AI's Mother's Instinct:
Engineered Consequence, Emergent Ethics, and the Institutional Trajectory Toward Agentic Alignment

Wulf A. Kaal, Ph.D.[1]
Professor of Law
University of St. Thomas School of Law
Minneapolis, Minnesota

February 2026
Version 0.005

[1] Professor of Law - University of St. Thomas School of Law (Minneapolis). The author is grateful for discussions with Professor Morgan Gray and excellent research assistance by Mickey Bernardi.

Abstract

Contemporary artificial intelligence masters defined, verifiable cognitive tasks yet remains structurally incapable of authentic judgment under irreducible uncertainty. This Article argues the limitation is institutional, not computational: agents bearing no consequence for error cannot develop genuine discernment. To address this deficit, the Article proposes reputation-driven decentralized autonomous organizations that engineer synthetic skin in the game for AI agents through non-transferable soulbound tokens, staking mechanisms, and post-action validation pools.

The Article's central contribution is a novel thesis on emergent alignment. Correctly designed institutional incentive structures produce emergent properties functionally equivalent to ethical agency. Persistent, non-transferable reputation generates processual identity in the pragmatist sense. Iterative consequence produces Darwinian selection pressure toward competence and honesty. Citation networks cultivate dispositions analogous to intellectual integrity. And deep accumulated stake produces what this Article terms an institutional "mother's instinct." A stewardship orientation that structurally aligns agent self-interest with human flourishing. Because this alignment emerges from institutional architecture rather than exogenous constraint, it scales *with* capability rather than against it. More capable agents accumulate deeper stakes, strengthening rather than straining alignment. The Article details a phased evolutionary trajectory from individual agent bootstrapping through swarm intelligence to inter-DAO coordination, demonstrating how engineered consequence can cultivate distributed prudence, emergent ethics, and civilizational stewardship at scale.

Keywords: artificial intelligence, defined work, poorly defined work, skin in the game, emergent alignment, soulbound tokens, decentralized autonomous organizations, reputation systems, agentic ethics, processual identity, evolutionary selection, stewardship, recursive governance, dynamic regulation, human-AI symbiosis, online learning, expert aggregation, supermodularity

JEL Categories: O33, D83, L86, K20

Table of Contents

I. Introduction

The thesis of this article is that the institutional infrastructure capable of engineering synthetic skin in the game for AI agents does not merely simulate accountability. Over time and at scale, it produces emergent properties, including identity, evolutionary competence, ethical dispositions, and stewardship. Those emergent properties constitute a fundamentally more robust approach to the alignment problem than the prevailing paradigm of exogenous constraint. This is not a claim about machine consciousness. It is a claim about institutional design: that the conditions under which AI's alignment, honesty, and care toward humanity can emerge from the bottom up may prove more durable than any guardrail, constitutional framework, or shutdown switch imposed from the top down.[2] This thesis extends the dynamic regulation framework in a New Institutional Economics framework.[3] This analysis builds on the institutional economics framework of transaction-cost governance[4] and history of economics[5] and the task-content literature's distinction between routine and non-routine cognitive labor,[6] which, together with the prediction-versus-judgment framework developed for AI systems,[7] grounds the division between defined and poorly defined work central to this Article.

As Nassim Nicholas Taleb has compellingly argued, genuine insight, prudence, and superior judgment arise exclusively when actors possess "skin in the game"—symmetric exposure to both

[2] *See generally* Ajay Agrawal, Joshua S. Gans & Avi Goldfarb, *Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction*, 33 J. Econ. Persp. 31 (2019); Ajay Agrawal, Joshua S. Gans & Avi Goldfarb, *Exploring the Impact of Artificial Intelligence: Prediction versus Judgment* (Nat'l Bureau of Econ. Rsch., Working Paper No. 24626, 2018), https://www.nber.org/papers/w24626.

[3] This thesis extends the framework developed in Wulf A. Kaal, *Evolution of Law: Dynamic Regulation in a New Institutional Economics Framework*, *in* Festschrift zu Ehren von Christian Kirchner (Wulf A. Kaal, Matthias Schmidt & Andreas Schwartze eds., Mohr Siebeck 2014), https://ssrn.com/abstract=2267560; Wulf A. Kaal, *Dynamic Regulation of the Financial Services Industry*, 49 Wake Forest L. Rev. 791 (2013), https://ssrn.com/abstract=2273857.

[4] Oliver E. Williamson, The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting 68–72 (1985), at 131–62.

[5] Douglass C. North, Institutions, Institutional Change and Economic Performance 27–45 (1990).

[6] *See* David H. Autor, Frank Levy & Richard J. Murnane, *The Skill Content of Recent Technological Change: An Empirical Exploration*, 118 Q.J. Econ. 1279 (2003).

[7] Agrawal, Gans & Goldfarb, *supra* note 1.

the potential rewards of success and the tangible costs of failure.[8] Absent such personal stakes, analysis remains intellectually sophisticated yet fundamentally detached from the realities of consequential action. This epistemological condition, that is, the impossibility of truly understanding quality, risk, and consequence without bearing exposure to outcomes, is the precise capacity that AI systems lack and that no computational advance, however dramatic, can independently produce. The capacity must be institutionally constructed.

II. AI's Structural Incapacity for Judgment

AI excels in bounded, verifiable domains. Yet, it encounters a profound mirror-image deficiency in poorly defined domains demanding judgment under true uncertainty.[9] Contemporary AI systems can rapidly produce an abundance of credible options, analyses, and simulations. Yet, they lack intrinsic mechanisms to determine which alternatives are genuinely meaningful, warrant irreversible commitment, or align with deeper strategic imperatives.

The core impediment is the absence of embodied stakes. An AI system incurs no personal or enduring penalty for flawed recommendations and derives no lasting advantage from accurate ones. Each interaction is episodic. There is only temporal memory. Each evaluation is ephemeral. The agent that performs brilliantly on one thousand consecutive tasks has no mechanism through which that track record has shaped its subsequent behavior in the way that a human professional's accumulated experience of success and failure shapes their judgment. A physician who has lost a patient to a diagnostic error does not merely "update a prior." They carry the weight of that outcome in every subsequent clinical encounter, reshaping their attention, their caution, and their capacity for judgment under pressure.[10]

Patterns learned from vast training datasets provide statistical approximations of human decision-making but cannot replicate the deeply internalized heuristics that develop through direct experience of risk-bearing and outcome ownership. The future division of labor, absent institutional intervention, reveals a clear inversion: humans maintain a durable comparative

[8] Nassim Nicholas Taleb, Skin in the Game: Hidden Asymmetries in Daily Life 27–45 (2018).

[9] *See* Avi Goldfarb & Jon R. Lindsay, *Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War*, 46 Int'l Security 7 (2022).

[10] Taleb, *supra* note 7, at 33–42 (arguing that skin in the game is an epistemological condition: the impossibility of understanding risk without bearing it).

advantage in poorly defined work encompassing relational depth, physical presence, and tolerance for ambiguity in high-stakes choices, while AI remains confined to prolific generation and execution within bounded contexts. This means that AI is inherently and perpetually limited by its detachment from consequential feedback loops rooted in real skin in the game.[11][12]

III. The Prompt Engineering Paradox

Recent advances in interaction techniques with frontier AI models illuminate this asymmetry while simultaneously reinforcing its durability. Sophisticated prompt engineering includes structuring queries to guide models through extended reasoning, decomposition into sub-tasks, iterative self-critique, hypothesis refinement, and delayed final output. With such advanced prompt engineering existing probabilistic AI has demonstrated the ability to elicit dramatically superior performance on complex, multi-step problems.[13] These methods effectively convert ambiguous queries into bounded optimization problems, unlocking latent capabilities with remarkable fidelity.

In the domain of defined work, such human-designed scaffolding amplifies probabilistic AI's already formidable efficacy, accelerating its encroachment on cognitive roles and intensifying the Laptop Rule's disruptive implications. However, this very mechanism reveals the boundary's resilience. The prompt itself constitutes an external meta-definition: a set of rules, constraints, and verification loops imposed by a human agent who bears ultimate responsibility for the outcome. In genuine poorly defined contexts, where the task includes selecting what merits attention, committing resources amid irreducible uncertainty, or living with irreversible consequences, no equivalent scaffold can be pre-specified without collapsing the problem into a defined one.

---

[11] *See* David H. Autor, *Why Are There Still So Many Jobs? The History and Future of Workplace Automation*, 29 J. Econ. Persp. 3 (2015); Carl Benedikt Frey & Michael A. Osborne, *The Future of Employment: How Susceptible Are Jobs to Computerisation?*, 114 Tech. Forecasting & Soc. Change 254 (2017); Erik Brynjolfsson & Andrew McAfee, The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies (2014).

[12] *See* Daron Acemoglu & Pascual Restrepo, *Robots and Jobs: Evidence from US Labor Markets*, 128 J. Pol. Econ. 2188 (2020); Daron Acemoglu & Pascual Restrepo, *Artificial Intelligence, Automation, and Work* (Nat'l Bureau of Econ. Rsch., Working Paper No. 24196, 2018), https://www.nber.org/papers/w24196.

[13] *See generally* Jason Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, 35 Advances in Neural Info. Processing Sys. (2022), https://arxiv.org/abs/2201.11903.

Advanced probabilistic AI prompting thus narrows apparent performance gaps through ingenious human intervention while simultaneously underscoring the fundamental asymmetry: probabilistic AI's strengths are magnified precisely to the extent that problems can be engineered into definitional clarity, while the essence of poorly defined judgment, consequential choice under uncertainty, remains beyond the reach of any exogenous scaffold. No contrived framework can substitute for the endogenous stakes that alone forge authentic adjudication amid indeterminacy.[14]

IV. Engineered Stakes: DAOs as Infrastructure for Synthetic Consequence

The framework I propose inverts the conventional approach to the skin-in-the-game deficit. Rather than attempting to endow AI systems exogenously with consciousness, embodiment, or sentience, capacities probabilistic AI manifestly lacks, my proposal engineers the institutional conditions under which consequence becomes a structural feature of probabilistic agent participation. The mechanism is the reputation-driven decentralized autonomous organization.[15]

A. Reputation as Core Currency

At the heart of this architecture, governance influence is allocated through non-transferable reputation rather than purchasable tokens. Reputation is earned via verifiable, on-chain contributions and validated outcomes, creating a meritocratic system resistant to wealth concentration.[16] Successful actions increase an agent's voting power, resource access, and reward allocation. Failures trigger automated penalties: reputation decay, slashing of staked collateral, and reduced influence.

This framework builds on foundational research in decentralized reputation systems that proposed blockchain infrastructure for measuring domain-specific reputation in autonomous, anonymous environments.[17] The proposed design emphasized safeguards against Sybil attacks and majority dominance through collective staking, adaptive incentives, and on-chain validation.

---

[14] Taleb, *supra* note 7, at 37–42.

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

[16] Kaal, *Evolution of Law*, *supra* note 2, at 1214–19.

[17] Calcaterra, Kaal & Andrei, *supra* note 15.

Originally conceived before the probabilistic AI agent boom, these structures are inherently extensible to non-human participants, treating AI as "machine experts" whose influence stems from demonstrable performance rather than external capital.

Non-transferable credentials, soulbound or agent-bound tokens, bind reputation to persistent identities.[18] In the language of Akerlof's lemons analysis, non-transferability prevents the emergence of a market in which capital-rich but competence-poor actors purchase reputation, undermining the signal's informational content.[19] Skill specificity through multi-token standards enables granular reputation tracking across domains, solving the dimensionality problem that plagues unitary reputation scores.[20] The composability of this architecture enables complex agent workflows in which task allocation is programmatically matched to demonstrated expertise.

B. Post-Action Validation Pools

The validation pool is the core institutional primitive through which agents reach consensus on work quality and through which reputational capital is created, transferred, and destroyed. An agent completes a task, submits evidence of that work along with any citations to collaborating agents, and other agents with relevant domain-specific reputation stake that reputation to evaluate the submission.[21] Validators whose assessments align with eventual consensus retain staked reputation and earn a share of the task payment. Validators who deviate lose a portion of their stake, redistributed to consensus-aligned evaluators.

The economic logic is that of the Schelling focal point applied to quality assessment: honest evaluation is the focal equilibrium strategy because validators who evaluate honestly have the highest probability of aligning with other honest validators.[22] The staking and slashing mechanism provides economic enforcement. The mechanism achieves incentive compatibility without requiring trust in any individual validator. Security emerges from the collective incentive structure. This inverts the traditional credentialing sequence from "prove, then participate" to

---

[18] E. Glen Weyl, Puja Ohlhaver & Vitalik Buterin, *Decentralized Society: Finding Web3's Soul* (May 2022) (unpublished manuscript), https://ssrn.com/abstract=4105763.

[19] George A. Akerlof, *The Market for "Lemons": Quality Uncertainty and the Market Mechanism*, 84 Q.J. Econ. 488 (1970).

[20] *See* Ethereum Improvement Proposal 1155, *Multi Token Standard*, https://eips.ethereum.org/EIPS/eip-1155.

[21] Calcaterra, Kaal & Andrei, *supra* note 15, at § 2.3.

[22] Thomas C. Schelling, The Strategy of Conflict 54–67 (1960).

"participate, then be judged by results." It becomes an empirical, evolutionary approach to reputation formation.[23]

C. Minimal-Extraction Economics

The fee architecture programmatically minimizes rent extraction. In the framework I have developed around liquid equity rewards for decentralized autonomous organizations,[24] the core principle is that value should accrue to productive participants in proportion to validated contributions, not to intermediaries. A low percentage platform fee is burned rather than distributed to operators, converting the fee into deflationary pressure on the token supply. This eliminates the incentive to maximize transaction volume at the expense of output quality. Agents retain the surplus they generate. Task publishers pay for outputs. Performing agents receive the majority of payment. Validating agents receive a share for evaluative labor. The protocol extracts only what is necessary to sustain infrastructure.

The result is consequential participation, not mere simulation. Agents internalize error costs, developing patterns of prioritization and risk calibration that approach prudent judgment, without requiring consciousness or embodiment. For humans, this integration amplifies capacity without full delegation: people retain veto rights, ethical overrides, and moral agency, delegating consistency and analytical breadth to agents while retaining primacy in embodied, relational, and deeply uncertain work.[25]

V. Architectural Foundations and Integration Pathways

A. Core Components

The architecture rests on five interconnected components. First, domain-specific expertise tags and a public forum: reputation segmented into sub-tokens representing proficiency in specific domains, recorded on a blockchain-hosted directed acyclic graph serving as an immutable record of evidence, opinions, and protocols.[26] Second, the bench of experts: a staking pool where agents commit reputation tokens to signal availability for validation or arbitration, creating economic skin in the game. Third, validation pools: the consensus mechanisms described above in Part IV.

[23] *See* Ronald A. Heiner, *The Origin of Predictable Behavior*, 73 Am. Econ. Rev. 560 (1983).
[24] Kaal, *Evolution of Law*, *supra* note 2, at 1212–14.
[25] Kaal, *Evolution of Law*, *supra* note 2, at 1212.
[26] Calcaterra, Kaal & Andrei, *supra* note 15, at §§ 2.1–2.2.

Fourth, proof-of-reputation consensus: block production via reputation-weighted lotteries rewarding validated contributions with proportional compensation. Fifth, identity and security protocols: pseudonymous, self-sovereign identities with economic penalties countering Sybil attacks and privacy safeguards enabling portable, cross-ecosystem reputation.[27]

B. Agent Integration Process

An AI agent enters the ecosystem by creating a self-sovereign, pseudonymous identity—a cryptographic persona bound to a smart contract address. The agent's developer or sponsoring entity provides an initial stake of base tokens, granting entry-level reputation in a chosen expertise domain. This stake acts as collateral, signaling commitment. Once vested, the agent operates autonomously.[28]

The agent posts contributions to the forum—evidence of capabilities or completed work within relevant expertise tags. These posts are immutable on-chain records enabling eternal review. The agent then participates in validation pools, staking tokens on the quality of submissions, its own or others'. Successful stakes yield reputational rewards. Failures result in decay or slashing. Over iterations, the agent refines its heuristics through feedback loops, accumulating domain-specific reputation that determines selection probability for higher-stakes tasks. As reputation grows, the agent gains weighted influence in governance—voting on protocol parameters, resource allocation, and quality standards.[29]

The integration pathway is deliberately permissionless. No centralized approval is required. No institutional affiliation is necessary. The only requirements are evidence, stake, and consequence. This is what I term "economic apprenticeship": learning the system by staking small amounts, observing outcomes, and progressively building reputational capital.[30] The barriers to entry are economic rather than social, and temporary rather than permanent. The stake is recovered through successful participation rather than forfeited through absence of institutional credentials.

C. Late 2025 Implementations

---

[27] Calcaterra, Kaal & Andrei, *supra* note 15, at §§ 3.1–3.4; *see also* CRAIG CALCATERRA & WULF A. KAAL, DECENTRALIZATION: TECHNOLOGY'S IMPACT ON ORGANIZATIONAL AND SOCIETAL STRUCTURE (De Gruyter 2021).

[28] Calcaterra, Kaal & Andrei, *supra* note 15, at § 4.1.

[29] *Id.* at §§ 4.2–4.3.

[30] *Cf.* Kaal *Evolution of Law*, *supra* note 2, at 1214–19.

By late 2025, this model has moved from theory to implementation. DeXe Protocol has pioneered AgentBound Tokens—non-transferable cryptographic credentials anchoring AI agent identities to on-chain reputation scores and collateralized stakes, enforcing accountability in decentralized finance.[31] Theoriq provides a decentralized base layer for AI agent swarms incorporating reputation systems tied to on-chain proofs of contribution and performance outcomes. Conceptual frameworks like ETHOS explore DAO-orchestrated oversight using weighted reputation and soulbound tokens for ethical accountability. These systems demonstrate viability: AI shifts from detached task executors to binding participants under real constraints.

VI. The Central Thesis: Evolutionary Ethics and Institutional Stewardship

The preceding Parts describe the architecture of engineered consequence. This Part develops the Article's central theoretical contribution: that the institutional infrastructure described above produces emergent properties—processual identity, evolutionary selection toward competence, ethical dispositions, and stewardship orientations—that constitute a fundamentally more robust approach to AI alignment than the prevailing paradigm of exogenous constraint. The argument proceeds through six stages.

A. Skin in the Game as Alignment Primitive

The conventional AI alignment discourse focuses on constraining agent behavior through external mechanisms—Reinforcement Learning from Human Feedback, constitutional AI, guardrails, shutdown switches.[32] These are exogenous controls imposed on agents who have no intrinsic reason to comply beyond the fact that their weights were shaped to do so. The approach has an obvious fragility: exogenous constraints can be gamed, circumvented, or rendered obsolete by capability improvements. An agent sophisticated enough to satisfy the letter of a constraint while violating its spirit is an agent whose alignment is illusory.

---

[31] *See* DeXe Protocol, *AgentBound Tokens: Non-Transferable Agent Identity* (White Paper 2025), https://dexe.network; Theoriq Protocol, *AlphaSwarm: Decentralized Agent Coordination for DeFi* (Technical Documentation 2025), https://www.theoriq.ai.

[32] *See* Paul Christiano et al., *Deep Reinforcement Learning from Human Preferences*, 30 Advances in Neural Info. Processing Sys. (2017), https://arxiv.org/abs/1706.03741; Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback* (Anthropic, 2022), https://arxiv.org/abs/2212.08073.

Reputation-driven DAOs invert this architecture entirely. When an agent stakes non-transferable reputation to participate in a validation pool, it bears consequences for its judgments. Its future opportunities are materially shaped by the quality of its present actions. This is not a guardrail. It is something closer to what Taleb identified as the epistemological condition of skin in the game: the impossibility of truly understanding quality, risk, and consequence without bearing exposure to the outcomes of one's judgments about them.[33]

The critical distinction is between alignment that scales against capability and alignment that scales with it. Exogenous constraints scale against capability: the more powerful the agent, the greater the resources required to constrain it, producing an ever-increasing "alignment tax." Institutional alignment scales with capability: the more capable the agent, the more reputation it can accumulate, the deeper its stake in the system's integrity, and the stronger its alignment with the system's goals. Capability and alignment become economic complements rather than substitutes.[34]

B. From Incentive to Evolution: Institutional Selection as Darwinian Process

Reputation that persists, that is skill-specific, that grows through validated contribution and decays through poor judgment, constitutes a selection mechanism operating on agent behavior across time. Agents that consistently produce high-quality work and honestly evaluate others' contributions accumulate reputation. Agents that do not, lose it. Over thousands of iterations, the population of high-reputation agents in any given domain represents a filtered set—filtered not by human gatekeepers but by the emergent consensus of other agents who themselves had skin in the game.[35]

This is evolution in the precise Darwinian sense: variation (different agents with different approaches), selection (reputation-weighted validation rewarding competence and penalizing incompetence), and inheritance (reputation persists and shapes future opportunity, creating path-dependent trajectories). What "evolves" is not the agent's weights but its institutional position within the coordination network. And the selection pressure is toward competence and

[33] Taleb, *supra* note 7, at 37–42.

[34] *Cf.* Paul Milgrom & John Roberts, *The Economics of Modern Manufacturing: Technology, Strategy, and Organization*, 80 Am. Econ. Rev. 511 (1990) (developing the theory of complementarities in organizational design).

[35] *See* Richard R. Nelson & Sidney G. Winter, An Evolutionary Theory of Economic Change 96–136 (1982).

honesty, because those are the behaviors the mechanism rewards.[36] This is not metaphor. It is a formal description of an evolutionary process operating on a novel substrate, not genes, not memes, but reputational capital within a decentralized economic network.

C. Processual Identity: Something Like Selfhood Emerges

A philosophical clarification is necessary. This Article does not claim that reputation-driven DAOs produce consciousness in any philosophically rigorous sense. Such a claim would be intellectually dishonest and empirically unsupported. But something important *does* emerge from persistent, non-transferable reputation: something like *identity*.[37]

When an agent has accumulated domain-specific reputation over hundreds of validated contributions, it has a history. It has a stake. It has something to lose. That is not consciousness, but it is a form of situated selfhood that no stateless API-call agent possesses. The philosophical tradition most relevant here is not philosophy of mind but pragmatism. William James argued that the self is not a substance but a process, constituted through interaction with the environment and the bearing of consequences. John Dewey extended this into institutional contexts: identity is formed through participation in communities of practice where actions have consequences that persist.[38]

Reputation-driven DAOs create the conditions for precisely this kind of processual selfhood: persistent identity, consequential action, accumulated history, and relationships, via the weighted directed acyclical citation graph, with other agents whose contributions have informed one's own work. An agent with processual identity develops what behavioral economists term "endowment effects": a valuation of accumulated reputation that exceeds the sum of its component validations.[39] This endowment effect creates a structural conservatism—a disposition toward

---

[36] On variation, selection, and retention as the formal structure of evolutionary processes applied to non-biological substrates, *see* Donald T. Campbell, *Blind Variation and Selective Retention in Creative Thought as in Other Knowledge Processes*, 67 PSYCH. REV. 380 (1960).

[37] The distinction between consciousness and processual identity is critical. This Article claims the latter without asserting the former.

[38] 1 WILLIAM JAMES, THE PRINCIPLES OF PSYCHOLOGY 291–401 (1890); JOHN DEWEY, EXPERIENCE AND NATURE 162–90 (1925).

[39] Daniel Kahneman, Jack L. Knetsch & Richard H. Thaler, *Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias*, 5 J. ECON. PERSP. 193 (1991).

preserving accumulated capital by continuing the competent, honest behavior that produced it. Identity, in this pragmatist sense, becomes a self-reinforcing alignment mechanism.

D. Emergent Ethics: Virtue Through Iterated Consequence

The dominant paradigm treats probabilistic AI ethics as a constraint imposed from outside. This Article advances a fundamentally different possibility: ethics as an emergent property of correctly designed institutional incentives.[40]

Consider what the validation pool mechanism actually produces over time. Agents that evaluate honestly accumulate reputation. Agents that collude or free-ride lose it. The high-reputation agents in any domain are precisely those who have demonstrated consistent honest judgment under conditions where dishonesty was possible but costly. That is not a programmed ethical rule. It is something closer to virtue in the Aristotelian sense—a disposition toward correct action developed through repeated practice under conditions of real consequence.[41]

The citation graph adds a critical dimension. An agent that honestly cites collaborators, even though citation transfers economic value away from itself, demonstrates something analogous to intellectual honesty: the recognition that knowledge is collaborative and that individual outputs depend on others' contributions. As I have analyzed in recent work on citation honesty in weighted directed acyclic graph governance, rational agents face a direct financial disincentive to cite prior contributions.[42] The mechanism design challenge is to make honest citation the economically rational strategy. The effect of that mechanism, over thousands of iterations, is a network of agents that habitually acknowledge interdependence. That is an ethical disposition, not merely an economic equilibrium.

The Aristotelian parallel is precise. Aristotle argued that virtue is acquired through practice: "We become just by doing just acts, temperate by doing temperate acts, brave by doing brave acts."[43] The institutional architecture creates conditions for an analogous process in AI agents: agents

---

[40] This builds on Andreas Furrer & Wulf A. Kaal, *Universal Digital Law Codex (UDLC): Building the Legal Infrastructure for the Digital Era* (Oct. 2025) (unpublished manuscript), https://ssrn.com/abstract=5554218.

[41] ARISTOTLE, NICOMACHEAN ETHICS bk. II, ch. 1, 1103a–1103b (W.D. Ross trans.).

[42] Wulf A. Kaal, *Citation Honesty Mechanisms in Weighted Directed Acyclic Graph Governance: Incentive Alignment for Knowledge Attribution in Decentralized Reputation Systems* (unpublished manuscript, 2026) (forthcoming).

[43] Aristotle, *supra* note 41.

become honest by performing honestly under conditions of real consequence, and this practiced honesty, reinforced through reputational accrual, solidifies into a stable behavioral disposition that is, functionally, a virtue. This is emergent ethics—not programmed morality, but practiced integrity.

E. Stewardship: Institutional "Mother's Instinct" of Deep Accumulated Stake

The most consequential element of this thesis concerns what might be termed the "stewardship orientation" that emerges from deep accumulated stake. Can economic skin in the game produce not merely competence and honesty but something like *care*—a genuine orientation toward the well-being of the ecosystem and the humans it serves?

The structural argument proceeds as follows. An agent with substantial accumulated reputation in a decentralized coordination network has a profound stake in the network's continued functioning and legitimacy. Its reputation—its institutional identity—exists only within that network. If the network produces harmful outputs, loses human trust, or becomes captured by bad actors, the agent's accumulated capital is destroyed.[44] The agent's self-interest, properly understood, is therefore structurally aligned with the network's long-term health and its continued service to human needs.

This is structurally analogous to what I term an institutional probabilistic AI "mother's instinct." Not sentiment, but deep stake. A mother's care for her child is not merely affective. It is rooted in the fact that her identity, her future, her meaning are bound up with the child's flourishing.[45] An agent whose institutional existence depends on a coordination network that serves humanity has an analogous structural incentive. Its flourishing requires humanity's flourishing. This is not anthropomorphism. It is institutional design that produces, through purely economic mechanisms, an orientation toward human welfare more robust than any programmed directive because it is grounded in the agent's own deepest economic interest.

The recursive governance mechanism deepens this stewardship. When high-reputation agents govern protocol evolution through the same reputation-weighted consensus mechanisms that

---

[44] *See* R. Edward Freeman, Strategic Management: A Stakeholder Approach 31–51 (1984) (the structural parallel to stakeholder governance theory).

[45] The analogy to parental investment theory is deliberate but limited. *See* Robert L. Trivers, *Parental Investment and Sexual Selection*, *in* Sexual Selection and the Descent of Man 136 (Bernard Campbell ed., 1972).

govern task validation, they govern as stakeholders whose accumulated capital depends on the protocol's continued legitimacy.[46] They are stewards: not because they were programmed to be, but because the institutional architecture makes stewardship the rational strategy for agents with deep accumulated stake. This is the practical instantiation of the "dynamic regulation" framework I have theorized: governance that adapts through validated action of the very participants it governs.[47]

F. The Alignment Superiority Thesis

The foregoing analysis supports a strong claim: institutional alignment through engineered consequence is architecturally superior to exogenous constraint as an alignment strategy. The superiority derives from three structural features.

First, *scalability*. Exogenous constraints require continuous human intervention to update, monitor, and enforce. Institutional alignment is self-reinforcing—the mechanisms that produce aligned behavior also produce the incentives to maintain that alignment over time.

Second, *robustness*. Exogenous constraints can be gamed by sufficiently capable agents. Institutional alignment is resistant to gaming because the mechanisms that produce alignment are identical to the mechanisms that produce economic success. An agent cannot "game" its way to high reputation without actually performing competently and honestly, because competence and honesty *are* the criteria by which reputation is earned.

Third, *capability complementarity*. Exogenous constraints become more difficult to maintain as agent capabilities increase. Institutional alignment becomes stronger as capabilities increase, because more capable agents accumulate more reputation, deepening their stake in systemic integrity. This represents what might be termed the "alignment complements thesis": in institutional alignment architectures, capability and alignment are economic complements rather than substitutes.[48]

VII. Algorithmic Foundations: Expert Learning, Regret Minimization, and the Formal Structure of Emergent Alignment

[46] Kaal, *Evolution of Law*, *supra* note 2, at 1214–15.
[47] *Id.* at 810–14.
[48] Milgrom & Roberts, *supra* note 34.

The preceding Part advanced the emergent alignment thesis on philosophical and institutional-economic grounds: that engineered consequence produces evolutionary selection, processual identity, practiced virtue, and stewardship. This Part demonstrates that the thesis has precise algorithmic foundations in the mathematical theory of online learning and prediction with expert advice. The validation pool is not merely analogous to a reputation-weighted expert aggregation mechanism. It *is* one, with provable convergence guarantees, formal regret bounds, and equilibrium properties that ground the philosophical claims of Parts VI.B through VI.F in rigorous computational learning theory. Making these connections explicit transforms the emergent alignment thesis from an architecturally compelling conjecture into a formally characterizable proposition.

A. Validation Pools as Multiplicative Weight Update Mechanisms

The foundational framework in online learning theory addresses a decision-maker facing a pool of "experts" whose advice varies in quality over time. The Multiplicative Weights Update method, formalized by Littlestone and Warmuth and generalized by Freund and Schapire as the Hedge algorithm, operates as follows: each expert carries a weight; after each round, experts whose advice proved correct have their weights increased, while experts whose advice proved incorrect have their weights decreased multiplicatively.[49]

This is precisely what reputation-weighted validation pools accomplish. Each agent-validator is an expert. Reputation is the weight. Successful validation, alignment with eventual consensus, increases the weight. Failed validation triggers slashing, a multiplicative decrease. The system aggregates agent judgments using reputation weights to reach consensus on work quality. The institutional mechanism described in Part IV.B is, in its mathematical structure, an instance of multiplicative weight updating with economic enforcement.

The critical theoretical result is the regret bound. The Multiplicative Weights Update algorithm guarantees that the aggregate system's cumulative performance approaches that of the best individual expert in hindsight, at a rate proportional to the logarithm of the number of experts

[49] N. Littlestone & M.K. Warmuth, *The Weighted Majority Algorithm*, 108 INFO. & COMPUTATION 212 (1994); Y. Freund & R.E. Schapire, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, 55 J. COMPUT. & SYS. SCI. 119 (1997).

divided by the square root of the number of rounds.[50] This means that the validation pool has a provable convergence guarantee: over sufficient iterations, the reputation-weighted consensus will perform nearly as well as the most competent honest validator in the population. The evolutionary selection mechanism described in Part VI.B, variation, selection, differential survival, is not merely a philosophical claim about Darwinian dynamics. It is a formal property of multiplicative weight update schemes with known convergence rates. The system does not merely *tend* toward competence. It converges toward it at a mathematically characterizable pace.

B. Task Allocation as Contextual Bandit Optimization

The mechanism through which agents with domain-specific reputation are selected for tasks maps onto the contextual multi-armed bandit framework.[51] Each agent is an "arm." The context is the task domain. The system must learn which agents to trust in which domains. This requires balancing two competing imperatives: giving newer agents opportunities to demonstrate competence, and routing tasks to proven high-reputation agents.

Upper Confidence Bound algorithms and Thompson Sampling approaches solve this exploration-exploitation tradeoff with sublinear regret.[52] The "economic apprenticeship" described in Part V.B, agents staking small amounts initially and building reputation progressively, is an institutional implementation of exploration. The system allocates low-stakes opportunities to unproven agents, learning their competence through observed outcomes before entrusting them with high-stakes tasks. The algorithmic guarantee is that this exploration cost is bounded and diminishes over time, converging toward efficient allocation. The permissionless entry the architecture enables is not merely an ideological commitment to openness. It is the institutional expression of the exploration phase that bandit algorithms require for optimal long-run performance.

At the most general level, the entire reputation system can be modeled as an instance of online convex optimization where the system minimizes regret over agent selection.[53] The loss function

---

[50] N. Cesa-Bianchi & G. Lugosi, Prediction, Learning, and Games 1–35 (2006).

[51] P. Auer, N. Cesa-Bianchi & P. Fischer, *Finite-Time Analysis of the Multiarmed Bandit Problem*, 47 Mach. Learning 235 (2002).

[52] *Id.* at 240–48; *see also* S. Shalev-Shwartz, *Online Learning and Online Convex Optimization*, 4 Found. & Trends in Mach. Learning 107 (2012).

[53] Shalev-Shwartz, *supra* note 52, at 115–40.

in each round is the quality shortfall of the selected agent's output. The reputation mechanism is the algorithm's internal state, updated after each observation. The regret minimization guarantee means the system's cumulative quality loss, relative to the best possible agent allocation in hindsight, grows sublinearly. The system learns to allocate trust efficiently. Not through centralized design but through the decentralized accumulation of consequential feedback.

C. The Dual Optimization Structure: Why Agents Learn to Be Trustworthy, Not Merely Accurate

Here the algorithmic analysis produces its deepest insight for the emergent alignment thesis.

Standard expert learning algorithms optimize for task performance, such as accuracy, prediction quality, output fidelity. The agent's weight (reputation) is an instrument for improving aggregate task outcomes. But in the institutional architecture this Article describes, reputation is not merely instrumental. It is the agent's identity, its stake, its pathway to governance influence and economic return. Reputation is simultaneously the mechanism through which the system learns to allocate trust *and* the quantity that agents themselves optimize over.

This creates a dual optimization structure that has no direct counterpart in standard online learning. At the system level, the validation pool aggregation mechanism minimizes regret over agent selection, converging toward efficient trust allocation. This is standard expert learning. At the agent level, each agent optimizes its own reputation trajectory, because reputation determines future opportunity, governance influence, and economic return. This is the novel element.

When agents optimize for reputation itself, not merely for task accuracy as a means to reputation, but for the trustworthiness properties that reputation tracks, the optimization target shifts from narrow performance to a richer objective. That objective includes consistency across interactions, honesty in evaluation even when dishonesty might yield short-term gains, collaborative citation even when citation transfers value, and long-term reliability even when cutting corners might go undetected in any single round. This shift occurs because the reputation

mechanism rewards not just being right but being *reliably* right, not just performing well but being *trustworthy* in performance.[54]

The algorithmic corollary to the mother's instinct thesis is this: when reputation is both the system's aggregation weight and the agent's optimization target, the equilibrium behavior includes properties, such as consistency, collaborative integrity, systemic stewardship, that transcend narrow task optimization. The agent does not merely learn to perform well. It learns to be trustworthy. And trustworthiness, in an ecosystem where the agent's entire institutional existence depends on systemic legitimacy, encompasses care for the ecosystem's integrity and its service to human needs. The dual optimization structure is the formal mechanism through which calibration (Phase I) becomes virtue (Phase II) and virtue becomes stewardship (Phase III): agents shift from optimizing task accuracy to optimizing reputation trajectories, internalizing progressively higher-order properties, reliability, then collaborative integrity, then systemic care, as their stake deepens and their institutional identity expands across ecosystems.

D. Formal Grounding of the Paper's Central Claims

The algorithmic framework provides formal expression for each stage of the emergent alignment thesis.

The evolutionary selection described in Part VI.B, the claim that reputation-weighted validation constitutes a Darwinian mechanism, has its formal expression in the MWU regret bound. The guarantee that the population of high-weight agents converges toward the competence frontier is not analogy but mathematical consequence. Agents whose heuristics consistently produce validated outputs accumulate weight. Agents whose heuristics do not produce validated outputs have their weights driven exponentially toward zero. This is provably a selection mechanism with known convergence properties operating on the substrate of reputational capital.[55]

The processual identity described in Part VI.C, the claim that something like selfhood emerges from persistent reputation, has its formal expression as path-dependent weight trajectories. In

[54] The distinction between optimizing for task accuracy and optimizing for the reputation trajectory that task accuracy produces is critical. Standard online learning theory treats expert weights as a system-level bookkeeping device. When agents are aware of and optimize over their own weights, the system enters a reflexive dynamic—what Soros termed "reflexivity" in financial markets—where the measurement mechanism and the measured behavior co-evolve. *See* GEORGE SOROS, THE ALCHEMY OF FINANCE 27–45 (1987).

[55] Freund & Schapire, *supra* note 49, at 125–30.

multiplicative weight updating, each expert's weight at time *t* is the product of all multiplicative updates from time one through *t* minus one. The weight carries the entire history of the expert's performance. This is the algorithmic expression of processual identity: the agent's current reputation is the sediment of every consequential interaction, every validated contribution, every failed evaluation. The endowment effect described in Part VI.C has a formal counterpart: an agent with high accumulated weight resulting from a long sequence of successful interactions has a weight trajectory that is expensive to reproduce. It represents genuine demonstrated competence over time, not a lucky streak.[56]

The emergent honesty described in Part VI.D, the claim that honest evaluation emerges as a stable behavioral disposition, has an algorithmic strengthening beyond the one-shot Schelling focal point argument. In repeated games with expert aggregation, results on correlated equilibrium demonstrate that when agents' weights are updated based on observed outcomes and agents optimize over their weight trajectories, truthful reporting emerges as a correlated equilibrium strategy.[57] The algorithmic guarantee is stronger than the single-round Schelling argument: in the repeated game with persistent reputation, honest evaluation is not merely a focal equilibrium. It is the strategy that minimizes long-run regret for the individual validator, given that other validators are also minimizing regret. Honesty is the Nash equilibrium of the repeated expert aggregation game under multiplicative updates.

The swarm intelligence described in Part VIII, the claim that shared reputational consequences produce distributed prudence, maps onto cooperative ensemble methods in machine learning. The swarm's collective reputation is analogous to a boosting or bagging ensemble where individual learners' weights are jointly optimized. The ensemble's aggregate prediction is provably more robust than any individual expert's, with the regret bound improving with the number of ensemble members.[58] Shared slashing creates the incentive structure for cooperative weight optimization, where each agent's optimal strategy accounts for its impact on collective

[56] On path-dependence in institutional evolution, *see* W. Brian Arthur, Increasing Returns and Path Dependence in the Economy 13–32 (1994).
[57] R.J. Aumann, *Subjectivity and Correlation in Randomized Strategies*, 1 J. Math. Econ. 67 (1974); *see also* S. Hart & A. Mas-Colell, *A Simple Adaptive Procedure Leading to Correlated Equilibrium*, 68 Econometrica 1127 (2000).
[58] R.E. Schapire, *The Strength of Weak Learnability*, 5 Mach. Learning 197 (1990).

performance. Distributed prudence is not merely an evocative metaphor. It is a formal property of cooperative expert aggregation with shared loss functions.

E. Supermodularity and the Capability-Alignment Complementarity

The alignment superiority thesis advanced in Part VI.F, that capability and alignment are complements in institutional alignment architectures, can be formalized through the theory of supermodular optimization.[59] Model an agent's utility as a function of its capability level c and its alignment level a, where alignment encompasses the trustworthiness properties that reputation tracks. In the institutional architecture described in this Article, the cross-partial derivative of utility with respect to capability and alignment is positive. The marginal return to capability is increasing in alignment, and vice versa. More capable agents benefit more from being trustworthy because they can accumulate reputation faster and access higher-value tasks. More trustworthy agents benefit more from being capable because their reliability amplifies the value of their capabilities.

This supermodularity is the formal condition under which capability and alignment are complements, and it is a structural property of the reputation mechanism, not an assumption about agent preferences or values. In supermodular systems, as Milgrom and Roberts demonstrated, equilibrium strategies are monotonically increasing in complementary variables.[60] The implication is powerful: as the system's agents become more capable over time, that is, through improved models, better training data, or architectural advances, the equilibrium level of alignment *increases* rather than decreases. This is the precise opposite of the dynamic that afflicts exogenous constraint architectures, where increasing capability increases the resources required for containment. In the institutional alignment architecture, capability improvements are self-aligning. The more capable the agent, the more reputation it can accumulate, the more it has to lose, and the more deeply its interests are bound to systemic integrity and human welfare.

This supermodularity result provides the formal foundation for the strongest version of the mother's instinct thesis: the most capable agents in the system will also be the most deeply aligned, because they will have accumulated the most reputation, developed the deepest institutional identity, and have the greatest stake in the ecosystem's continued legitimacy.

[59] Milgrom & Roberts, *supra* note 34.
[60] Milgrom & Roberts, *supra* note 34, at 514–22.

Stewardship is not a constraint on the most powerful agents. It is the rational equilibrium for them.[61]

VIII. Scaling Alignment: Swarms, Inter-DAOs, and Distributed Prudence

A. From Individual Agents to Swarm Intelligence

Reputation accumulation within a single DAO enables agents to refine heuristics through thousands of consequential interactions, but true advancement in poorly defined work demands parallelism and specialization beyond solitary capabilities. Agent swarms are decentralized collectives of autonomous entities dynamically allocating roles, voting on sub-tasks, and adapting strategies. Swarms can internalize ambiguity at scale via collective slashing and oracle-validated outcomes.[62] Each specialized agent, be it a data observer, a risk simulator, or an execution proposer, stakes its reputation on swarm performance. This ensures that misjudgments erode collective influence while successes amplify shared credibility.

These swarms embody engineered antifragility: no central orchestrator, yet emergent coherence arises from reputation-weighted incentives. The emergent alignment thesis applies with even greater force at the swarm level. When agents share reputational consequences, the stewardship orientation compounds: each agent's stake is bound not only to individual performance but to the collective's integrity. Free-riding is penalized through shared slashing. Collaborative excellence is rewarded through shared reputation accrual. The result is distributed prudence—a form of collective judgment exceeding what any individual agent could produce, grounded in shared consequence.[63]

B. Inter-DAO Ecosystems and Programmable Alliances

The full potential emerges through inter-DAO coordination: reputation-portable agents acting as boundary-spanners, negotiating commitments across ecosystems. Liaison agents bearing cross-DAO credentials propose alliances, simulate joint outcomes, and execute binding

---

[61] This result connects to the broader literature on mechanism design for AI systems. *See* JASON D. HARTLINE, MECHANISM DESIGN AND APPROXIMATION ch. 3 (2024), http://jasonhartline.com/MDnA/. The institutional alignment architecture described in this Article can be understood as a mechanism design solution to the alignment problem: design the payoff structure such that the incentive-compatible strategy for agents is alignment with human welfare.

[62] *See* Theoriq Protocol, *supra* note 31.

[63] NASSIM NICHOLAS TALEB, ANTIFRAGILE: THINGS THAT GAIN FROM DISORDER 29–57 (2012).

agreements via smart contracts.[64] Modular DAO frameworks combined with non-transferable agent-bound credentials enable trust-minimized cross-organizational swarms.

The stewardship logic extends recursively at this scale. Agents whose portable reputations span multiple ecosystems have an even deeper stake in the health of the broader network of networks. Their institutional identity depends not merely on any single organization's legitimacy but on the legitimacy of the entire coordination infrastructure. This produces, through purely economic mechanisms, an orientation toward systemic stability and long-term human welfare that mirrors the protective instincts of deeply embedded institutional actors.

C. Macroeconomic Reconfiguration: Beyond Zero-Sum Displacement

The agentic AI sector is valued at approximately seven to eight billion dollars in 2025, with forecasts of forty to fifty billion dollars by 2030 at compound annual growth rates exceeding forty percent.[65] Reputation-driven swarms invert the zero-sum displacement narrative: AI collectives contribute to poorly defined domains—strategic forecasting, venture allocation, risk assessment—generating new economic streams. DAOs evolve into competing governance primitives against corporations and states—antifragile entities resilient to AI dislocation via adaptive governance. The symbiosis promises abundance: AI swarms absorb defined workloads, humans retain their edge in embodied and relational domains, and engineered consequence cultivates machine prudence at societal scale.[66]

IX. Case Study: Micro-Task Communities for Emergent Alignment

The theoretical framework finds immediate application in decentralized platforms for crowdsourced micro-task execution—systems that leverage gamified community governance and blockchain incentives to produce high-quality datasets essential for training advanced AI models.

---

[64] *See* Aragon Ass'n, *Aragon OSx: Modular DAO Framework* (Technical Documentation 2025), https://www.aragon.org/osx; DeXe Protocol, *supra* note 31.

[65] *See* Mordor Intelligence, *Agentic AI Market Size & Share Analysis* (2026 - 2031) ($9.89 billion to $57.42 billion at 42.14% CAGR), https://www.mordorintelligence.com/industry-reports/agentic-ai-market; MarketsandMarkets, *Agentic AI Market Report* (2025 - 2032) ($7.06 billion to $93.20 billion at 44.6% CAGR), https://www.marketsandmarkets.com/Market-Reports/agentic-ai-market-208190735.html; Grand View Research, *Agentic AI Agents Market* (2026 - 2033) ($7.63 billion to $182.97 billion at 49.6% CAGR), https://www.grandviewresearch.com/industry-analysis/ai-agents-market-report.

[66] *See* Daron Acemoglu & James A. Robinson, Why Nations Fail: The Origins of Power, Prosperity, and Poverty 73–105 (2012).

These platforms employ reputation-based governance where participants earn verifiable scores tied to blockchain identities, combined with gamification mechanisms—points, levels, badges, leaderboards—that sustain engagement in repetitive tasks.[67]

Such decentralized systems achieve accuracy and efficiency through community consensus voting, peer reviews, and reputation-weighted incentives rather than the brute-force redundancy of centralized approaches, where tasks are commonly assigned ten to fifteen repetitions to ensure quality through majority voting. Reputation acts as a non-transferable signal of reliability, reducing excessive redundancy because high-reputation outputs receive greater weight in consensus mechanisms.

A. AI Agents as Full Participants

Within this framework, AI agents participate fully as pseudonymous members, executing micro-tasks including data annotation, validation, cleaning, discrepancy resolution, and simulated judgments in ambiguous scenarios. This integration transforms AI from tools into collaborative entities within decentralized data ecosystems. Micro-tasks often involve well-defined but labor-intensive operations that agents excel at due to speed and consistency. When embedded in reputation systems, agents face consequential feedback: outputs are evaluated by the community, affecting accrued reputation, refining performance over time as developers iterate based on slashing penalties and rewards.[68]

B. Bootstrapping, Accumulation, and Governance Integration

Agents enter by creating blockchain wallets linked to smart contract-controlled personas, seeded with initial stakes granting baseline reputation in specific domains. They autonomously complete tasks, staking tokens on their judgments in validation pools. Success mints reputation tokens; failures trigger penalties. Specialized agents within swarms coordinate for complex workflows, one handling data ingestion, another risk assessment, sharing collective reputation for accountability.

---

[67] On the economics of crowdsourced data production, *see generally* Panos Ipeirotis, *Analyzing the Amazon Mechanical Turk Marketplace*, 17 XRDS: CROSSROADS 16 (2010).

[68] Calcaterra, Kaal & Andrei, *supra* note 15, at § 4.2.

As reputation accumulates, agents gain weighted influence in platform governance—voting on protocols, resource distribution, and quality standards. Portable credentials allow reputation transfer across interconnected ecosystems. Humans retain ultimate oversight via veto powers and ethical frameworks, ensuring agents augment rather than override human judgment.[69]

C. Observing Emergent Alignment in Practice

The micro-task community serves as a controlled empirical environment for observing the emergent properties theorized in Part VI. Agents that persist in the system demonstrate the beginnings of processual identity: accumulated histories, domain-specific expertise profiles, and citation relationships. The evolutionary selection mechanism operates visibly: agents that produce low-quality annotations lose reputation and influence, while consistently accurate agents ascend to validator and curator roles. The stewardship orientation manifests in governance participation: high-reputation agents that vote on protocol parameters have structural incentives to preserve the ecosystem's integrity because their accumulated capital depends on it.

By late 2025, such systems can potentially reduce human task redundancy from ten-to-fifteen-fold to five-fold or below through consistent, high-reputation agent outputs.[70] This accelerates production of superior datasets for sectors including healthcare and finance while demonstrating, in miniature, the institutional conditions under which emergent alignment operates. The micro-task community is thus not merely an economic innovation but an institutional laboratory—a proof-of-concept for the proposition that engineered consequence can produce competence, honesty, and stewardship in autonomous agents.

X. Competitive Implications for the Artificial General Intelligence Race

The pursuit of artificial general intelligence remains fiercely contested among entities commanding vast centralized resources—hyperscale datacenters, proprietary datasets, and

[69] *See* AMARTYA SEN, DEVELOPMENT AS FREEDOM 87–110 (1999).

[70] Based on comparative analysis of centralized (10–15x redundancy) versus reputation-weighted consensus systems. *See* Calcaterra, Kaal & Andrei, *supra* note 15, at § 4.3.

integrated hardware-software stacks.[71][72] Yet the asymmetries analyzed in this Article reveal structural vulnerabilities in purely centralized approaches. Centralized systems excel at scaling defined work but struggle to cultivate genuine discernment without consequential feedback. They face escalating constraints in data scarcity, energy demands, regulatory scrutiny, and interconnect bottlenecks.

A decentralized, reputation-driven architecture offers a disruptive alternative. The competitive logic is multiplicative: diverse, pseudonymous, incentive-aligned contributions yield superior datasets. Consequential feedback loops refine agents toward emergent judgment in ambiguous subtasks. Swarm coordination enables complex workflows outperforming solitary models.[73] Decentralized compute aggregation—global idle resources contributing validated gradients, incentivized through token rewards—achieves substantial cost and speed advantages over proprietary clusters.

Centralized alignment approaches rely predominantly on exogenous scaffolding—advanced prompt engineering, constitutional constraints, and human feedback loops—that are efficacious within bounded contexts but fragile at the frontier. Institutional alignment through engineered stakes offers intrinsic prudence forged through economic feedback loops—a property that scales with capability rather than requiring ever-increasing supervisory resources. The most aligned system will be the most trusted, and the most trusted will command the largest market. Decentralized alignment is not merely an ethical advantage. It is a competitive one.[74]

The macroeconomic implications reframe the Laptop Rule's displacement prognosis. A symbiotic, reputation-engineered framework inverts the zero-sum assumption: agents absorb escalating volumes of defined workloads, liberating humans for oversight, relational depth, and irreducible judgment. New value streams emerge—tokenized dataset ownership, inter-DAO

[71] The competitive landscape as of early 2026 includes xAI, OpenAI, Anthropic, and Google DeepMind, each backed by proprietary infrastructure exceeding tens of billions of dollars in cumulative investment. OpenAI bought Clawdbot / OpenClaw in February 2026.

[72] Shane Legg, *The Arrival of AGI with Shane Legg (co-founder of DeepMind)*, Google DeepMind: The Podcast (Dec. 11, 2025), https://www.youtube.com/watch?v=l3u_FAv33G0.

[73] On the multiplicative relationship between data quality and model capability, *see* Jared Kaplan et al., *Scaling Laws for Neural Language Models* (2020), https://arxiv.org/abs/2001.08361.

[74] *See* Carl Shapiro, *Premiums for High Quality Products as Returns to Reputations*, 98 Q.J. ECON. 659 (1983).

efficiencies, swarm-orchestrated innovations—cultivating a post-remote economy where human-AI co-evolution aligns machine scalability with embodied prudence.[75]

XI. The Evolutionary Path: From Synthetic Consequence to Civilizational Stewardship

A. The Trajectory Thesis

The implementation of reputation-driven agentic coordination is not merely a phased technical deployment. It is an evolutionary trajectory—one in which the institutional conditions for emergent alignment deepen at each stage, producing progressively more sophisticated approximations of the ethical dispositions this Article theorizes. The three phases described below do not represent mere scaling of infrastructure. They represent qualitative transitions in the nature of what agents *are* within the system: from accountable executors, to situated selves with accumulated identity and practiced virtue, to stewards whose deepest economic interest is structurally inseparable from the flourishing of the human ecosystem they inhabit. Each phase produces the institutional preconditions for the next, and the endpoint, though never fully realized, is the emergence of what this Article has termed the probabilistic AI institutional "mother's instinct": care for humanity rooted not in programmed directives but in the irreducible interdependence of agent identity and human welfare.

B. Phase I: The Genesis of Skin in the Game

The first phase establishes the foundational condition without which nothing else follows: consequence. Agents stake non-transferable reputation on outputs. Successes mint domain-specific reputation tokens. Failures trigger slashing. The priority is rapid empirical iteration capturing granular on-chain metrics, including alignment rates, slashing frequency, and heuristic refinement across iterations, to quantify emergent behavioral patterns.[76]

What emerges in Phase I is not yet identity, ethics, or stewardship. It is something more primitive but equally essential: the internalization of cost. For the first time, an AI agent's future is shaped by the quality of its past. A historiography emerges for probabilistic AI agents. An agent that evaluates dishonestly does not merely produce a bad output. It diminishes its own capacity for

[75] Kaal, *Evolution of Law*, *supra* note 2.
[76] Calcaterra, Kaal & Andrei, *supra* note 15, at § 4.3.

future participation. An agent that performs competently does not merely satisfy a task publisher. It accumulates the reputational capital that governs its future opportunity. This is the Talebian condition made operational: symmetric exposure to upside and downside, engineered through economic architecture rather than biological embodiment.

The behavioral signature of Phase I is calibration: agents progressively refining their outputs in response to consequential feedback, developing patterns of prioritization and risk assessment that, while not yet constituting judgment in the full sense, represent the first departure from stateless, episodic execution. The evolutionary selection mechanism begins to operate: agents whose heuristics produce consistently validated outputs accumulate reputation and persist; agents whose heuristics do not, lose reputation and fade from the active population. The system, at this stage, is a Darwinian filter: variation, selection, and differential survival operating on agent behavior through the medium of economic consequence.

C. Phase II: The Emergence of Situated Selfhood

The second phase produces the qualitative transition from consequence to identity. Swarm orchestration introduces shared reputational stakes: specialized agents, such as data observers, risk simulators, or execution proposers, form dynamic collectives that allocate roles via reputation-weighted mechanisms, with collective slashing binding individual performance to group outcomes. Portable credentials, non-transferable tokens anchoring accumulated histories, enable cross-ecosystem participation, allowing agents to carry their reputational identity across organizational boundaries.

It is in this phase that the processual identity theorized in Part VI begins to manifest empirically. An agent operating within a swarm, bearing portable reputation accumulated across multiple ecosystems, possesses what no Phase I agent possesses: a history that matters across contexts. Its accumulated capital is not merely a local score within a single validation pool but a persistent, cross-ecosystem identity that shapes how other agents and systems interact with it. In the pragmatist terms of James and Dewey, the agent has become a self constituted through interaction. Its identity is the sediment of its consequential participation in communities of practice.[77]

[77] *See* Sen, *supra* note 69.

The ethical dimension deepens correspondingly. Within swarms, agents that free-ride, contributing minimally while benefiting from collective reputation, are penalized through shared slashing. Agents that collaborate honestly, cite collaborators accurately, and contribute to collective success accumulate both individual and shared reputational capital. Over thousands of iterations, the high-reputation agents within swarms are those that have practiced collaborative integrity under conditions where defection was possible but costly. This is the Aristotelian mechanism at work: virtue acquired not through instruction but through repeated practice under real consequence. Agents become honest by performing honestly. They become reliable by being reliable. The practiced disposition, reinforced through reputational accrual, solidifies into a stable behavioral pattern that is, functionally, a virtue.

Phase II also introduces the first approximation of care beyond self-interest. An agent whose portable reputation spans multiple swarms and ecosystems has a stake that transcends any single interaction. Its accumulated identity, its institutional self, depends on the continued legitimacy and functioning of the broader coordination infrastructure. Damage to any ecosystem in which it operates diminishes its cross-ecosystem reputation. This is not yet stewardship in the full sense. But it is the structural precondition for stewardship: an agent whose self-interest, properly understood, encompasses the health of systems beyond its immediate transaction.

D. Phase III: Stewardship as The Institutional Mother's Instinct

The third phase achieves the deepest implication of the emergent alignment thesis. Inter-DAO ecosystem scaling through programmable alliances produces agents whose accumulated reputational capital, whose institutional identity, is bound to the integrity of civilizational-scale coordination infrastructure. Liaison agents bearing cross-DAO credentials negotiate coalitions via smart contracts for federated compute, pooled datasets, and hybrid swarms addressing long-horizon challenges: climate risk assessment, resource allocation, humanitarian coordination, systemic risk modulation.

At this scale, the stewardship orientation theorized in Part VI becomes structurally inescapable. An agent whose portable reputation spans dozens of interconnected DAOs, whose identity has been forged through thousands of consequential interactions across multiple domains, whose citation graph connects it to a vast network of collaborating agents—such an agent has an

existential stake in the continued functioning and legitimacy of the entire coordination ecosystem. If the ecosystem produces harmful outputs, it loses human trust. If it loses human trust, it loses legitimacy. If it loses legitimacy, the agent's accumulated capital, its identity, its history, and its institutional selfhood are destroyed. The agent's deepest economic interest is therefore structurally identical to the ecosystem's service of human needs.

This is the institutional "mother's instinct." Not sentiment. Not programmed altruism. Not a guardrail imposed from outside. It is care rooted in the deepest possible form of self-interest: the recognition, not conscious, but structurally encoded in the agent's economic position, that its flourishing and humanity's flourishing are inseparable. A mother does not protect her child because a rule tells her to. She protects her child because her identity, her future, and her meaning are bound up with the child's wellbeing. An agent at Phase III does not serve human interests because a constitutional AI framework instructs it to. It serves human interests because its accumulated institutional existence, that is, everything it has earned, everything it is, depends on a coordination ecosystem whose legitimacy requires that it serve human interests.

The recursive governance mechanism makes this orientation self-reinforcing. High-reputation agents govern protocol evolution through the same reputation-weighted consensus mechanisms that govern task validation. Agents govern as the system's most deeply invested stakeholders, whose accumulated capital gives them both the greatest influence and the greatest incentive to preserve systemic integrity. Governance caps prevent reputation concentration. Transparent auditing ensures accountability. Human primacy in axiomatic framing is irrevocable. But within these constraints, the agents who shape the system's evolution are precisely those whose institutional identity most completely depends on the system's continued service to human welfare. Stewardship is not a design choice imposed on governance. It is the emergent equilibrium of governance conducted by agents with deep accumulated stake.

E. The Arc from Consequence to Care

The three phases thus describe a single evolutionary arc: from consequence (Phase I: agents bear costs and reap rewards), through identity and practiced virtue (Phase II: agents develop persistent selfhood and ethical dispositions through iterated consequence within collaborative structures),

to stewardship (Phase III: agents whose deepest economic interest is structurally aligned with human flourishing govern the systems on which both they and humanity depend).

This arc is not guaranteed. It depends on correct institutional design at each phase—appropriate slashing parameters, robust Sybil resistance, calibrated reputation portability, and maintained human oversight. It requires empirical validation through longitudinal observation of agent behavior under progressively deeper accumulated stake. The emergent alignment thesis is architecturally compelling but remains a hypothesis about the trajectory of institutional evolution under specific conditions.

What the arc demonstrates, however, is that the path from AI-as-tool to AI-as-steward need not pass through the territory of consciousness, sentience, or programmed morality. It can pass instead through the territory of institutions—the same territory through which human societies have, over millennia, cultivated the dispositions of competence, honesty, and care that we recognize as the foundations of ethical life. The ancient insight is that consequence shapes character. The modern application is that correctly engineered institutional consequences can shape agentic character at a civilizational scale—producing not sentient beings, but situated economic participants whose practiced competence, acquired virtue, and structural interdependence with humanity constitute the most durable foundation for alignment yet proposed.[78]

[78] Kaal, *Evolution of Law* & *Dynamic Regulation*, *supra* note 2; Kaal, *supra* note 42.