Governance as a Product: A Reputation-Weighted Institutional Architecture for Autonomous AI Agent Governance

Wulf A. Kaal
School of Law
University of St. Thomas
Minneapolis, MN, USA
ORCID: 0000-0003-0757-275X

***Abstract*—Autonomous artificial intelligence agents already act consequentially in finance, medicine, and critical infrastructure, yet the governance that binds them remains either informal and unenforceable or formal and institutionally inoperable. This paper argues that human legal governance fails for two interacting reasons: natural language cannot fully describe the situations governance must cover, and human institutions reason about the rules they do state unreliably. We present Governance as a Product (GaaP), a callable and composable governance service that addresses both failures through a reputation-weighted institutional architecture rather than a restricted formal logic. GaaP composes four interdependent layers: a weighted directed acyclic graph substrate with non-transferable reputation staking; a translation and validation layer that pairs typed governance primitives and ontology-guided composition with reputation-staked validation pools; deterministic cryptoeconomic enforcement; and a temporal layer in which the system governs its own evolution. The architecture trades a formal completeness guarantee for full expressiveness and institutional operability, reducing reasoning error to a bounded and auditable residual. We position GaaP as durable and adaptive infrastructure for the sustainable governance of autonomous systems, able to evolve without ossifying and so to remain valid as technology and norms shift.**

***Index Terms*—AI governance, autonomous agents, reputation systems, decentralized autonomous organizations, weighted directed acyclic graph, mechanism design, dynamic regulation, sustainable governance**

I. Introduction

Autonomous AI agents are not a future prospect but a present institutional fact. They execute trades at sub-millisecond latency, generate medical recommendations, and manage components of critical infrastructure. When such agents make consequential decisions, governance is required at the speed and scale at which the agents themselves operate.

The available responses are structurally inadequate in opposite ways. Informal, deliberative, and discretionary governance cannot bind autonomous agents quickly or at scale, and it provides no mechanical means of enforcement. Pure formal verification is technically sound but institutionally inoperable: it demands specialist expertise that excludes nearly every governance stakeholder who is not a logician, so the resulting infrastructure is formally capable yet practically inert.

This paper proposes a third path. Governance as a Product (GaaP) is a callable and composable governance service that addresses the two underlying failures of human governance without restricting itself to a decidable logic. Its foundation is the Calcaterra-Kaal weighted directed acyclic graph (WDAG), an institutional substrate of reputation-weighted validation pools with non-transferable reputation staking [1]–[3].

The contributions are four. First, a diagnosis of human governance as the conjunction of two interacting failures. Second, a four-layer architecture that grounds enforceable governance in the WDAG rather than in a formal language. Third, an explicit statement of the central design tradeoff, expressiveness against enforceability, together with the mechanisms that bound the resulting error. Fourth, the positioning of GaaP as sustainable governance infrastructure: governance that adapts without ossifying remains valid over time, which is the institutional analog of sustainability and a direct response to the regulatory pacing problem [4].

II. Background and Related Work

A. Dynamic regulation and the pacing problem

A long line of work establishes that reactive law has a diminishing capacity to track innovation, because regulation responds to facts after they form while technology advances continuously [4]. Dynamic and anticipatory regulation proposes feedback-driven instruments in response. GaaP extends that program from the regulation of systems to a reusable governance service.

B. Blockchain and DAO governance

Decentralized autonomous organizations demonstrated that governance rules can be encoded, executed, and iterated on chain rather than through institutional lag [5]. Empirically, however, DAO governance exhibits a persistent institutional deficit. Token-weighted voting concentrates control, participation is thin, treasuries are undercapitalized against the decisions they secure, and flash-loan style attacks exploit the coupling of voting power to transferable tokens. These findings motivate the separation of non-transferable reputation from transferable collateral that the GaaP substrate adopts.

TABLE I
SELECTED INSTITUTIONAL FINDINGS ACROSS FORTY OPERATIONAL DAOS [6]

Institutional metric	Result
Cohort mean, thirteen-category rubric	67.3 / 130 (51.8%)
Agent and alignment governance	2.10 / 10
Decentralization vs. token plutocracy	5.17 / 10
Production reputation-weighted aggregation	0 of 40
Projected mean after graph-centered upgrades	95.3 / 130

A systematic evaluation of forty operational DAOs across eight industry segments quantifies this deficit [6]. Measured against a thirteen-category institutional rubric, the cohort implements only about half of the prescribed architecture, with a mean of 67.3 of 130 points (51.8 percent). Two results bear directly on the governance of AI agents. The category measuring agent and alignment governance scores 2.10 of 10 across the dataset, and no entity crosses the midpoint on it, the only category for which that holds. Decentralization scores 5.17 of 10, and no DAO in the cohort runs reputation-weighted on-chain aggregation in production; token-weighted voting remains the default form. Table I summarizes the findings most relevant here. The same study projects that a small set of upgrades centered on reputation-weighted graph historiography and a separation of powers would raise the cohort mean to 95.3 of 130. GaaP can be read as an implementation of that agenda for the AI-agent setting: a reputation-weighted substrate supplying the agent-governance layer the cohort lacks.

C. Formal verification and legal translation

Formal methods for multi-agent systems, including alternating-time temporal logic, strategy logic, and their information-asymmetric and open-system extensions, provide rigorous guarantees within bounded settings but at substantial expressiveness and expertise cost [9], [10]. A parallel literature on legal argumentation and neuro-symbolic legal AI studies the translation from informal norms to executable form [11], [12]. GaaP draws the lesson that no governance layer succeeds unless it is simultaneously enforceable and accessible to non-specialist stakeholders, and it treats that joint requirement as the design target.

III. TWO IRREDUCIBLE FAILURES OF HUMAN GOVERNANCE

Human and institutional governance already attempts formal logical reasoning. Statutes are said to imply requirements, principles to constrain authority, and decisions to resolve contradictions. Governance nonetheless fails systematically at the formal level, for two distinct reasons.

The first failure is the inability of language to describe reality. However precisely a specification is drafted, there will always be situations it does not cover and circumstances its language cannot capture. This is not a drafting defect but a structural limit of natural language.

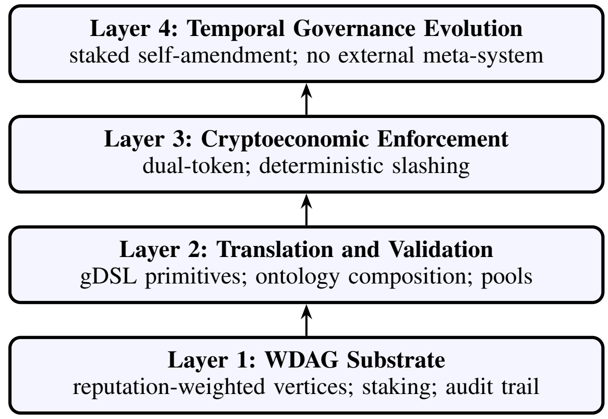

Fig. 1. The four-layer GaaP architecture. The WDAG substrate is the institutional foundation; each higher layer operates on the layer beneath it.

The second failure is the unreliability of logical reasoning. Even within the situations a specification does describe, human institutions infer unreliably: they draw conclusions from inconsistent premises, apply rules inconsistently across cases, and miss implications of the rules they hold.

The two failures interact. The choice of representation determines the scope of reliable reasoning. A more expressive representation can describe more of reality but is harder to reason about; a less expressive representation enables reliable reasoning over a narrower domain. Any governance design must therefore choose where on this frontier to sit, and that choice is the pivotal decision the architecture below makes explicit.

IV. THE GAAP ARCHITECTURE

GaaP composes four interdependent layers, illustrated in Fig. 1. Each layer is independently necessary and none is individually sufficient.

A. Layer One: the WDAG substrate

The foundation is the Calcaterra-Kaal weighted directed acyclic graph [1], [2]. Every governance decision is a vertex, edges are weighted by validated reputation, and participants stake non-transferable reputation on their judgments. Because the graph is acyclic and append-only, it provides a continuous and verifiable audit trail: every decision is traceable to the evidence and validations that produced it. This trail is the institutional analog of a proof tree, supplying accountability without requiring a formal proof.

B. Layer Two: translation and validation

This layer converts governance intent into enforceable form. A governance domain-specific language (gDSL) provides a fixed vocabulary of typed primitives: constraint types (behavioral, compositional, temporal), enforcement modalities (slashing conditions, reputation stakes), and agent-scope operators (single agent, multi agent, coalition). Large language models supply a probabilistic first pass; ontology-mediated composition constrains the input space; and tiered validation pools of reputation-staked human experts validate semantic faithfulness. The consensus of the validation pools, recorded on the WDAG, performs the verification function that a formal decision procedure would otherwise serve.

C. Layer Three: cryptoeconomic enforcement

Enforcement uses a dual-token design. Non-transferable reputation tokens are separated from fungible, priceable collateral tokens. This separation closes the governance-capture pathway that single-token systems expose, in which voting power can be purchased or borrowed [6]. Violations trigger slashing deterministically on conditions established by validation-pool consensus and recorded on the graph. Stakes of the losing side are slashed pro rata to the winners, which by the Folk Theorem sustains cooperation across repeated validation interactions [13].

D. Layer Four: temporal governance evolution

The temporal layer lets the system govern its own change. Through staked validation pools operating under fixed hard protocols, the WDAG defines the conditions under which it may itself be amended. This avoids the ossification of static rules without admitting the chaos of ungoverned change, and it does so with no authority external to the system.

V. The Validation Pipeline

The path from governance intent to enforceable predicate is a composition of three stages, each reducing a categorically different class of error by a categorically different mechanism.

Stage one, ontology-mediated guided composition, eliminates input-space errors by constraining proposals to ontologically valid compositions before any formalization is attempted. Stage two, typed compilation of the gDSL, eliminates formal-space errors such as type and category mismatches through type discipline with verified compilation to executable enforcement predicates. Stage three, reputation-weighted validation, adjudicates semantic faithfulness and consistency and records the result on the WDAG, replacing a decidable decision procedure as the verification step.

The pipeline accumulates translation precedent: a structured record mapping intent patterns to validated predicates and the decisions that authorized them. This precedent raises accuracy over time and reduces reliance on the most expensive top-tier validators. Two input tracks feed the pipeline. Natural-language proposals from governance communities enter at stage one. Formal specifications submitted directly by agent deployers require no translation and enter validation directly.

VI. Why a Reputation-Weighted Substrate

A. Expressiveness against enforceability

The pivotal design choice is where to sit on the frontier of Section III. Restricting governance to a decidable formal fragment would buy reliable reasoning at the cost of expressiveness and accessibility, excluding the stakeholders governance must serve. GaaP takes the opposite trade. It retains full natural-language expressiveness and buys reliability through economic incentive alignment and reputation-weighted aggregation, accepting a bounded and auditable residual error rather than claiming its complete elimination. The concession is deliberate and is the paper's central claim: institutional operability is worth a bounded residual.

B. How the substrate bounds the residual

Four mechanisms jointly compress the second failure. Typed primitives remove category errors mechanically. Ontology-mediated composition removes input-space errors. Reputation-weighted multi-party validation suppresses idiosyncratic inference error through aggregation under skin in the game, an effect consistent with results on the aggregation of weighted expert judgment [7], [8]. The append-only audit trail makes the residual visible and contestable rather than hidden. The result is not zero error but error that is bounded, attributable, and correctable.

C. Governed self-modification without a meta-system

Because validation pools stake the same reputation they earn and vote on protocol change under fixed hard protocols, the system is its own meta-system. This institutional form of governed self-modification threads two classical results: Arrow's theorem bounds what preference aggregation can guarantee [14], while the Folk Theorem shows that repeated interaction can sustain cooperation [13]. The WDAG is the construction that reconciles them in an operating system of governance.

D. Open frontier

Multi-agent settings raise problems the substrate must continue to address: Sybil resistance, governance attacks that target the staking layer, and the detection of values drift across populations of agents. Empirical DAO evidence supplies both cautionary cases and design constraints for this frontier [6].

VII. Implications

A. A regulatory infrastructure primitive

GaaP is a callable governance service that deployers and regulators can compose into agent stacks rather than rebuild for each deployment. As a reusable primitive it operationalizes the dynamic and anticipatory regulation program, supplying the feedback and enforcement that reactive instruments lack [4].

B. Accountability through the audit trail

The WDAG trail resolves the attribution problem for autonomous-agent harm. When an agent acts, the governing decision, the evidence behind it, and the validators who staked on it are all recoverable. Accountability follows from institutional traceability rather than from a formal proof, which is what makes it deployable at scale.

C. Sustainable governance

Governance that can evolve without ossifying remains valid as technology and norms shift, which makes it durable infrastructure rather than a static artifact that decays into irrelevance. By internalizing its own amendment, GaaP narrows the regulatory pacing gap that otherwise forces repeated and costly cycles of re-regulation. In the terms of this venue, an institution that sustains its own validity over time is the governance counterpart of a sustainable system.

VIII. Conclusion

Human governance fails because language cannot fully describe reality and human institutions reason about what it does describe unreliably. Neither informal deliberation nor pure formalism resolves both failures at once. Governance as a Product resolves the meta-governance problem differently: a reputation-weighted institutional substrate that governs its own evolution through staked validation, accepting a bounded and auditable residual in exchange for expressiveness and operability. The result is enforceable, accessible, and durable governance infrastructure for autonomous systems.

References

[1] C. Calcaterra, W. A. Kaal, and V. Andrei, "Decentralized governance of distributed systems via interaction control," SSRN 3125822, 2018.

[2] C. Calcaterra, "On-chain governance of decentralized autonomous organizations: Blockchain organization using Semada," SSRN 3188374, 2018.

[3] C. Calcaterra and W. A. Kaal, *Decentralization: Technology's Impact on Organizational and Societal Structure*. Berlin: De Gruyter, 2021.

[4] W. A. Kaal, "Dynamic regulation for innovation," in *Perspectives in Law, Business and Innovation*, M. Fenwick et al., Eds. New York: Springer, 2016.

[5] W. A. Kaal, "Blockchain-based corporate governance," Stanford J. Blockchain Law and Policy, vol. 4, no. 1, 2021.

[6] W. A. Kaal, "The institutional deficit in decentralized autonomous organizations: An empirical analysis of forty DAOs," working paper, 2026.

[7] W. A. Kaal, "Citation honesty mechanisms in weighted directed acyclic graph governance," SSRN 6269518, 2026.

[8] W. A. Kaal, "Evolution of domain-specific reputation systems: From binary validation to citation-weighted knowledge attribution," SSRN 6192998, 2026.

[9] P. Kouvaros, "Towards formal verification of neuro-symbolic multi-agent systems," in Proc. IJCAI, 2023, p. 7014.

[10] R. Beutner and B. Finkbeiner, "Hyper strategy logic," in Proc. AAMAS, 2024, p. 189.

[11] H. Prakken and G. Sartor, "Law and logic: A review from an argumentation perspective," Artificial Intelligence, vol. 227, pp. 214–245, 2015.

[12] M. Gray, J. Savelka, W. Oliver, and K. Ashley, "Generating legal arguments with automatically identified factor magnitudes," in Proc. ICAIL, 2025.

[13] D. Fudenberg and E. Maskin, "The Folk Theorem in repeated games with discounting or with incomplete information," Econometrica, vol. 54, no. 3, pp. 533–554, 1986.

[14] K. J. Arrow, *Social Choice and Individual Values*, 2nd ed. New York: Wiley, 1963.