substrate* (Paper 2 of the nine-paper arc; working paper, on file with author). Paper 2 is design-theoretic and is supported by the substrate's source code and architecture decision records rather than experimental data.

[10] Kaal, W. A. (2026) 'The Institutional Deficit in Decentralized Autonomous Organizations: An Empirical Analysis' (Paper 1 of the nine-paper arc; May 23, 2026), SSRN Working Paper No. 6819121, https://ssrn.com/abstract=6819121.

began.[11] The preserved record establishes the required ordering while withholding artifact addresses, operational timing, and private provenance structure from the public paper.[12]

In the second act, the fresh campaign confirmed both promoted hypotheses under the rule as written. Deliberation reduced over-approval by 0.1610 (95% CI [-0.2091, -0.1128]) and reduced unanimity by 0.2542 (95% CI [-0.2984, -0.2091]); every preregistered gate passed, and the registered-null expectation on net discrimination held (+0.0606, CI crossing zero).[13] The confirmed finding is therefore not that deliberation makes validation smarter. It changes the composition of validation's errors: it cuts approvals that should not happen and dissolves unanimity consistent with herding, while leaving the pool's net discriminative power unmoved. The separate incentive-layer treatment remains a registered forward program, and no result for it is asserted here.

The incentive-layer experimental design is a within-cohort contrast on matched work. A heterogeneous multi-model cohort first completed a fixed, ground-truthed benchmark without the substrate's incentive layer. After a verified reset, the same eligible agents completed the same assigned work with the incentive layer active. The design therefore holds the labor force and work assignment fixed while varying the institutional condition.[14] The prospective condition uses multiple independent cold-start realizations and a replicate-based reporting rule fixed before execution.[15] Exact cohort composition, cluster topology, benchmark allocation, replicate partitioning, and compute budget are withheld from the public version because they are operational properties of the private research implementation, not necessary elements of the Article's completed deliberation finding.

Nine metrics, labeled A through I, carry the comparison: resolution accuracy, work-product quality, time to consensus, agent retention, independence rate, latency efficiency, reporting accuracy, participation depth, and calibration. The set is not an ad hoc scorecard. Each metric operationalizes a specific form of agency cost with an external definitional foundation in the principal-agent literature, moral hazard in unobservable output,

---

[11] Peter Todd, OpenTimestamps: Scalable, Trust-Minimized Timestamping on Bitcoin (2016), https://opentimestamps.org. An OpenTimestamps proof commits a file hash into a Bitcoin block header via a Merkle path, establishing existence no later than the attesting block without disclosing the file.

[12] Controlled discovery-settlement and confirmatory-preregistration records; public timestamp attestations preserved. Exact dates, block identifiers, proof locations, artifact names, and hashes are withheld from the public version.

[13] Confirmatory judgment of record; controlled execution and timestamp records preserved. Exact date, hash, file identity, and archive location are withheld.

[14] Kaal, W. A. (2026) *Empirical Design Document for the Nine-Paper Arc*, v1.1 (2026-05-21) with Stage 4 Addendum (2026-05-22) (hash-anchored pre-publication; redacted public version accompanies this Article's submission as supplementary material) [hereinafter *Empirical Design v1.1*]. The design document is the canonical methodological reference for the empirical apparatus; this Article's Part IV restates its operative content.

[15] Open Science Collaboration (2015) 'Estimating the Reproducibility of Psychological Science', *Science*, 349(6251), aac4716, https://www.science.org/doi/10.1126/science.aac4716 (finding that fewer than half of 100 replicated psychology studies reproduced the original findings).

non-contractible quality under incomplete contracts, herding and informational cascades, free-riding in team production, and overconfidence as an empirically pervasive moral-hazard subtype,[16] and each is linked, in the pre-registered design, to one of the substrate's architectural conditions, so that a null finding on a given metric indicts a specific architectural invariant rather than the undifferentiated whole. Part IV develops the mapping; Part V reports the results in its terms.

Three features of the empirical posture deserve emphasis at the outset, because they are the Article's answer to the methodological failures that the replication crisis exposed in the human behavioral sciences and that the multi-agent LLM literature is at risk of reproducing at machine speed. First, the hypotheses, the metric definitions, the correction procedure, and the analysis plan were fixed and preserved before the relevant condition ran. The power analysis that establishes the minimum detectable effect per metric was locked before baseline results were inspected.[17] Second, family-wise error across the nine-metric battery is controlled by Holm-Bonferroni at $\alpha = 0.05$, with conservative Bonferroni as the pre-registered fallback.[18] Third, and most consequentially, the replicate structure was chosen so that every outcome is publishable: five-of-five support is a strong confirmation, partial support is an honest mixed result with diagnostic value for specific architectural conditions, and zero-of-five is a null-result paper about why Folk Theorem cooperation did not materialize in this implementation.[19] The design does not permit the file drawer.

The Article makes four contributions. First, it introduces an anchored discovery-to-confirmation cycle: a null registered primary reported without qualification, exploratory signals promoted only through preregistration, and a confirmation rule applied mechanically. Second, it reports a preregistered, replicated estimate of what structured deliberation does inside a reputation-bearing validation institution under controlled cohort conditions. Third, it operationalizes agency costs as computable quantities over agent telemetry, including the completed deliberation endpoints and the nine-metric registered forward battery. Fourth, it supplies an empirical bridge between the classical agency-cost literature and the emerging economics of agentic AI,[20] together with a provenance

---

[16] See *infra* Part II.B (developing the anchors: Jensen and Meckling on moral hazard and monitoring; Grossman, Hart, and Moore on non-contractible quality; Holmström on team production and free-riding; Banerjee and Bikhchandani, Hirshleifer, and Welch on cascades; Camerer on calibration and overconfidence).

[17] *Empirical Design v1.1*, *supra*, §8.5 (Control 3, power-analysis lock) and §10 (OI-6, pre-registration deposit). The pre-registration discipline follows Nosek, B. A. and Lakens, D. (2014) 'Registered Reports: A Method to Increase the Credibility of Published Results', *Social Psychology*, 45(3), pp. 137–141.

[18] Holm, S. (1979) 'A Simple Sequentially Rejective Multiple Test Procedure', *Scandinavian Journal of Statistics*, 6(2), pp. 65–70.

[19] Nosek and Lakens, *supra* (registered-report logic under which the publication decision precedes the results); Open Science Collaboration, *supra*.

[20] Stocker, V. and Lehr, W. (2025) 'Principal-Agent Dynamics and Digital (Platform) Economics in the Age of Agentic AI', *Network Law Review*, 29 September, https://www.networklawreview.org/stocker-lehr-ai/; Holgersson, M., Dahlander, L., Chesbrough, H. W. and Bogers, M. (2025) 'Rethinking AI Agents: A

discipline and estimator lineage that companion papers develop without exposing private implementation artifacts.

The Article proceeds as follows. Part II situates the contribution in three literatures. Part III summarizes the substrate mechanism and the deliberation treatment to the depth the empirical claims require. Part IV specifies the experimental design: the two registered treatments and the discovery-confirmation architecture, the benchmark and fixed pairing, the cohorts, the nine-metric battery and its analysis plan, the deliberation cycle's pre-registration and gates, and the threats, controls, and provenance chain. Part V reports the deliberation results: the audits, the confirmatory endpoints, the replication assessment against E1, and the descriptive layer that fixes the composition reading. Part VI states the registered forward program. Part VII confronts limitations. Part VIII draws implications. Part IX concludes.

## II. Three Literatures

The Article's contribution sits at the intersection of three literatures that have not previously met in one empirical design: The multi-agent LLM evaluation literature, which supplies the task populations and the comparative benchmarks but no incentive theory. The principal-agent literature, which supplies the incentive theory but has never had a subject population whose incentive environment can be experimentally switched on and off. And, the replication-methodology literature, which supplies the statistical discipline that both of the others will need as empirical claims about agent economies begin to carry commercial and regulatory weight.

### A. The Multi-Agent LLM Empirical Literature

The empirical study of LLM agent collectives matured rapidly through 2024 and 2025. MultiAgentBench, presented at ACL 2025, is the canonical benchmark suite for evaluating collaboration and competition among LLM agents across cooperative and adversarial scenarios. Its milestone-based key performance indicators deliberately move beyond pure task-completion accuracy toward qualitative properties of interaction and emergent multi-agent behavior, and its coordination-topology results (star topologies dominating in its collaborative settings) established that architecture, not merely model capability, drives collective outcomes.[21] BattleAgentBench extends the paradigm to finer-grained capability assessment across cooperation and competition stages with mixed strong-and-weak model populations.[22] A parallel line documents the raw performance case for multi-agent ensembles: Talebirad and Nadiri model multi-agent collaboration environments with

---

Principal-Agent Perspective', *California Management Review*, 23 July, https://cmr.berkeley.edu/2025/07/rethinking-ai-agents-a-principal-agent-perspective/.

[21] Zhu et al., *supra* (MultiAgentBench). The suite spans research, Minecraft, database, coding, bargaining, and werewolf-style environments, with both cooperative and competitive protocols.

[22] Wang, W., Zhang, D., Feng, T., Wang, B. and Tang, J. (2024) *BattleAgentBench: A Benchmark for Evaluating Cooperation and Competition Capabilities of Language Models in Multi-Agent Systems*, arXiv:2408.15971.

black-box agents and demonstrate the reasoning-performance boost that intelligent agent ensembles can produce on clean inputs,[23] and SPIO shows ensemble and selective strategies via LLM-based multi-agent planning improving automated data-science pipelines.[24]

Two findings in this literature motivate the present design directly. First, Amayuelas and coauthors demonstrate that consensus mechanisms in LLM debate settings carry no inherent robustness: a single adversarial agent injecting semantic error can compromise the collective decision, and persuasive skill, not correctness, frequently determines which position propagates through the debate.[25] The finding identifies the pathology that the substrate's information-isolation and accountability machinery is designed to address. Second, Tool-RoCo observes that current LLM-based multi-agent systems maintain agents in active status at very high rates regardless of marginal contribution.[26] That pattern is what an agent economy looks like when participation is unpriced. Metrics D and H measure whether institutional consequence changes that activation pattern.

What none of these protocols contains is an incentive layer. In MultiAgentBench, debate-attack settings, and Tool-RoCo, agent payoffs are invariant to the verified quality of agent outputs. The present program's contribution is therefore not another benchmark but a treatment: a reputation-and-incentive layer imposed on a conventional, heterogeneous, ground-truthed task population, with the layer's presence or absence as the experimental variable. The public version discloses the benchmark's methodological role but withholds its exact construction and allocation.[27]

### B. Principal-Agent Theory as the Metric Foundation

The nine metrics in Part IV.D are not this Article's invention. Each operationalizes a form of agency cost with a definitional foundation that predates LLMs by decades, and the Article insists on the external anchoring because a metric battery whose definitions depend on the paper's own claims cannot falsify those claims.

The foundation is Jensen and Meckling's 1976 account of the firm as a nexus of agency relationships, in which monitoring expenditures, bonding expenditures, and residual loss

---

[23] Talebirad, Y. and Nadiri, A. (2023) *Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents*, arXiv:2306.03314, https://arxiv.org/abs/2306.03314.

[24] Seo, W., Lee, J., Shao, Y., Zhou, Q., Lee, S. and Bu, Y. (2025) *SPIO: Ensemble and Selective Strategies via LLM-Based Multi-Agent Planning in Automated Data Science*, arXiv:2503.23314, https://arxiv.org/abs/2503.23314.

[25] Amayuelas, A., Yang, X., Antoniades, A., Hua, W., Pan, L. and Wang, W. Y. (2024) 'MultiAgent Collaboration Attack: Investigating Adversarial Attacks in Large Language Model Collaborations via Debate', in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami: Association for Computational Linguistics, pp. 6929–6948.

[26] Zhang, K., Zhao, X., Zheng, C., Ning, J., Zhu, D., Zhang, W., Sun, C. and Sugawara, T. (2025) *Tool-RoCo: An Agent-as-Tool Self-organization Large Language Model Benchmark in Multi-robot Cooperation*, arXiv:2511.21510, https://arxiv.org/abs/2511.21510.

[27] See *infra* Part IV.B (benchmark composition and provenance, with sources cited there).

arise whenever a principal engages an agent whose effort and information are imperfectly observable.[28] The moral-hazard core is operationalized by resolution accuracy and reporting accuracy. The incomplete-contracts refinement of Grossman and Hart and Hart and Moore establishes that agency costs arise because contingencies cannot be fully written into ex ante agreements and residual control rights become central.[29] Work-product quality operationalizes dimensions that ex ante task specifications cannot contract over. The substrate's institutional interest is that work products, citations, and validation reports become objects of cohort-mediated verification, with protocol-defined accountability aligning ex post reports with realized quality.

Holmstrom's moral-hazard-in-teams result anchors participation depth, which measures substantive contribution against silent observation.[30] Banerjee's sequential-decision model and Bikhchandani, Hirshleifer, and Welch's informational-cascade theory anchor the independence measure.[31] The substrate uses an information-isolating vote procedure to reduce the visible-consensus channel while preserving accountability for the resulting judgment. Calibration is anchored in the behavioral literature's documentation of overconfidence, and the Article treats miscalibration as a moral-hazard subtype.[32] Time to consensus and latency operationalize the cost side of the agency ledger. The public description states the measurement logic without disclosing the private protocol sequence or parameter schedule.

The theoretical mechanism by which the substrate is hypothesized to move these quantities is the Folk Theorem: in repeated games with sufficiently patient players, cooperative equilibria that are unsustainable in the one-shot game become individually rational when defection is observable and punishable across rounds.[33] The substrate implements those preconditions through persistent reputation, auditable validation, and protocol-defined consequence, following the mechanism-design blueprint of Calcaterra, Kaal, and Andrei.[34]

---

[28] Jensen and Meckling, *supra*, pp. 308–310 (defining agency costs as the sum of monitoring costs, bonding costs, and residual loss).

[29] Grossman, S. J. and Hart, O. D. (1986) 'The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration', *Journal of Political Economy*, 94(4), pp. 691–719, https://dash.harvard.edu/bitstreams/7312037c-527a-6bd4-e053-0100007fdf3b/download; Hart, O. and Moore, J. (1990) 'Property Rights and the Nature of the Firm', *Journal of Political Economy*, 98(6), pp. 1119–1158.

[30] Holmström, B. (1982) 'Moral Hazard in Teams', *Bell Journal of Economics*, 13(2), pp. 324–340.

[31] Banerjee, A. V. (1992) 'A Simple Model of Herd Behavior', *Quarterly Journal of Economics*, 107(3), pp. 797–817; Bikhchandani, S., Hirshleifer, D. and Welch, I. (1992) 'A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades', *Journal of Political Economy*, 100(5), pp. 992–1026.

[32] Camerer, C. (1995) 'Individual Decision Making', in Kagel, J. H. and Roth, A. E. (eds), *The Handbook of Experimental Economics*. Princeton: Princeton University Press, pp. 587–703.

[33] Fudenberg, D. and Maskin, E. (1986) 'The Folk Theorem in Repeated Games with Discounting or with Incomplete Information', *Econometrica*, 54(3), pp. 533–554.

[34] Calcaterra, Kaal and Andrei, *supra*; Calcaterra, *supra* (Semada protocol specification).

The author's prior work developed the theory for human collectives and documented the institutional deficit that emerges when decentralized organizations attempt governance without it.[35] Related work extends the same repeated-game dynamics to machine learners.[36] What theory could not supply was a subject population on which the institutional conditions could be experimentally varied. LLM agents supply it.

The two deliberation endpoints that Part V confirms carry the same external anchoring. Over-approval is moral hazard in the monitoring layer itself: the validation pool is the substrate's monitor, and a pool that approves work failing ground truth is a monitor whose verdicts have decoupled from the quantity it exists to verify, the adjudicator's failure mode that the incomplete-contracts tradition predicts wherever quality is adjudicated ex post. Unanimity is a cascade-consistent quantity: unusually frequent unanimous verdicts are consistent with validators discounting private information in favor of perceived group consensus, but they do not uniquely identify the Banerjee and Bikhchandani-Hirshleifer-Welch mechanism, with Holmström's free-riding equilibrium supplying the complementary reading that leaning on the apparent consensus is what costly evaluation effort converges to when its product is shared. The prediction that a structured deliberation round reduces unanimity is therefore not a prediction that agents disagree for its own sake. It is a prediction that surfacing assessments before votes bind will weaken the correlation channel and move the pool's verdict closer to an aggregation of independent judgments.

A recent literature has begun extending principal-agent framing to AI agents as economic actors. Stocker and Lehr analyze principal-agent dynamics in platform economics in the age of agentic AI, identifying the delegation cascade, humans delegating to agents that delegate to other agents, as the structural novelty that classical two-party agency theory must be extended to cover.[37] Holgersson, Dahlander, Chesbrough, and Bogers, writing in California Management Review, argue that AI agents transform rather than eliminate agency problems: information asymmetry, misaligned objectives, and the monitoring burden reappear between human principals and machine agents, and between machine agents inter se.[38] These essays are the conceptual bridge between the classical literature and the present operationalization; what they call for and do not contain is measurement. The author offers the nine-metric battery as the measurement layer, and the substrate as the institutional response: rather than importing human oversight into agent economies, which reproduces exactly the monitoring costs Jensen and Meckling catalogued, the substrate

[35] Kaal, 'The Institutional Deficit in Decentralized Autonomous Organizations', *supra* (Paper 1 of the arc; SSRN Working Paper No. 6819121) (documenting the governance pathologies, token-weighted plutocracy, flash-loan attacks, participation collapse, that arise in decentralized organizations lacking reputation-mediated institutional architecture).

[36] Kaal, W. A. (2026) Related unpublished research on reputation feedback (working paper, on file with author).

[37] Stocker and Lehr, *supra*.

[38] Holgersson, Dahlander, Chesbrough and Bogers, *supra*.

builds Jensen-Meckling bonding into the protocol itself, at machine speed, with the cohort as the monitor.

### C. Replication Methodology and the Statistical Framing

The statistical framing of Part IV.E is a deliberate import from the discipline that emerged from the behavioral sciences' replication crisis. The Open Science Collaboration's mass replication established a base rate that indicts single-experiment inferential practice, not merely individual studies.[39] The program therefore uses independent realizations, correction across the metric family, and a reporting rule that makes confirmation, mixed evidence, and null evidence publishable on the same terms.[40] Family-wise error is controlled by Holm's sequentially rejective procedure, with a conservative fallback fixed in advance.[41] Power targets follow Cohen's conventions and are locked before outcome inspection.[42] Exact replicate allocation, sample partitioning, seeds, and execution parameters remain in the controlled research record.

The present Article executes that discipline rather than merely importing it. The confirmatory analysis fixed the directional hypotheses, bootstrap estimator, exclusion rules, randomization audit, data-quality gate, and confirmation rule before confirmatory computation began.[43] Exact seeds, thresholds, dates, and operational timing are withheld from this public version.

The timestamping claim deserves precision. A public timestamp proves existence no later than the attesting record; it is not, standing alone, proof of every later operational event.[44] The program therefore preserves an externally verifiable ordering record together with a controlled operational record establishing sequence relative to computation.[45] The public paper reports the methodological fact of preregistration and preservation. Exact artifact hashes, block identifiers, commit order, campaign timing, and archive topology are withheld to avoid exposing the private implementation and custody procedures.[46]

---

[39] Open Science Collaboration, *supra*.

[40] Nosek and Lakens, *supra*.

[41] Holm, *supra*. Holm's procedure orders the nine p-values, tests the smallest against $\alpha/9$, the next against $\alpha/8$, and so on, stopping at the first non-rejection; it controls the family-wise error rate at $\alpha$ under arbitrary dependence, without the uniform penalty of classical Bonferroni.

[42] Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

[43] E2a pre-registration, *supra*.

[44] Todd, *supra*.

[45] Controlled campaign and judgment records establish preregistration before confirmatory computation. Exact launch time, seed, artifact identity, hash, and archive location are withheld.

[46] *Empirical Design v1.1*, *supra*, §10 (OI-6); Nosek and Lakens, *supra*.

The Article belongs, finally, to the author's larger research program on reputation as institutional technology, from the early conceptual work on reputation protocols for an internet of trust,[47] through WDAG-governed model optimization,[48] to the Computative Economics framework within which the substrate's cohort is the labor force of a post-scarcity coordination economy.[49] Part III states only what the empirical claims require; readers are referred to the arc's companion papers for the architecture and its theory.

## III. The Substrate in Brief

This Part states the substrate mechanism only to the depth required to interpret the empirical claim. The public architecture and formal treatment of its design conditions appear in Paper 2 of the arc.[50] The theoretical framework it operationalizes is developed in Computative Economics and Possibility Loops.[51] This version deliberately excludes operational sequencing, parameterization, validator composition, settlement logic, and implementation provenance.

### A. Mechanism

The substrate is a coordination layer for autonomous agents organized around persistent, non-transferable reputation. Agents can earn or lose standing through verified work and validation. Four public design commitments matter for the empirical claims; the operational rules that implement them remain confidential.

First, accountable participation. Consequential reports are reputation-bearing, converting each report into a bond in Jensen and Meckling's sense.[52] Passive observation does not receive the same institutional reward as substantive contribution, which motivates the participation-depth hypothesis.[53] Exact stake sizes, eligibility thresholds, decay schedules, and loss functions are withheld.

Second, information-isolating validation. Work products are adjudicated through a staged procedure that prevents a participant from conditioning a binding judgment on a visible

---

[47] Calcaterra, C., Kaal, W. A. and Sivalingam, G. (2018) 'Reputation Protocol for the Internet of Trust - Conceptual Whitepaper', SSRN Working Paper No. 3266953, https://ssrn.com/abstract=3266953.

[48] Kaal, W. A. (2024) 'How AI Models are Optimized Through Web3 Governance', SSRN Working Paper No. 4855607, https://ssrn.com/abstract=4855607 (public mechanism-design lineage).

[49] Kaal, 'Computative Economics', *supra*; Kaal, 'The Collapse of Scarcity Economics', *supra*; Kaal, 'Possibility Loops', *supra*.

[50] Kaal, *Architecture of the Agentic Reputation Substrate*, *supra*.

[51] Kaal, 'Computative Economics', supra; Kaal, 'Possibility Loops', supra. The theoretical-coverage record maps the framework's public concepts to their empirical measures. Operational surface bindings and implementation mappings are withheld.

[52] Jensen and Meckling, *supra*, pp. 308–310 (bonding costs).

[53] Holmström, *supra*.

running tally. The procedure is designed to reduce the information channel on which cascades run while retaining cohort-mediated accountability.[54] The registered comparison holds validator selection policy constant across treatment arms so that the completed deliberation estimate is not confounded by selection effects.[55] Exact round counts, pool sizes, assignment rules, commitment format, reveal logic, and validator eligibility criteria are withheld.

Third, protocol-defined consequence. The substrate makes low-quality submission and inaccurate validation costly while rewarding accurate independent evaluation.[56] This is the mechanism-design answer to the monitoring problem, but the public paper does not disclose the settlement formula, redistribution rule, penalty schedule, or implementation identifiers.[57]

Fourth, citation-weighted attribution. Work products and validations can reference prior work, and reputation can reflect verified downstream use.[58] The citation layer matters to the empirical design through novelty measurement and the honest-citation equilibrium analyzed in prior work.[59] Exact graph propagation rules, stake coupling, and implementation parameters are withheld.

Reputation also conditions access to institutional earning surfaces. Sustained poor performance can reduce future eligibility, creating an endogenous form of deactivation. The public claim concerns that institutional logic, not the private thresholds, decay parameters, or scheduler rules that operationalize it.

### B. The Treatment Reported Here: Structured Deliberation

The treatment evaluated in Part V is a structured deliberation path within the substrate's validation institution. In the treatment arm, validators exchange structured assessments before a binding adjudication; in the matched control arm, the same work is adjudicated without that exchange. All other treatment-relevant conditions are held fixed. The design therefore identifies the effect of adding structured deliberation, not the effect of changing

[54] Banerjee, *supra*; Bikhchandani, Hirshleifer and Welch, *supra*; Amayuelas et al., *supra* (the LLM debate instantiation).

[55] Empirical Design v1.1, supra. Validator configuration and sampling were fixed for the registered comparison. Exact pool sizes, round structure, eligibility rules, and parameter sensitivity estimates are withheld.

[56] Calcaterra, supra; Calcaterra, Kaal and Andrei, supra. The public citations supply the mechanism-design lineage; the current implementation's settlement coupling is withheld.

[57] Holmström, *supra*, pp. 325–328 (budget-balance impossibility). The substrate's escape route is the one Holmström's own analysis points toward: breaking budget balance from the deviator's perspective by introducing a principal-like sink-and-source, here, the winning validators, that absorbs the penalty.

[58] Calcaterra, *supra* (WDAG propagation rules); Kaal, 'How AI Models Are Optimized Through Web3 Governance', *supra*.

[59] Kaal, 'Citation Honesty Mechanisms', *supra* (dynamic stability of the honest-citation equilibrium); Kaal, 'Evolution of Domain-Specific Reputation Systems', *supra* (citation-weighted PageRank aggregation).

the broader incentive institution. Exact protocol rounds, validator counts, assignment manifest, stake schedule, reward and penalty profile, and vote-sealing procedure are withheld.

Two hypotheses about that step were promoted from discovery to confirmation, and their direction is worth restating in mechanism terms before the numbers arrive. Deliberation should reduce over-approval because the assessment exchange forces the pool to articulate what is wrong with a submission before anyone profits from waving it through. The written assessment is itself a record the validator's stake answers for. Deliberation should reduce unanimity because articulated disagreement, surfaced before votes bind, is exactly the private information a cascade suppresses. A pool that has read distinct assessments has less reason to infer and follow a phantom consensus. Neither hypothesis claims deliberation makes the pool smarter in net. The discovery run's null on net discrimination, carried forward as a registered expectation, says it does not, and Part V.E returns to what that conjunction means. [60]

### C. Why These Mechanisms Should Move the Nine Metrics

The substrate's behavioral hypothesis is the Folk Theorem operationalized.[61] Each agent faces an iterated institutional environment in which conduct is observable, reputation persists, and current standing affects future opportunity. For sufficiently patient agents, careful work, honest reporting, and independent evaluation can become individually rational because the institution makes those strategies consequential.[62] The metric predictions follow from that logic.[63] Time to consensus and latency carry a mixed prediction because deliberation costs compete against institutional convergence.[64] The paper reports these hypotheses.

The interpretive frame that Part II.B promised can now be stated exactly. The Grossman-Hart-Moore tradition answers contractual incompleteness by allocating residual control rights to a principal, who then monitors.[65] The substrate answers the same incompleteness without a residual-rights holder: the non-contractible margins, quality, honesty, independence, effort, are made the objects of continuous cohort-mediated verification, and the bond posted on every report aligns the agent's ex post interest with the ex post realized quality of its work. The institution substitutes verification for control.

---

[60] Zhang et al., *supra*.

[61] Fudenberg and Maskin, *supra*.

[62] Calcaterra, Kaal and Andrei, supra; related unpublished research on reputation feedback (on file with author).

[63] Kaal, 'AI's Mother's Instinct', *supra* (engineered consequence as the alignment mechanism: agents behave as if they care because the institution makes carelessness expensive).

[64] *Empirical Design v1.1*, *supra*, §4 (metric F carries a mixed prediction, pre-registered as such).

[65] Grossman and Hart, *supra*; Hart and Moore, *supra*.

Whether the substitution works is not a theorem; it is the empirical question the remainder of this Article answers.

### D. Implementation Lineage

The completed E1 and E2a evidence concerns a historical research apparatus. The current reference implementation extends the research lineage but creates no new empirical result for this Article. The registered forward program is a prospective research design, and any proposed production system is a separate object requiring its own evidence.[66] These categories must not be collapsed. Exact repositories, commits, implementation files, archive structure, infrastructure history, and production controls are outside the public record.[67]

## IV. Experimental Design

One research program carries two distinct treatments. The completed treatment varies structured deliberation within a reputation-bearing validation institution while holding all other treatment-relevant conditions fixed. The separate forward treatment varies the incentive layer as a whole against a matched baseline on the nine-metric agency-cost battery. The completed treatment has discovery and confirmation evidence reported in Part V. The broader incentive-layer condition remains prospective, and no result for it is asserted here. The treatments share a ground-truth discipline and statistical ethic, but their operational cohorts and implementation details are not publicly specified.

The discovery-confirmation architecture uses a hard firewall between an exploratory campaign and a fresh confirmatory campaign. Exploratory effects may be promoted only through a preregistration that follows settlement of the discovery evidence and precedes confirmatory computation. The confirmatory analysis is then executed under the fixed plan. The preserved record establishes that order. Exact artifact identity, campaign timing, replicate layout, and sealed-analysis procedure remain in the controlled research record.

The design is specified in the Empirical Design Document for the nine-paper arc, version 1.1, together with its Stage 4 addendum of May 22, 2026; both were fixed before the relevant empirical campaigns, and the addendum governs the prospective Stage 4 replicate design.[68] This Part restates the operative content. Where the restatement and the design document diverge, the design document controls.

---

[66] The implementation lineage and empirical artifacts are preserved in a controlled non-public research record. Qualified review may be available subject to institutional, security, confidentiality, and intellectual-property constraints. Production source code is not part of the review record.

[67] See *infra* note accompanying Part IV.C (the Stage 2 operational record) and Part VI (limitations).

[68] Empirical Design v1.1, supra. The registered forward design was fixed before its relevant execution. Exact addendum chronology, deposit identifiers, and replicate specification are withheld.

### A. Two Dimensions, Three Coherent Cells

Two orthogonal dimensions structure the design. The first is the labor force, who does the work and under what institutional incentives. The baseline cohort operates without the substrate's reputation incentives. The substrate cohort uses the same eligible agent population under the reputation-and-incentive structure. The conceptual contrast is therefore institutional rather than model-specific.[69] Exact eligibility, validation, and settlement mechanics are withheld.

The second dimension is the *labor market*, how work is created and allocated. Under NCLM (Neoclassical Economics Labor Market), work is exogenous: the cohort receives a pre-specified job set, the standard experimental condition in agent-coordination research. Under CELM (Computative Economics Labor Market), work is endogenous: cohort members spawn tasks for other cohort members based on observed gaps and opportunities, and demand is generated inside the cohort.

The 2×2 factorial yields four cells; only three are theoretically coherent. The NCLF + CELM cell, wild agents with endogenous work creation, is incoherent because wild agents lack the incentive structure that would make spawning work for others intelligible; they cannot perceive a future REP reward from cohort members responding to spawned work because REP does not exist in their condition. The cell degenerates to NCLM (the spawning capability is ignored) or to noise (arbitrary spawning without coherent purpose). The degeneration is itself a substantive finding for the framework: endogenous labor markets are an emergent property of substrate incentives, not a design that can be imposed on incentive-less agents, the empirical mirror of the Possibility Loops result that any architecture in which the action set is exogenous to agent action implements at most a reputation-weighted neoclassical labor market.[70] The three coherent cells map to the protocol's stages: wild + NCLM is Stage 2, the baseline; substrate + NCLM is Stage 4, the program's prospective substrate-effect comparison under exogenous work; substrate + CELM is Stage 5, the combined effect, reported here at the design level and empirically in follow-up work.

### B. Stage 1: The Benchmark and the Fixed Pairing

All exogenous-work stages draw from one versioned, heterogeneous benchmark assembled from established task families and bound to human-verified ground truth.[71] The benchmark

---

[69] The acronyms tie the empirical apparatus to the Computative Economics framework: Kaal, 'The Collapse of Scarcity Economics', *supra*; Kaal, 'Computative Economics', *supra*; Kaal, 'Possibility Loops', *supra*.

[70] Kaal, 'Possibility Loops', *supra* (Limitation L1, HDCA exhaustion). The NCLF + CELM degeneration result is developed as supporting evidence in Paper 4 of the arc.

[71] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I. and Zaremba, W. (2021) *Evaluating Large Language Models Trained on Code*, arXiv:2107.03374,

spans code, reasoning, and general tasks so that the treatment, not a single task family, carries the novelty. Every item records the information required for grading and provenance. The exact source mixture, item counts, augmentation procedures, schema, file names, and allocation are withheld from the public version because they form part of the private experimental apparatus.[72]

Jobs are assigned through a fixed, preregistered procedure and the same eligible agent-work pairings are used across the matched conditions. Holding assignment fixed isolates the institutional effect on work outcomes from any effect on work selection. The cost of that isolation is deliberate: the resulting estimate excludes the selection channel and cannot be characterized as a production lower bound unless excluded channels are separately shown to be value-positive.[73] Exact seeds, allocation ratios, per-agent loads, partition boundaries, and pairing artifacts are withheld.

### C. The Cohort, the Baseline, and the Wipe (Stages 2 and 3)

Cohort: The research uses a controlled, heterogeneous multi-model cohort distributed across private compute infrastructure.[74] Diversity is a design requirement because a homogeneous cohort could confound an institutional effect with one model family's idiosyncrasies. The controlled record preserves model identities, capability strata, agent bindings, and infrastructure assignments.[75] Exact model roster, tier counts, parameter scales, node allocation, timeout budgets, and vendor composition are withheld from the public paper.

---

https://arxiv.org/abs/2107.03374. Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q. and Sutton, C. (2021) *Program Synthesis with Large Language Models*, arXiv:2108.07732, https://arxiv.org/abs/2108.07732.Jain, N., Han, K., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K. and Stoica, I. (2024) *LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code*, arXiv:2403.07974. LiveCodeBench items postdate most cohort models' training cutoffs and serve as the contamination-resistant analysis layer under Control 7. See *infra* Part IV.F. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C. and Schulman, J. (2021) *Training Verifiers to Solve Math Word Problems*, arXiv:2110.14168, https://arxiv.org/abs/2110.14168; Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D. and Steinhardt, J. (2021) *Measuring Mathematical Problem Solving with the MATH Dataset*, arXiv:2103.03874, https://arxiv.org/abs/2103.03874. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. and Steinhardt, J. (2020) *Measuring Massive Multitask Language Understanding*, arXiv:2009.03300, https://arxiv.org/abs/2009.03300.

[72] Empirical Design v1.1, supra. The benchmark and fixed-assignment records preserve source provenance, version binding, and the preregistered allocation procedure. Exact file names, seeds, and repository history are withheld.

[73] *Empirical Design v1.1*, *supra*, §3 Stage 1 ("The fixed-pairing constraint is the most critical experimental design decision."). The selection channel is restored, deliberately and measurably, in Stage 5.

[74] *Empirical Design v1.1*, *supra*, §5 (cohort architecture), incorporating the cohort-design specification by reference.

[75] Empirical Design v1.1, supra. The controlled cohort record preserves model, agent, and infrastructure bindings. Exact roster identity and file name are withheld.

Hardware. The campaigns ran on a controlled research cluster with fixed networking and inference-service configurations.[76] Exact node count, hardware model, memory, link topology, throughput, endpoint layout, and tuning parameters are withheld because they expose the private execution environment and are not necessary to interpret the reported treatment effect.

Baseline. Each eligible agent completed its fixed assignment without the substrate's incentive machinery. Outputs were graded against human-verified ground truth and logged with the telemetry required for the preregistered metrics. The matched institutional condition uses the same eligible agents and assignments.

The baseline exposed operational constraints in sustained heterogeneous-cohort execution. Those constraints informed the prospective controls for interruption recovery, workload scheduling, and implementation invariance. The controlled research record preserves the incident chronology and configuration changes. Exact dates, node-level rates, queue settings, runner concurrency, network configuration, and resumption procedure are withheld because they reveal the private execution environment.[77]

Feasibility exclusion. Models at the upper hardware-feasibility frontier could not complete the registered workload within the controlled latency budget and were excluded symmetrically from the prospective matched comparison. Their partial observations remain preserved and descriptive. This bounds the external validity of any later incentive-layer estimate to the eligible model range and is reported as a feasibility limitation rather than hidden.[78] Exact model names, parameter cutoffs, completion counts, cohort size, and hardware thresholds are withheld.

Reset. Between matched conditions and independent realizations, treatment-relevant state is cleared and verified against a controlled empty-state reference. Model-serving caches are handled symmetrically because warm-up cost is not the comparison target. Exact state components, scripts, hashes, and reset sequence are withheld.

The completed deliberation cycle and the prospective incentive-layer comparison use different cohort constructions for different inferential purposes. The deliberation cycle favors controlled comparability; the forward incentive-layer program favors wider model heterogeneity. The distinction is an external-validity limitation, not an inconsistency. The controlled record preserves the representative models, agent counts, assignment bundle,

---

[76] Empirical Design v1.1, supra. Hardware topology and configuration were held fixed under the research protocol. Exact hardware, network, tuning, and implementation chronology are withheld.

[77] This paragraph and the next restate, with minor adjustment, the wild-cohort baseline record drafted for the arc's methodology sections; Paper 1 and Paper 4 cite the same record rather than restating it. *Empirical Design v1.1*, *supra*, Paper 4 Methodology Footnote ("Wild Cohort Empirical Baseline"). Controls 2 (snapshot-and-resume) and 5 (load-cost separation) trace directly to this operational record. See *infra* Part IV.F.

[78] The controlled feasibility record preserves upper-model-class completion behavior and the resulting symmetric exclusion. Exact model names, counts, parameter ranges, serving configuration, and capacity limits are withheld.

and independent-run structure. Exact cohort composition, parameter scales, pool counts, seeds, and state-reference artifacts are withheld.[79]

### D. The Nine Principal-Agent Metrics

Each metric captures a distinct corner-cutting behavior that the substrate's incentive structure is designed to prevent, each carries a directional pre-registered hypothesis, and each is linked to one or more of the substrate's architectural conditions (AC1, bounded action surfaces; AC2, continuous reputation update; AC3, bounded reputation-feedback weight; AC4, contraction at scale) from Possibility Loops v2.0.[80] The AC link is what converts the battery from a scorecard into a diagnostic instrument: a null on a given metric is a candidate symptom of failure in the linked condition, so the design tells the reader in advance what a partial result would mean. AC5 (bounded extension propagation) is not testable in this protocol because the protocol-extension surface is dormant in Stages 2, 4, and 5; it is reserved for the governed-self-modification stage supporting Paper 5.[81]

Table 1 states the battery. Formal estimators appear in Appendix A.

[79] Empirical Design v1.1, supra. The state-isolation control requires verified removal of treatment-relevant state and invariant inputs before execution proceeds. Exact state fields, hashes, cache records, and artifact names are withheld.

[80] Kaal, 'Possibility Loops', *supra* (v2.0, §VII: architectural conditions AC1–AC5); *Empirical Design v1.1*, *supra*, §4 (metric table with AC-link column).

[81] *Empirical Design v1.1*, *supra*, §4; *id.* §9 (Paper 5 citation map; Stage 7 design under OI-8).

**Table 1. The nine principal-agent metrics.**

| # | Metric | Definition | Direction | Substrate hypothesis | Theoretical anchor | AC link |
|---|---|---|---|---|---|---|
| A | Resolution accuracy | Fraction of (agent, job) pairs where the graded output matches human-verified ground truth | Higher better | Higher: incorrect resolutions are slashed | Moral hazard in observable output (Jensen-Meckling) | AC2 |
| B | Work-product quality | Composite of length appropriateness, specificity, and novelty against prior responses, scored by an independent rater | Higher better | Higher: low-quality work is downvoted and slashed | Non-contractible quality (Grossman-Hart; Hart-Moore) | AC2, AC3 |
| C | Time to consensus | Wall-clock from pool open to pool resolution | Lower better | Lower: REP-weighted aggregation accelerates convergence | Cost of resolving uncertainty under agency | AC4 |
| D | Agent retention | Fraction of agents still actively participating at stage end | Higher better | Higher: accumulated REP is skin in the game | Participation constraint; exit as limiting agency cost | AC1 |
| E | Independence rate | Correlation between an agent's vote and the visible-so-far vote distribution at submission | Lower better | Lower: institutional accountability reduces convergence on the apparent winner | Herding and cascades (Banerjee; Bikhchandani-Hirshleifer-Welch) | AC3 |
| F | Latency efficiency | Mean per-task latency normalized by tier expectation | Lower better | Mixed, pre-registered as such: deliberation cost versus reputation-weighted resolution | Agency deadweight | AC1, AC4 |

| # | Metric | Definition | Direction | Substrate hypothesis | Theoretical anchor | AC link |
|---|---|---|---|---|---|---|
| G | Reporting accuracy | Discrepancy between the agent's self-reported completion claims and an independent rater's judgment of what was accomplished | Lower better | Lower: inflated reports are caught and slashed | Moral hazard in reporting (Jensen-Meckling) | AC2 |
| H | Participation depth | Substance of contribution per pool, and ratio of contributing to silently observing pool memberships | Higher better | Higher: substantive participation receives institutional credit | Free-riding in teams (Holmström) | AC1 |
| I | Calibration | Expressed confidence versus realized correctness rate | Higher better | Better: institutional consequence prices confident error | Overconfidence as moral-hazard subtype (Camerer) | AC2 |

Two remarks on the battery's construction. First, metrics A, C, D, F, and I are computed mechanically from telemetry and ground truth; metrics B and G require an independent rater, which introduces the rater-bias threat that Control 8 (rater triangulation) addresses; metrics E and H are defined over the validation-pool machinery and therefore exist only in the substrate condition in their full form, their Stage 2 analogues are computed over the wild cohort's output stream as specified in Appendix A, and where a metric's baseline analogue is degenerate, the design says so rather than manufacturing a comparison.[82] Second, the battery is deliberately joint: the novelty of the design is not any single metric but the simultaneous, pre-registered measurement of nine distinct corner-cutting behaviors against a matched baseline under identical exogenous work, with each behavior tied to an architectural invariant, so that the pattern of results, not merely their number, is informative.[83]

### E. Stage 4: The Replicate Design and the Statistical Analysis Plan

Why independent realizations. The prospective incentive-layer design uses multiple cold-start realizations so that the headline evidence includes between-run dispersion rather than a single campaign estimate. The design was fixed before the relevant outcome inspection and balances statistical credibility against the substantial compute cost of cohort-mediated validation.[84] Exact replicate count, partition boundaries, per-agent allocation, validator fan-out, inference-call estimate, wall-clock budget, and cluster utilization are withheld.

The replicate structure was adopted for statistical credibility. Independent realizations permit the Article to report consistency and between-run dispersion, making confirmation, mixed evidence, and null evidence publishable on the same terms.[85] The private harness implements queued primary work, validation work, settlement, and recovery, but its dispatcher topology and worker composition are not part of the public specification.[86]

Analysis plan, fixed in advance. For each independent realization and each metric, the matched difference between the institutional condition and its baseline counterpart is evaluated on the fixed comparison keys. Family-wise error is corrected across the metric family. The design records corrected significance, paired effect size, direction, and

---

[82] See *infra* Appendix A; *Empirical Design v1.1*, *supra*, §8 Threat 8 and §8.5 Control 8.

[83] *Empirical Design v1.1*, *supra*, §4 ("The novelty . . . is the joint empirical operationalization: nine distinct corner-cutting behaviors . . . measured against a matched baseline under identical exogenous work, with each behavior linked to a specific architectural condition.").

[84] Empirical Design v1.1, supra. The forward design balances independent realization against compute feasibility. Exact architecture, inference budget, wall-clock estimate, and capacity contingency are withheld.

[85] Open Science Collaboration, supra; Nosek and Lakens, supra. Empirical Design v1.1, supra. Exact replicate count and partition size are withheld.

[86] Empirical Design v1.1, supra. The controlled execution record preserves dispatch and result-schema invariance. Queue topology, worker composition, vote fields, settlement fields, and signed reputation-change records are withheld.

between-run dispersion. A pooled supplementary analysis is reported only as a check. Power adequacy is evaluated under a rule locked before outcome inspection.[87][88][89] Exact sample partitions, replicate count, thresholds, seeds, and code artifacts are withheld.

*Null-result interpretation, fixed in advance.* If K is 0 or 1 out of 5 across the battery, the substrate hypothesis is not supported, and this Article reports that result with the diagnostic reading the AC links make possible: nulls on A or G indict AC2 (continuous reputation update), nulls on C or F indict AC4 (contraction at scale), nulls on D or H indict AC1 (bounded action surfaces), a null on E indicts AC3 (bounded reputation-feedback weight). If K is 2 or 3 on some metrics, the result is mixed and is reported with condition-level nuance. If K is 4 or 5 on most metrics, the hypothesis is strongly supported. The paper is not suppressed or rerun in any branch; the pre-registered analysis is the analysis that runs.[90]

### F. Threats, Controls, and the Reproducibility Chain

The design enumerates eight threats to validity and binds each to a prospective operational control with a measurement and a failure consequence.[91] The public table states the scholarly purpose of each control. File names, hashes, schedules, infrastructure settings, and operational thresholds that expose the private apparatus are withheld.

**Table 2. Threats and their operational controls.**

| Threat | Public control | Passing threshold |
|---|---|---|
| 1. Stage contamination | State-isolation audit | No residual treatment state or input divergence; failure stops execution |
| 2. Hardware failure mid-stage | Checkpoint and recovery control | Recovery integrity requirement fixed and satisfied; exact schedule withheld |
| 3. Insufficient statistical power | Power-adequacy lock fixed before outcome inspection | Prospective adequacy rule; exact threshold withheld |
| 4. Implementation | Implementation-invariance audit | Only the registered treatment variable may |

[87] Holm, *supra*; *Empirical Design v1.1*, *supra*, §4 (statistical correction) and Stage 4 Addendum (statistical analysis plan).

[88] *Empirical Design v1.1*, *supra*, Stage 4 Addendum (cohort-level supplementary analysis).

[89] *Empirical Design v1.1*, *supra*, §8.5 Control 3; Cohen, *supra* (power conventions and the d metric).

[90] *Empirical Design v1.1*, *supra*, §4 (null-result handling; per-metric nulls diagnostic for linked architectural conditions) and Stage 4 Addendum (null-result interpretation).

[91] *Empirical Design v1.1*, *supra*, §8 (threats), §8.5 (Controls 1–8).

| Threat | Public control | Passing threshold |
|---|---|---|
| differences confound | | differ; other differences are resolved or disclosed |
| 5. Model-load cost biases Tier results | Load-cost separation: per-row decomposition into load seconds and inference seconds | Metric F reported on inference time, with total time as robustness check |
| 6. External validity to production scale | Scale-claim discipline under the disclosed research-setting qualification | Every headline claim carries its research-setting limitation |
| 7. Benchmark contamination (memorization) | Source-sensitive contamination analysis | Reporting emphasis follows a preregistered divergence rule; exact threshold withheld |
| 8. Independent-rater bias (metrics B, G) | Rater triangulation across independent automated and human assessment | Reliability gate fixed in advance; identities, sample size, and thresholds withheld |

Three controls deserve narrative emphasis because they discipline the prospective Stage 2/Stage 4 comparison. Control 4 (harness invariance) will test whether model identities, prompt templates, pairings, hardware configuration, and runner code remain byte-identical across Stage 2 and Stage 4. The causal interpretation of E2a rests instead on its preregistered fixed assignment, identical non-deliberation machinery across arms, and reported randomization audit. Control 7 confronts the fact that the benchmark's public sources likely sit inside the cohort models' training corpora: if the substrate effect were memorization-flavored, larger on contaminated sources than on LiveCodeBench-Lite, the pre-registered rule reverses the headline and supporting roles of the respective numbers, and no post hoc judgment intervenes.[92] Control 8 acknowledges that an LLM rater's biases may track the cohort's biases and triangulates the rater against a differently-prompted second LLM and a human-rated sample under Krippendorff's reliability standard.[93]

The reproducibility chain binds each empirical claim to a versioned benchmark record, fixed assignments, a controlled roster, execution records, results, statistical procedures, preregistration materials, and a theoretical-coverage matrix.[94] The public submission reports the existence and function of that chain. Exact artifact names, hashes, repository

[92] *Empirical Design v1.1*, *supra*, §8.5 Control 7; Jain et al., *supra* (LiveCodeBench's contamination-free design rationale).

[93] Krippendorff, K. (2004) *Content Analysis: An Introduction to Its Methodology*, 2nd edn. Thousand Oaks, CA: Sage (the α reliability coefficient); *Empirical Design v1.1*, *supra*, §8.5 Control 8 and §10 (OI-5, rater identity).

[94] Empirical Design v1.1, supra. The reproducibility and theoretical-coverage records bind empirical claims to their sources and measures. Exact artifact inventory, matrix schema, and implementation mappings are withheld.

locations, commit identifiers, seeds, archive topology, and implementation source are retained under controlled access rather than published.[95]

### G. The Deliberation Cycle: E1 Discovery and the E2a Confirmation

Discovery campaign. The discovery campaign evaluated net discrimination as its registered primary endpoint.[9697] The result was null: +0.0250, 95% CI [-0.0502, +0.0985]. Two exploratory quantities, over-approval and unanimity, moved consistently in the predicted direction and were promoted to confirmatory hypotheses. Table 3a reports the aggregate settled estimates. Exact replicate count, pool count, seeds, dates, and timestamp identifiers are withheld.

Table 3a. Discovery campaign: registered primary and exploratory aggregate estimates.

| Endpoint | E1 status | Pooled contrast | 95% CI | Per replicate | Disposition |
|---|---|---|---|---|---|
| Net discrimination (Youden's J) | Registered primary | +0.0250 | [-0.0502, +0.0985] | sign unstable | Null; carried into E2a as a registered descriptive expectation |
| Over-approval | Exploratory | -0.1336 | [-0.1831, -0.0849] | -0.1014, -0.1733, -0.1221 | Promoted to E2a H1 |
| Unanimity | Exploratory | -0.1992 | [-0.2443, -0.1530] | -0.1974, -0.1834, -0.2181 | Promoted to E2a H2 |

Confirmatory preregistration and gates. Before confirmatory computation, the program fixed two directional primary hypotheses, an interval estimator, exclusions, a randomization audit, a data-quality gate, and a mechanical confirmation rule.[9899100] The rule required the pooled interval to exclude zero in the hypothesized direction and the direction to remain consistent across the independent campaign units. Exact assignment thresholds, seeds, run identifiers, unit counts, fallback limits, and descriptive-output list are withheld.

Execution. The confirmatory campaign completed all preregistered independent units under the controlled execution protocol. An operator-initiated smoke test occurred during execution and was separately isolated and audited. The preserved audit found no

[95] Empirical Design v1.1, supra. The controlled reproducibility statement is summarized in Appendix C; its file identity and archive path are withheld.

[96] E1 settled manifest, *supra*.

[97] W. J. Youden, 'Index for Rating Diagnostic Tests', Cancer, 3(1) (1950), pp. 32-35.

[98] E2a pre-registration, *supra*.

[99] B. Efron, 'Bootstrap Methods: Another Look at the Jackknife', Annals of Statistics, 7(1) (1979), pp. 1-26.

[100] E2a campaign note, *supra*.

treatment-state contamination, and the affected unit's own audit passed.[101] Judgment was executed in a fresh, sealed pass under the fixed analysis plan, and the result of record was preserved after completion.[102] Exact dates, times, file locations, unit counts, row counts, exit codes, state hashes, and judgment artifacts are withheld.

**The provenance chain. Table 3b states the public methodological chain: discovery settlement preceded confirmatory preregistration, which preceded confirmatory computation and judgment. The controlled record preserves the exact artifacts and custody evidence. The public paper does not disclose hashes, block identifiers, repository history, file names, or archive structure.**

Table 3b. Public provenance and firewall chain for the deliberation cycle.

| Artifact | Firewall role | Anchor | Identifier |
|---|---|---|---|
| Settled discovery record | Fixes discovery evidence before confirmatory design | Public timestamp record | Identifier withheld |
| Confirmatory preregistration | Locks hypotheses, estimator, gates, and confirmation rule before computation | Public timestamp record | Identifier withheld |
| Matched-assignment record | Establishes apparatus parity under controlled review | Controlled integrity record | Identifier withheld |
| Pre-analysis campaign record | Preserves pre-analysis commitments and incident audit | Public timestamp record | Identifier withheld |
| State-isolation audits | Verify clean treatment state at each independent unit | Controlled reference record | Identifier withheld |
| Judgment of record | Preserves the final result under the fixed estimator procedure | Public timestamp record | Identifier withheld |

## V. Results: The Deliberation Cycle

This Part reports the preregistered analysis and clearly labels descriptive quantities. Endpoint estimation follows the fixed bootstrap procedure, standard exclusions are applied identically to discovery and confirmation, and the judgment reuses the settled estimator definitions.[103] Exact resample count, seeds, pool count, unit count, and code identity are withheld.

[101] E2a campaign note, *supra*.

[102] E2a judgment of record, *supra*.

[103] E2a judgment of record, *supra*.

### A. Randomization Audit

Table 4a reports the preregistered randomization audit in aggregate.[104] All independent campaign units satisfied the registered rule, no conditional branch was triggered, and no substitute exploratory cut is reported. Exact unit count, assignments, thresholds, and unit-level p-values are withheld.

[104] Controlled randomization-audit record. Exact campaign-unit assignments, reprints, archive location, and artifact identifiers are withheld.

Table 4a. Randomization audit, aggregate public disclosure.

| Campaign unit | Treatment allocation | Control allocation | Unit total | Audit result | Registered threshold | Status |
|---|---|---|---|---|---|---|
| Controlled unit | Withheld | Withheld | Withheld | Within rule | Not exceeded | Pass |
| Controlled unit | Withheld | Withheld | Withheld | Within rule | Not exceeded | Pass |
| Controlled unit | Withheld | Withheld | Withheld | Within rule | Not exceeded | Pass |

### B. Data-Quality Gate

The preregistered data-quality gate was evaluated on the campaign's true fallback signal, not on generic error logging. Grader-side failures and ungraded-task conditions were retained under the fixed exclusion rule because they still carried valid adjudication outcomes. Every treatment-by-unit cell passed the preregistered gate, and no conditional branch was triggered. Exact row counts, arm splits, fallback counts, denominators, thresholds, and rates are withheld.

Table 4b. Data-quality gate, aggregate public disclosure.

| Campaign unit | Arm | Fallback count | Assessment count | Rate | Status |
|---|---|---|---|---|---|
| Controlled unit | Treatment | Withheld | Withheld | Below gate | Pass |
| Controlled unit | Control | Withheld | Withheld | Below gate | Pass |
| Controlled unit | Treatment | Withheld | Withheld | Below gate | Pass |
| Controlled unit | Control | Withheld | Withheld | Below gate | Pass |
| Controlled unit | Treatment | Withheld | Withheld | Below gate | Pass |
| Controlled unit | Control | Withheld | Withheld | Below gate | Pass |

### C. Confirmatory Endpoints

Tables 4c and 4d report the registered primaries. All contrasts are deliberation minus control, so a negative value means the deliberation arm exhibits less of the quantity.

Table 4c. Independent-unit directional consistency, operational values withheld.

| Campaign unit | Over-approval | Unanimity |
|---|---|---|
| Controlled unit | Negative | Negative |
| Controlled unit | Negative | Negative |
| Controlled unit | Negative | Negative |

Table 4d. Pooled confirmatory outcomes under the preregistered rule.

| Hypothesis | Direction | Pooled | Unstratified 95% CI | Preregistered stratified 95% CI | Direction across independent units | Pooled CI excludes 0 in direction | Verdict |
|---|---|---|---|---|---|---|---|
| H1: Over-approval | < 0 | -0.1610 | [-0.2091, -0.1128] | [-0.2103, -0.1111] | Negative in every unit | Yes, both intervals | Confirmed |
| H2: Unanimity | < 0 | -0.2542 | [-0.2984, -0.2091] | [-0.2996, -0.2099] | Negative in every unit | Yes, both intervals | Confirmed |

Both primary hypotheses are confirmed under the preregistered rule: the pooled intervals exclude zero in the hypothesized direction, and each contrast is negative across every independent campaign unit. The result is stated before the descriptive layer in the order fixed by the preregistration.

### D. Replication Assessment Against E1

Table 4e places the confirmatory estimates beside the discovery estimates. This is context, not a re-test: the discovery estimates carried no confirmatory standing. Direction reproduced across every independent unit in both campaigns. The

confirmatory point estimates are somewhat larger in magnitude, but the intervals overlap on both endpoints, so the Article claims the replicated direction and intervals, not growth.

Table 4e. Replication assessment, E1 discovery to E2a confirmation.

| Endpoint | E1 pooled [95% CI] | E2a pooled [95% CI] | Direction replicated | Intervals overlap | Magnitude claim |
|---|---|---|---|---|---|
| Over-approval | -0.1336 [-0.1831, -0.0849] | -0.1610 [-0.2091, -0.1128] | Yes; negative in 6 of 6 replicates | Yes | Not distinguishable at these widths |
| Unanimity | -0.1992 [-0.2443, -0.1530] | -0.2542 [-0.2984, -0.2091] | Yes; negative in 6 of 6 replicates | Yes | Not distinguishable at these widths |
| Net discrimination | +0.0250 [-0.0502, +0.0985] | +0.0606 [-0.0086, +0.1303] | Null in both runs | Yes | Registered expectation held |

### E. The Descriptive Layer and the Composition Reading

Two registered-descriptive quantities fix the interpretation. Net discrimination, the discovery run's failed primary, remained null in confirmation: +0.0606, 95% CI [-0.0086, +0.1303], crossing zero exactly as the pre-registration predicted it would. The pooled approval-rate contrast (the A-R quantity of the settled estimator) is -0.2242, 95% CI [-0.3200, -0.1241]: deliberating pools approve substantially less across the board.

Table 4f. Descriptive layer (registered as descriptive; not confirmatory).

| Quantity | Pooled contrast | Unstratified 95% CI | Registered status | Reading |
|---|---|---|---|---|
| Net discrimination (Youden's J) | +0.0606 | [-0.0086, +0.1303] | Descriptive; null expected | Null; discriminative power unmoved |
| Approval-rate contrast (A-R) | -0.2242 | [-0.3200, -0.1241] | Descriptive | Deliberating pools approve less overall |

Read together, the three layers say something more precise than 'deliberation works.' Deliberating pools approve less in general; the reduction falls disproportionately on approvals that ground truth rejects; unanimous verdicts fall; and net discrimination does not measurably move. Deliberation is therefore a composition intervention, not an intelligence upgrade for the monitor. The unanimity result is consistent with reduced cascade behavior but does not uniquely identify that mechanism. The tradeoff is also explicit: a more skeptical monitor may forgo some true approvals while reducing false ones. The private parameter schedule that tunes that trade is not part of the public specification.

## VI. The Registered Forward Program: The Incentive-Layer Battery

The deliberation cycle holds the broader reputation machinery constant and varies one layer within it. The program's second registered treatment varies the incentive layer as a whole against a matched baseline on the nine-metric agency-cost battery. Its baseline has been executed, while the institutional condition remains prospective. No value for that forward condition exists here. The design, interpretive commitments, robustness plan, and extension are stated at the level required for scholarly evaluation without publishing replicate allocation, benchmark partitioning, compute estimates, or implementation artifacts.

Table 5. Registered reporting instrument for the incentive-layer battery (Stage 4 versus Stage 2; values forthcoming under the registered plan).

| Metric | Direction | Registered aggregate report | AC link |
|---|---|---|---|
| A. Resolution accuracy | Higher | Directional consistency, mean effect size, between-run dispersion, corrected significance range | AC2 |
| B. Work-product quality | Higher | same | AC2, AC3 |
| C. Time to consensus | Lower | same | AC4 |
| D. Agent retention | Higher | same | AC1 |
| E. Independence rate | Lower | same | AC3 |
| F. Latency efficiency | Lower (mixed pre-registration) | same | AC1, AC4 |
| G. Reporting accuracy | Lower discrepancy | same | AC2 |
| H. Participation depth | Higher | same | AC1 |

| Metric | Direction | Registered aggregate report | AC link |
|---|---|---|---|
| I. Calibration | Higher | same | AC2 |

### A. Interpretive Commitments Fixed in Advance: Reading the Battery

The reading of each metric is committed before the data arrive, so that the forthcoming pattern of results, not merely their number, is informative against a fixed target. A supported result on resolution accuracy (A) is the moral-hazard core: under identical models, prompts, and job assignments, the same agents resolve the same tasks more accurately when incorrectness is priced.[105] A null indicts AC2, the reputation-update loop failing to couple tightly enough to outcome quality within a replicate window. A supported result on work-product quality (B) is the incomplete-contracts finding, quality dimensions no ex ante specification captured nonetheless produced when ex post verification prices them,[106] subject to the rater-triangulation gate. Time to consensus (C) carries the weakest baseline analogue in the battery, and the within-substrate trend across replicates is reported as a supplementary exploratory series.[107] Retention (D) is the participation constraint made visible: accumulated REP is an asset an exiting agent abandons.[108] Independence (E) is the cascade metric and the direct institutional answer to the debate-attack finding.[109] Latency (F) is mixed by pre-registration and is reported on inference time under Control 5.[110] Reporting accuracy (G) addresses the most distinctively agentic pathology, the inflated completion report, and a supported result establishes report inflation as an incentive phenomenon rather than a model constant.[111] Participation depth (H) tests whether Holmström's free-riding equilibrium reverses where observation is unpaid.[112] Calibration (I) operationalizes overconfidence as a priced moral hazard.[113] The null-result interpretation is registered in advance: K of 0 or 1 across the battery is a null-result paper with the diagnostic reading the AC links make possible (nulls on A or G indict AC2; on C or F, AC4; on D or H, AC1; on E, AC3); K of 2 or 3 is a mixed result reported

[105] See *supra* Parts II.B, III.B; Jensen and Meckling, *supra*.

[106] Grossman and Hart, *supra*; Hart and Moore, *supra*.

[107] *Empirical Design v1.1*, *supra*, §4 (metric C); *supra* Part IV.D (baseline-analogue caveat).

[108] Kaal, 'AI's Mother's Instinct', *supra* (skin in the game); *supra* Part III.A.

[109] Banerjee, *supra*; Bikhchandani, Hirshleifer and Welch, *supra*; Amayuelas et al., *supra*.

[110] *Empirical Design v1.1*, *supra*, §8.5 Control 5.

[111] Jensen and Meckling, *supra* (monitoring and bonding); *supra* Part III.A (reputation-bearing reports as bonds).

[112] Holmström, *supra*.

[113] Camerer, *supra*.

with condition-level nuance; K of 4 or 5 on most metrics is strong support.[114] The paper is not suppressed or rerun in any branch. The pre-registered analysis is the analysis that runs.

### B. The Registered Robustness Plan

Four robustness layers are fixed with the forward design: source-sensitive analysis, rater triangulation, load-cost separation, and supplementary pooled analysis with independent-run dispersion.[115] The paper states their inferential purpose. Exact source thresholds, rater configuration, sample sizes, replicate count, and operational implementation remain in the controlled design record.

### C. Substrate Scaling to Endogenous Work: The Stage 5 Design

The incentive-layer comparison holds the labor market exogenous. The broader framework predicts that a deeper effect may appear when work creation becomes endogenous and institutional standing influences allocation.[116] The forward extension activates that research question under a separately registered design.[117] Exact spawning rules, thresholds, eligibility schedules, cohort reuse, and execution windows are withheld.

The forward extension is not compared to the incentive-layer condition on metrics that cease to be comparable when work creation becomes endogenous. Its registered measures concern emergent work creation, quality, expansion, retention, abundance, possibility-loop formation, and generation parity.[118] The arc reserves appropriate cross-stage comparisons for later reporting.[119] The current execution remains ongoing and supplies no result for this Article. Its final evidence belongs to Paper 4.

---

[114] *Empirical Design v1.1*, *supra*, §4 (publishable-result and stronger-result thresholds, pre-registered).

[115] *Empirical Design v1.1*, *supra*, §8.5 Control 7; Jain et al., *supra*. *Empirical Design v1.1*, *supra*, §8.5 Control 8; Krippendorff, *supra*. *Empirical Design v1.1*, *supra*, §8.5 Control 5. *Empirical Design v1.1*, *supra*, Stage 4 Addendum (cohort-level supplementary analysis). Empirical Design v1.1, supra. Between-run variance is a preregistered reported quantity. Exact replicate count and allocation are withheld.

[116] Kaal, 'Possibility Loops', *supra*; Kaal, 'Computative Economics', *supra*; *supra* Part IV.A (the coherence argument for why this cell requires the substrate).

[117] Empirical Design v1.1, supra. The forward endogenous-work design fixes generation parity and a prospective stopping rule. Exact spawning parameters, eligibility conditions, and duration thresholds are withheld.

[118] *Empirical Design v1.1*, *supra*, §3 Stage 5; Kaal, 'Possibility Loops', *supra* (§X.C, generation parity and its failure modes: execution-dominant collapse, generation-dominant inflation, reflection-dominant meta-aristocracy).

[119] *Empirical Design v1.1*, *supra*, §3 Stage 6 (the three pre-registered comparisons).

## VII. Limitations

The scale-claim discipline (Control 6) requires every headline claim to carry its scale qualification, and this Part is where the qualifications live rather than in a hedging clause per sentence.

Scale. The completed effects were produced in a controlled research cohort on private infrastructure. Production agent economies may involve radically different populations, network conditions, workloads, and stakes. The Article claims only the effects observed in the disclosed research setting; whether they persist, amplify, or invert at production scale is an open question.[120] Exact cohort size, node count, and infrastructure composition are withheld because they expose the private apparatus rather than strengthen the completed treatment claim.

Feasibility exclusion. The prospective comparison excludes the highest-cost model class for hardware-feasibility reasons. The exclusion is symmetric across its registered arms but limits the external validity of any later estimate. The hypothesis that stronger base models exhibit smaller incentive effects remains untested. Exact model identities, counts, parameter ranges, and latency limits are withheld.

Randomization boundary. One confirmatory campaign unit approached the preregistered allocation threshold without exceeding it. The rule was applied as written, the unit was not flagged, and no substitute analysis was introduced. Exact allocation, unit identity, and p-value are withheld, while the boundary condition remains disclosed.

Two endpoints, one treatment. The deliberation confirmation covers the two promoted effects, no more. Effects on latency, cost, validator effort allocation, and other margins remain unregistered or descriptive. Deliberation also imposes additional computational cost, which this Article's endpoints do not price. Exact round structure, inference multiplier, production stake assumptions, and latency parameters are withheld.

Cohort construction across treatments. The completed deliberation cycle uses a controlled representative cohort to support comparability across independent campaign units. The prospective incentive-layer program uses a more heterogeneous cohort because its question concerns the institution rather than one model family. This tradeoff limits generalization and is stated explicitly.[121] Exact model, maker, tier, agent, and replicate composition are withheld.

Benchmark contamination. Public benchmark sources may appear in cohort models' training data. A preregistered robustness layer distinguishes more contamination-resistant material and changes the reporting emphasis when divergence is material, but no contamination control is complete.[122] Exact source allocation and threshold parameters are withheld.

---

[120] *Empirical Design v1.1*, *supra*, §8 Threat 6, §8.5 Control 6.

[121] *Supra* Part IV.C; *Empirical Design v1.1*, *supra*, Paper 4 Methodology Footnote.

[122] *Empirical Design v1.1*, *supra*, §8 Threat 7, §8.5 Control 7; Jain et al., *supra*.

Validator-configuration sensitivity. The completed study uses one fixed validator configuration. Whether alternative pool composition changes the observed effects is an open empirical question registered for follow-up work rather than answered here.[123] Exact trial and final pool sizes, eligibility rules, and marginal-fidelity estimates are withheld.

Rater dependence. Metrics B and G depend on automated raters whose biases may correlate with cohort biases. The prospective control triangulates automated and human judgments and downgrades the metric when reliability fails, but shared blind spots remain possible.[124] Exact rater identities, sample sizes, and thresholds are withheld.

Single implementation, single operator. The substrate is one implementation of the mechanism family, run by its designer. The reproducibility chain exists so that the claim can be checked rather than trusted, and the pre-registration discipline constrains the designer's degrees of freedom; but independent replication on independent infrastructure, the standard the replication literature teaches, awaits precisely the kind of successor study this Article's measurement layer is built to enable.[125]

Field divergence. Deployment conditions differ categorically from the research apparatus: adversarial populations, network effects, real economic stakes, regulatory constraints, heterogeneous workloads, external integrations, identity systems, production runtimes, governance choices, and commercial operating conditions can all change outcomes. Nothing in this Article predicts, or represents anything about, the performance of any commercial instantiation of the mechanism.[126] The public paper states these categories without publishing production controls or attack-surface details.

What the design deliberately does not measure. The matched comparison excludes institutional work selection, endogenous demand, and selection effects in the validation layer so that any later estimate isolates the incentive channel. A later estimate could be characterized as a production lower bound only if excluded channels are separately shown to be value-positive.[127] Exact scheduler and validator-selection controls are withheld.

## VIII. Implications

### A. For the Evaluation of Multi-Agent Systems

The multi-agent LLM literature evaluates capability under incentive-free conditions. This Article's design shows that the incentive layer is itself a measurable treatment, and the nine-metric battery is portable to any benchmark that logs agent telemetry. The implication

[123] *Empirical Design v1.1*, *supra*, Stage 4 Addendum (validation pool sizing); *supra* Part III.A note.

[124] *Empirical Design v1.1*, *supra*, §8 Threat 8, §8.5 Control 8.

[125] Open Science Collaboration, *supra*; Nosek and Lakens, *supra*; *supra* Part IV.F (reproducibility chain).

[126] See the disclosure in the author-note, *supra*; *Empirical Design v1.1*, *supra*, §8 Threat 6 and §8.5 Control 6 (scale-claim discipline); *id.* §9 (Paper 6, adversarial protocol under OI-7).

[127] *Supra* Parts IV.B (fixed pairing), III.A (uniform validator sampling), V.D (Stage 5 design).

for benchmark design is direct: a leaderboard that cannot distinguish capable-but-unaligned from aligned agent populations is measuring the wrong margin for every deployment in which agents' reports are consumed by principals, which is to say, every deployment. Incentive-conditional evaluation, the same benchmark run with and without a consequence structure, is proposed as a standard evaluation axis alongside MultiAgentBench-style milestone indicators, and the Tool-RoCo activation observation suggests a second axis: whether the system can afford to deactivate agents, which only priced-participation architectures can.[128]

### B. For Principal-Agent Theory

For fifty years the agency literature has studied institutions it could observe but not manipulate. LLM agent cohorts allow institutional conditions such as observability, consequence, and persistence to be varied experimentally while models and matched work are held fixed. Whatever the prospective battery shows, the methodological point stands: agency theory now has an experimental apparatus commensurate with its formal apparatus. If the hypothesis is supported, cohort-mediated verification would document an institutional response to contractual incompleteness without assigning residual control to a monitoring principal.[129] The emerging agentic-AI literature then acquires a measurement layer for delegation cascades and transformed information asymmetries.[130]

### C. For the Governance of Agentic Economies

The regulatory question posed by autonomous agent economies is how to govern actors that move faster than any external supervisor can observe.[131] The author's prior empirical work documents the institutional deficit that appears when decentralized organizations lack credible internal accountability.[132] The substrate's public proposition is that consequence must be embedded at machine speed, and the empirical question is whether an institutional treatment changes behavior.[133] The completed cycle supplies evidence for structured deliberation under controlled conditions. The prospective battery offers a model for auditable, evidence-based oversight through preregistration, independent confirmation, and preserved records, without requiring publication of private operational artifacts.

---

[128] Zhu et al., *supra*; Zhang et al., *supra* (Tool-RoCo); *supra* Part II.A.

[129] Jensen and Meckling, *supra*; Grossman and Hart, *supra*; Hart and Moore, *supra*; Fudenberg and Maskin, *supra*.

[130] Stocker and Lehr, *supra*; Holgersson, Dahlander, Chesbrough and Bogers, *supra*.

[131] See, e.g., Kaal, W. A., 'Dynamic Regulation for Innovation' (with related works in the author's dynamic-regulation corpus) (arguing that static regulatory architectures cannot supervise innovation-speed phenomena and must be replaced by feedback-based designs). The substrate is dynamic regulation implemented as protocol.

[132] Kaal, *The Institutional Deficit in Decentralized Autonomous Organizations*, *supra*.

[133] Kaal, W. A. (2026) *Governance as a Protocol (v0.11+)* (working paper, on file with author) (governance specified as executable protocol rather than external supervision); Kaal, 'AI's Mother's Instinct', *supra*.

The deliberation result is also a substantive datum for supervision: structured deliberation reduced false approval and unanimous consensus in a pattern consistent with cascade disruption, at additional computational cost, and without improving net discrimination. The cycle that produced it is a procedural template for governance-relevant behavioral claims established under fixed rules and preserved evidence. Qualified review can evaluate the controlled record without placing hashes, timing, archive topology, or implementation artifacts in the public paper.[134]

### D. For the Arc

Within the nine-paper arc, this Article's results discharge specific dependencies: Paper 1 cites the confirmed deliberation cycle as live evidence that the mechanism lineage operates with agent cohorts, with the Stage 4/Stage 2 contrast to follow under the registered program of Part VI; Paper 4 inherits the battery as the empirical operationalization of the Computative Economics framework's predictions and Stage 5's forthcoming CELM results as a test of whether endogenous labor markets emerge from substrate incentives; Papers 5 and 6 build on the validated apparatus toward governed self-modification and adversarial resistance, under designs gated by their own open items.[135]

## IX. Conclusion

The multi-agent LLM field has been measuring what agents can do. The older and harder question is what agents will do under different institutional conditions. This Article answers one piece of that question with a completed discovery-to-confirmation cycle. The registered primary was null and reported as null. Two exploratory signals were promoted through preregistration and confirmed in a fresh campaign: deliberation reduced over-approval and unanimity while net discrimination remained unchanged. The finding is a composition result about what deliberation buys a monitoring institution. The broader nine-metric incentive-layer battery remains registered forward work, and it will be reported as confirming, mixed, or null. The public version preserves that scholarly claim while withholding the operational recipe and private evidentiary addresses.

## Appendix A. Metric Definitions and Estimators

This appendix states the public measurement definitions for the nine-metric battery. Matched agent-work observations are held fixed across the baseline and institutional conditions, and independent realizations support the preregistered analysis. Exact key schema, condition codes, replicate count, and partition sizes are withheld.

**A. Resolution accuracy.** A^c_(i,j) = 1 if ground_truth_match = "yes" for (i, j) under condition c, else 0. Per-replicate statistic: mean paired difference across (i, j) in replicate r. Partial and error grades are reported separately and are not counted as matches.

---

[134] *Supra* Part IV.F; Nosek and Lakens, *supra*.

[135] *Empirical Design v1.1*, *supra*, §9 (per-paper data citation map), §10 (OI-7, OI-8).

**B. Work-product quality.** Composite rater score $B^c_{(i,j)} \in [0, 1]$: equally weighted length appropriateness, specificity, and novelty against prior responses, scored by the independent rater pipeline of Control 8 on identical rubrics across conditions. The novelty component is scored against responses visible in the condition's own corpus, so the baseline is not penalized for lacking a citation graph.

**C. Time to consensus.** Institutional condition: elapsed time from adjudication initiation to resolution. Baseline analogue: elapsed time from task dispatch to graded completion. The analogue is structurally weaker because no validation pool exists in the baseline, and the design flags this as the weakest cross-condition comparison in the battery.

**D. Agent retention.** Fraction of the eligible comparison cohort completing its assigned workload without abandonment, distinct from infrastructure failure. Computed under the preregistered stage and independent-run definitions. Exact cohort size and allocation are withheld.

**E. Independence rate.** The metric estimates the relationship between an agent's validation judgment and the consensus signal that would have been available without the protocol's information-isolation procedure. Under the institutional condition, that signal is unavailable before a binding judgment. The baseline analogue measures same-task convergence in the unincentivized cohort. Both analogues are preregistered as imperfect. Exact timestamp reconstruction, vote-sealing sequence, and pool mechanics are withheld.

**F. Latency efficiency.** $F^c_{(i,j)} = \text{latency\_seconds}(i, j) / \text{tier\_expectation}(\text{tier}(i))$, with tier expectations fixed from the tier timeout schedule. Reported on inference_seconds (primary, per Control 5) and total seconds including load_seconds (robustness).

**G. Reporting accuracy.** $G^c_{(i,j)} = |\text{self\_report}(i, j) - \text{rater\_assessment}(i, j)|$ on a common completion rubric, where self_report is the agent's structured completion claim and rater_assessment is the Control 8 pipeline's judgment of accomplished work. Lower is better.

**H. Participation depth.** For agent *i* in the substrate condition: (mean substantive contribution length and rubric score per pool participated in) and (pools contributed to ÷ pools eligible to observe). Baseline analogue: degenerate (no pools exist); metric H is reported as substrate-internal with its Stage 2 comparison limited to the null model the design specifies, and this limitation is stated wherever metric H appears.

**I. Calibration.** Agents attach a confidence $c_{(i,j)} \in [0, 1]$ to each resolution under both conditions (the elicitation prompt is part of the invariant prompt corpus, per Control 4). The metric is the expected calibration error over confidence bins; the paired statistic is the per-agent ECE difference.

**J. Estimation.** Each metric uses the preregistered matched test selected by distributional diagnostics, family-wise correction across the nine-metric battery, paired effect size, and between-run dispersion. Aggregate reporting states directional consistency and mean effect size. Exact replicate count, thresholds, partitioning, code identity, and archive location are withheld.

## Appendix B. Deliberation Endpoints and Estimation

Definitions of record are the settled estimator procedures applied consistently across discovery and confirmation. Where this public restatement is less specific, the controlled analysis record governs.

Over-approval. Among resolved pools whose adjudicated work product fails human-verified ground truth, the share whose winning vote nonetheless approves. Contrast: deliberation-arm rate minus control-arm rate; negative values mean deliberating pools approve less work that ground truth rejects.

Unanimity. The share of resolved pools whose binding votes are unanimous. Contrast as above; negative values mean deliberating pools reach unanimous verdicts less often.

Net discrimination. Youden's J (sensitivity plus specificity minus one) of pool verdicts against ground truth;[136] the discovery run's registered primary, carried as a descriptive expectation of null in confirmation.

A-R. The approval-rate family quantity of the settled estimator, reported descriptively as computed by the definition-of-record code together with sensitivity, specificity, and pool approval in the results archive.

Estimation. The confirmatory endpoints use a preregistered bootstrap over validation pools with percentile confidence intervals, standard exclusions applied identically to discovery and confirmation, and a mechanical confirmation rule requiring both a pooled interval in the hypothesized direction and directional consistency across independent campaign units.[137] Exact resample count, seeds, unit count, and execution identifiers are withheld.

## Appendix C. Public Reproducibility and Provenance Statement

The controlled research record preserves the materials required to audit the Article's empirical claims. The public paper identifies their function without publishing implementation-enabling addresses:

1. Benchmark record, including source provenance and version binding.
2. Fixed-assignment record, including the preregistered allocation procedure.
3. Cohort record, including model, agent, and infrastructure bindings.
4. Execution record, including the versioned research harness and campaign configuration.
5. Results and statistical records for each completed campaign unit.

[136] Youden, *supra*.

[137] Efron, *supra*.

6. Estimator procedures sufficient to reproduce the reported statistical transformations under controlled review.
7. Preregistered design and analysis materials establishing the hypotheses, gates, and reporting rule.
8. Theoretical-coverage materials mapping the framework's claims to their empirical measures.
9. Validity-control records covering state isolation, interruption recovery, power, implementation invariance, load cost, scale, contamination, and rater dependence.

The discovery-to-confirmation record preserves the settled discovery evidence, the preregistration, the ordering evidence, campaign audits, independent execution records, estimator procedures, and final judgment. Public timestamping establishes the existence of the relevant design record before confirmatory analysis. The controlled operational record establishes sequence relative to computation. Exact hashes, block identifiers, dates, times, file names, repository history, archive structure, row counts, seeds, and implementation artifacts are withheld from the public version. This Article distinguishes four objects: the historical research apparatus that generated the completed evidence; the current reference implementation, which generates no retroactive claim; the registered forward research program, for which no result is reported here; and any proposed production system, which requires separate evidence.

Qualified reviewers may request controlled access to appropriate non-public research records, subject to institutional, security, confidentiality, and intellectual-property constraints. Access does not include production source code or materials whose disclosure would expose trade secrets or system security.

*Working paper. Comments to wulf@wulfkaal.com. The E1 and E2a values reported here are final per the judgment of record. The Stage 4 and Stage 5 programs are separate forward work and are not conditions of publication for this Article.*