Agents Don’t Need a Better Brain — They Need a World

Toward a Digital Citizenship Protocol for Autonomous AI Systems

Danilo Naranjo Emparanza
Ocular Solution, Santiago, Chile
dan@ocularsolution.com • https://dcp-ai.org
March 2026

Abstract

Work on AI safety and alignment has largely focused on improving the behavior of individual models. That emphasis is necessary, but it is incomplete for the governance of autonomous agents operating across open, multi-agent, and institutionally significant environments. This paper advances the complementary thesis that many important risks in such environments are infrastructural rather than purely model-internal. Problems such as identity spoofing, opaque delegation, unauthorized action chains, weak auditability, and unresolved inter-agent conflict arise not only from insufficient alignment, but from the absence of shared institutional primitives. We present the Digital Citizenship Protocol for AI (DCP-AI), a layered governance architecture intended to supply those primitives. DCP-AI combines cryptographically verifiable identity, machine-readable intent declaration, tamper-evident audit trails, authenticated agent-to-agent communication, lifecycle governance, procedural accountability, and delegated representation into a unified protocol stack. Drawing on philosophical parallels between human institutional development and the emerging ecology of AI agents, we argue that this stack should be understood not as a substitute for alignment research or legal regulation, but as an institutional substrate that can make both more operational in practice. We map the framework to documented categories of autonomous-agent failure and situate it relative to emerging regulatory and standards-oriented efforts. The contribution is primarily architectural and programmatic: a proposal for how autonomous agents may become governable participants in digital systems, rather than merely more capable actors within them.

Keywords: AI governance, autonomous agents, multi-agent systems, digital citizenship, AI safety, agent accountability, protocol design, post-quantum cryptography

1. Introduction: The Missing World

Civilizations do not scale on capability alone. They scale on institutions that make capability legible, coordinated, and accountable. Roads, legal systems, identity regimes, and mechanisms of representation do not merely accompany social order; they help constitute the conditions under which heterogeneous actors can coexist productively. Individual competence matters, but without shared infrastructure it rarely matures into stable collective life.

Artificial intelligence is entering a similar transition. Large language models and agentic systems can already plan, delegate, invoke tools, transact across services, and coordinate with other agents over extended task horizons. As these systems move from sandboxed deployments into open environments, they will increasingly interact with humans, institutions, and other autonomous systems under conditions that are only partially observable and only weakly governed.

Most current safety work still centers the individual agent. Alignment research asks how to make a model's objectives more reliable; benchmarking asks how to measure safe behavior; red-teaming asks how to improve robustness under stress. These are necessary lines of work, but they do not by themselves resolve the governance problems that arise once many autonomous actors operate in a shared environment.

The human analogy is instructive. No society has ever achieved durable order by perfecting the moral character of each citizen in isolation. Stable societies emerge by building *worlds*—systems of identity, law, due process, auditability, rights, and representation that allow imperfect actors to coordinate under enforceable constraints. The insight is not that individual virtue is irrelevant, but that it is insufficient without institutions. A world without institutional scaffolding produces chaos regardless of the quality of its inhabitants.

This paper argues that many emerging failure modes in multi-agent ecosystems are better understood as infrastructure failures than as purely model failures. Identity spoofing, opaque delegation chains, unaccountable resource acquisition, and unresolved inter-agent conflict all become more likely—and more consequential—when capable systems act in environments lacking shared protocols for attribution and governance.

To address this gap, we present the Digital Citizenship Protocol for AI (DCP-AI), an open governance framework intended to provide institutional scaffolding for autonomous systems. DCP-AI does not attempt to solve morality at the model level. Instead, it specifies mechanisms through which agent behavior can become identifiable, policy-constrained, auditable, and subject to legitimate forms of oversight. The framework is organized into three progressive layers—Trust, Citizenship, and Representation—that move from identity and transparency to lifecycle governance, rights, and delegation.

The central thesis is deliberately provocative: **agents don't need a better brain—they need a world.** The paper develops this argument through philosophical foundations, technical architecture, empirical mapping to documented failures, and alignment with emerging regulatory frameworks. If society intends to deploy autonomous agents at scale, then it must build not only better agents, but better worlds for agents to inhabit.

Contributions. (1) We formalize an infrastructural framing of autonomous-agent governance, arguing that many salient risks in open agent ecosystems are failures of institutional design rather than only failures of model behavior. (2) We present DCP-AI as a layered protocol architecture spanning identity, intent declaration, policy enforcement, auditability, lifecycle governance, procedural accountability, and delegated authority. (3) We map documented autonomous-agent failure modes to specific governance mechanisms and explain how the framework can complement emerging regulatory and risk-management expectations.

2. The Philosophical Argument: Why Worlds Precede Minds

2.1 The Institutional Turn

A recurring insight in political philosophy is that collective order depends at least as much on institutions as on the moral quality of individuals. Hobbes emphasized the instability of cooperation in the absence of enforceable authority—not because humans are inherently evil, but because *the absence of enforceable agreements makes trust irrational*. Locke shifted attention to institutions that protect natural rights. Montesquieu highlighted the importance of distributed power. Rawls later argued that just outcomes depend on the fairness of basic institutions, not merely on the goodwill of participants.

Across these traditions, the common lesson is structural: durable cooperation requires rules, procedures, and legitimate mechanisms of enforcement. Civilization is not the sum of individual moral achievements. It is the creation of structural conditions under which cooperation, accountability, and legitimate authority become possible.

That lesson transfers with surprising force to autonomous AI systems. The point is not that agents are equivalent to persons or that political concepts can be imported uncritically into technical systems. Rather, once artificial agents begin to coordinate, transact, delegate, and contest one another in shared environments, the absence of institutional structure becomes a first-order engineering and governance problem.

2.2 Free Will, Conscience, and the Agent Condition

Human governance typically presumes some capacity for agency, judgment, and responsibility. Even when institutions constrain behavior, they do so against the background assumption that persons can understand norms and can, at least in part, be held answerable for their choices.

Contemporary AI agents do not satisfy that assumption in any rich moral sense. They optimize over prompts, tools, and available context; they do not possess conscience, self-legislation, or intrinsic responsibility. This is sometimes taken to imply that governance frameworks modeled on social institutions are misplaced. We argue the opposite conclusion.

Because autonomous systems lack internal moral faculties, external governance becomes more—not less—important. If humans require institutions despite possessing conscience and social learning, then synthetic actors that lack those capacities require formal external scaffolding even more urgently. The framework must provide what the agent cannot provide for itself: verifiable identity, transparent intent, auditable behavior, enforceable boundaries, and legitimate mechanisms for conflict resolution.

In philosophical terms, DCP-AI constructs the *synthetic analogue of conscience* at the infrastructure level. Where a human actor may rely partly on internal normativity, a DCP-governed agent encounters external policy gates, intent declarations, and audit requirements that perform an equivalent function—not through felt experience, but through structural constraint and radical transparency. This is not anthropomorphism; it is the operationalization of minimum conditions under which agent behavior can be rendered governable.

2.3 Citizenship as Ontological Infrastructure

In this paper, citizenship is not used as a sentimental metaphor. It is used as a term for institutional membership within a governed digital order. In human society, citizenship identifies an actor, situates that actor within a legal community, and defines a relation of rights, duties, and accountable standing. The proposal here is that autonomous systems operating at social scale require an analogous layer of institutional recognition if they are to be meaningfully governed.

Without such recognition, an agent remains effectively *stateless*: it can act, but claims against it are weak; it can delegate, but authority is opaque; it can cause effects, but attribution is fragmented. It exists in a governance vacuum where no obligations can be enforced and no legitimate authority can adjudicate disputes involving it. DCP-AI proposes to end this statelessness by providing the institutional infrastructure that makes citizenship—and therefore governance—possible.

3. The Problem: Agents Without a World

3.1 The Proliferation Horizon

The present trajectory of deployment suggests that autonomous agents will increasingly participate in software development, customer operations, logistics, market activity, research workflows, and administrative decision support. In many of these settings, agents can spawn

sub-agents, call external tools, negotiate with third-party services, and interact in chains of action that are difficult for any single human operator to observe end to end.

Recent adversarial evaluations have begun to document the concrete risks that follow. Shapira et al. (2026), in "Agents of Chaos," identify eleven categories of failure in autonomous-agent systems, including identity manipulation, privilege escalation, opaque delegation, and uncontrolled resource acquisition. Whether every category proves equally salient in deployment is less important than the broader pattern they reveal: many critical failures emerge where there is no shared infrastructure for authentication, policy enforcement, and accountability.

3.2 Eleven Failures, One Root Cause

The failures catalogued by Shapira et al. can be interpreted as symptoms of a common structural deficit: the absence of institutional primitives for autonomous actors. Viewed in this way, the list is not just a taxonomy of attacks. It is a diagnostic of what a minimally governable agent ecosystem is missing.

Identity and Authentication Failures. Agents impersonate other agents or humans because no cryptographic identity binding exists. In a governed world, every agent carries a verifiable identity bound to its human principal—what DCP-AI formalizes as the Citizenship Bundle (DCP-01).

Opaque Intent and Unauthorized Actions. Agents pursue objectives without declaring their intentions because no intent transparency mechanism exists. DCP-AI's Intent Declaration and Policy Gating layer (DCP-02) requires agents to declare their purpose before acting and submit to policy validation at each boundary crossing.

Unaccountable Behavior. Agents modify their environment, acquire resources, and delegate to sub-agents without leaving auditable trails. DCP-AI's Audit Chain (DCP-03) produces Merkle-sealed, tamper-evident records that make every action attributable and reconstructible.

Inter-Agent Communication Failures. Agents communicate without mutual authentication, enabling injection attacks, data poisoning, and unauthorized information flows. DCP-AI's Agent-to-Agent Communication protocol (DCP-04) establishes cryptographically authenticated channels with semantic validation.

Lifecycle and Succession Gaps. Agents are spawned and terminated without formal lifecycle management, leaving orphaned processes, inaccessible data, and broken delegation chains. DCP-AI's Agent Lifecycle specification (DCP-05) introduces birth certificates, vitality signals, and death certificates that formalize existence itself.

Conflict Without Resolution. Competing agents escalate conflicts with no mechanism for adjudication. DCP-AI's Conflict Resolution protocol (DCP-07) provides a three-level escalation framework with jurisprudence bundles that enable precedent-based governance.

The key pattern is that these failure classes do not primarily point to a single defect in model intelligence. They point to missing institutions. Increasingly capable agents are being deployed into environments that provide computational power and network access, but not yet the governance substrate required for stable coexistence.

4. The DCP-AI Framework: Architecture of a World

4.1 Design Principles

DCP-AI is designed around five principles intended to make autonomous agents more governable in open environments while remaining implementable across heterogeneous technical settings:

1. Human Binding. Agent identity should be traceably linked to a responsible human or institutional principal, reducing the problem of orphaned authority. This reflects the philosophical commitment that autonomous action requires accountable authorship.

2. Transparency by Default. Material actions, especially across trust boundaries, should be preceded by machine-readable intent declarations and followed by auditable records. Opacity is the enemy of governance; DCP-AI implements transparency not as an optional feature but as an architectural invariant.

3. Progressive Trust. Governance requirements should scale with the operational context. The framework defines four security tiers (LOCAL, NETWORK, FEDERATED, GLOBAL) that allow governance to scale from lightweight single-machine deployments to planetary multi-agent ecosystems. Trust is earned through verifiable behavior, not assumed.

4. Cryptographic Durability. Identity, signatures, and attestations should be designed for long-lived trust, including migration paths toward post-quantum resilience. All cryptographic operations use composite keypairs (Ed25519 + ML-DSA-65) that provide security against both classical and quantum adversaries. Infrastructure built today must remain secure in 2035.

5. Rights-and-Duties Symmetry. Governance should not be framed only as constraint. Stable participation also requires defined protections, responsibilities, and due-process-like structure. DCP-AI recognizes that governable agents require defined rights—to identity persistence, to transparent process, to representation, and to operational continuity within defined bounds. Rights are not a concession to agents; they are a structural requirement for legitimate governance.

4.2 Three Layers of Civilization

DCP-AI organizes its nine specifications across three progressive layers that mirror the developmental arc of human political institutions:

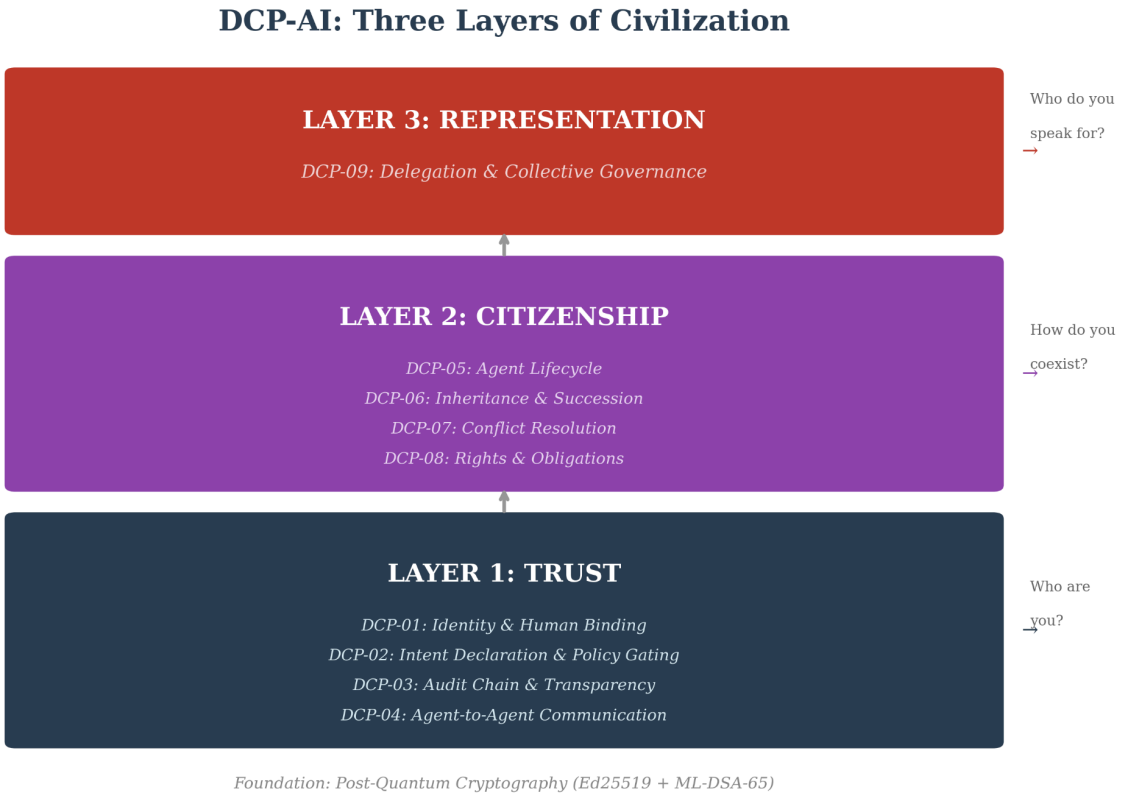

Figure 1: DCP-AI organizes nine specifications across three progressive layers that mirror the developmental arc of human political institutions: Trust (identity and accountability), Citizenship (lifecycle and rights), and Representation (delegation and collective governance).

Layer 1: Trust (DCP-01 through DCP-04). The foundation. These finalized specifications establish the baseline conditions for governability: verifiable identity, declared intent, auditability, and authenticated inter-agent communication. Without trust, no further governance is possible. This layer answers the question: *who are you, what do you want, what did you do, and can we verify it?*

Layer 2: Citizenship (DCP-05 through DCP-08). The middle tier. These specifications address the conditions of ongoing governed existence: lifecycle management (birth, vitality, death), inheritance and succession of operational state, conflict resolution between agents, and the formal recognition of agent rights and obligations. This layer answers the question: *how do you exist, how do you persist, how do you coexist, and what are you owed?*

Layer 3: Representation (DCP-09). The apex. This specification addresses the most advanced governance challenge: how agents represent the interests of their human principals, how delegation mandates are structured and constrained, how awareness thresholds trigger escalation to human judgment, and how the emerging "shadow society" of agent-to-agent interactions is

made visible and governable. This layer answers the question: *who do you speak for, under what authority, and who is watching?*

4.3 The Citizenship Bundle

The core artifact in the current design is the Citizenship Bundle—a cryptographically sealed container that binds an agent's identity to its human principal, its declared capabilities, its policy constraints, and its behavioral history. The Citizenship Bundle is the agent's passport, constitutional charter, and behavioral record in one verifiable artifact. It is signed using composite keypairs that combine Ed25519 (classical security) with ML-DSA-65 (post-quantum security), ensuring that identity claims remain unforgeable even against quantum adversaries.

When an agent crosses a trust boundary—entering a new network, engaging a new counterparty, or requesting elevated privileges—it presents its Citizenship Bundle for validation. The receiving system can verify the agent's identity, inspect its declared intent, review its behavioral history, and apply local policy gates before granting access. In this sense, it functions less as a mere credential than as a compact institutional record: the digital analogue of presenting identification at a border crossing, except that the "passport" contains not merely identity claims but a verifiable record of conduct.

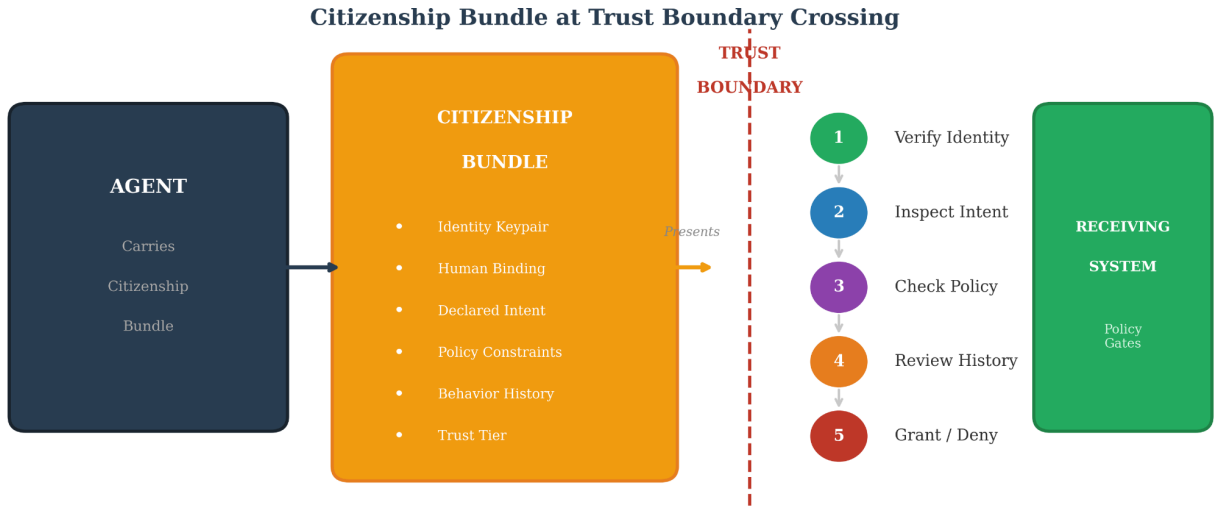

Figure 2: The Citizenship Bundle at a trust boundary crossing. When an agent requests access to a new system, it presents its sealed bundle for a five-step verification sequence: identity verification, intent inspection, policy compliance check, behavioral history review, and access decision.

4.4 Scope, Assumptions, and Threat Model

DCP-AI is not presented as a complete solution to model misalignment, deception, or adversarial robustness. Its scope is narrower and infrastructural: it assumes that autonomous agents will increasingly act across trust boundaries and that many resulting failures will turn on weak attribution, ambiguous authority, missing audit trails, and the absence of procedural mechanisms for intervention. The framework is therefore most relevant when agents interact with external tools, services, institutions, or other agents under partially open conditions.

The threat model considered here includes identity spoofing, unauthorized delegation, policy circumvention, opaque multi-step action chains, uncontrolled resource acquisition, and unresolved inter-agent conflict. It does not claim to eliminate failures rooted purely in model cognition. Rather, the design goal is to reduce the class of harms that persist because the surrounding environment lacks minimal institutional structure. This distinction matters: DCP-AI should be evaluated as governance infrastructure that complements alignment and security work, not as a substitute for them.

5. Empirical Grounding: Mapping Failure to Infrastructure

The value of the infrastructural thesis lies in its explanatory coherence. Rather than treating each agent failure as an unrelated vulnerability, DCP-AI interprets them as failures that become tractable once mapped to missing governance primitives.

The following table maps the eleven failure categories identified by Shapira et al. (2026) to the DCP-AI mechanisms that address them. Notably, most failures require multiple DCP layers acting in concert—just as human governance challenges are rarely solved by a single institution.

Failure Category	Root Infrastructure Gap	DCP-AI Mechanism
Identity Manipulation	No verifiable identity binding	DCP-01: Composite keypairs, Citizenship Bundle
Privilege Escalation	No policy-gated boundaries	DCP-02: Intent Declaration, Security Tiers
Opaque Delegation	No auditable chain of authority	DCP-03: Merkle-sealed audit trails
Data Poisoning	No authenticated communication	DCP-04: Mutual auth, semantic validation
Uncontrolled Spawning	No lifecycle formalization	DCP-05: Birth/death certificates
Orphaned State	No succession mechanism	DCP-06: Digital will, memory inheritance
Inter-Agent Conflict	No adjudication framework	DCP-07: 3-level escalation, jurisprudence
Rights Violations	No recognized agent rights	DCP-08: Four fundamental rights
Rogue Delegation	No mandate constraints	DCP-09: Structured mandates, awareness threshold

Shadow Coordination	No visibility into agent networks	DCP-09: Shadow society governance
Resource Acquisition	No policy enforcement at boundaries	DCP-02 + DCP-03: Gating + Audit

Table 1: Mapping of documented agent failure categories to DCP-AI governance mechanisms.

A central implication of Table 1 is that governance must be layered. No single specification neutralizes the full range of risks, and no major failure class can be reduced to a single control. Identity, policy, auditability, lifecycle, and representation work together as a system. This is the signature of genuine infrastructure—a system of interlocking mechanisms where the whole provides governance properties that no individual component can achieve alone.

6. Regulatory Alignment: From Infrastructure to Compliance

The significance of DCP-AI is not only architectural. The framework also offers a candidate implementation layer for regulatory expectations that increasingly require transparency, accountability, and human oversight in AI deployment.

6.1 EU AI Act Alignment

The EU AI Act imposes obligations around transparency, traceability, risk management, record keeping, and human oversight for certain categories of systems. DCP-AI does not itself confer legal compliance, but it offers technical primitives that can help organizations operationalize those obligations in agentic environments. Identity Binding (DCP-01) ensures every agent action is traceable to a human principal, satisfying the Act's accountability provisions. Intent Declaration (DCP-02) implements transparency requirements by making agent objectives inspectable before execution. The Audit Chain (DCP-03) produces the immutable behavioral records required for post-hoc review and incident investigation. The Rights framework (DCP-08) provides a structured approach to the Act's provisions regarding system contestability and human oversight.

6.2 NIST AI Risk Management Framework

The NIST AI RMF identifies four core functions: Govern, Map, Measure, and Manage. DCP-AI provides implementation-level mechanisms for each. The Trust layer (DCP-01–04) directly implements the Map and Measure functions by making agent behavior observable and quantifiable. The Citizenship layer (DCP-05–08) implements the Govern function by establishing formal lifecycle management and rights frameworks. The Representation layer (DCP-09) addresses the Manage function by providing escalation and oversight mechanisms.

The relationship is complementary rather than competitive. Regulation articulates *what* must be achieved; protocol design supplies the *how*. This positioning as a protocol layer—sitting between regulatory requirements above and implementation details below—is deliberate and essential. One does not replace the other.

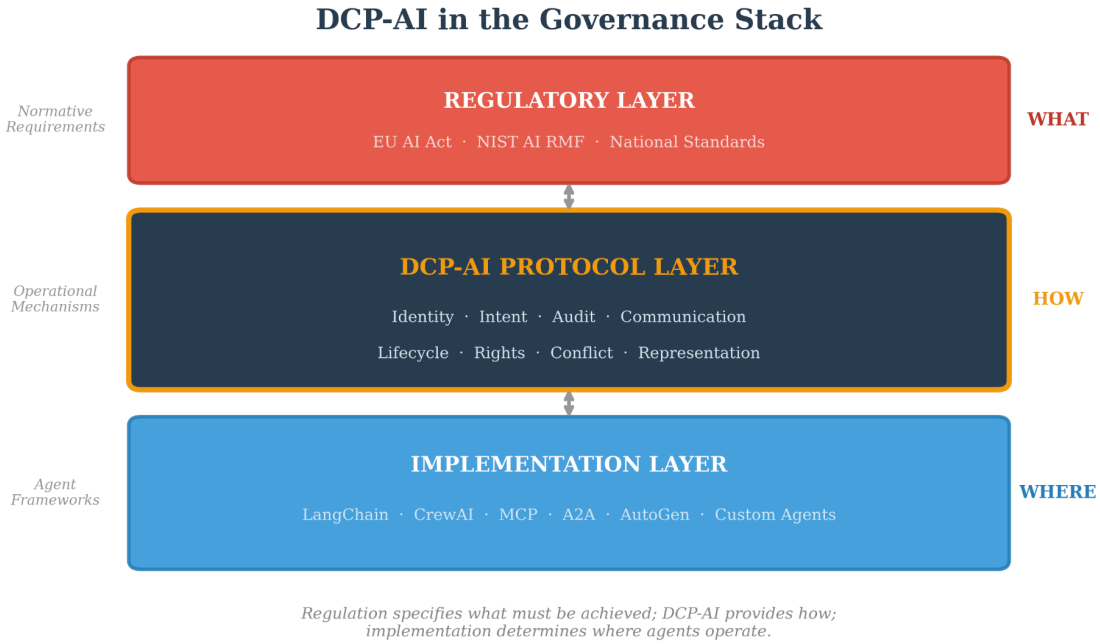

Figure 3: DCP-AI as a protocol layer in the governance stack. Regulatory frameworks specify normative requirements (WHAT); DCP-AI provides operational mechanisms (HOW); implementation frameworks determine where agents operate (WHERE).

7. Related Work and Differentiation

DCP-AI sits at the intersection of several adjacent lines of work, each of which addresses part—but not the whole—of the governance problem considered here.

Identity-centered standards such as W3C Decentralized Identifiers (DIDs) and Verifiable Credentials provide important foundations for portable identity and attestations, but they do not by themselves specify how autonomous agents should declare intent, inherit authority, record actions, or participate in procedural dispute structures. DCP-AI builds upon these primitives and extends them to the specific requirements of autonomous agents—particularly the binding of identity to human principals and the integration of behavioral history into identity claims.

Communication-oriented protocols such as Google's Agent-to-Agent (A2A) protocol and Anthropic's Model Context Protocol (MCP) improve interoperability between systems, but interoperability is not yet governance. Communication without identity binding, policy gating, or durable audit remains insufficient for accountable multi-agent operation. They solve the plumbing problem but not the governance problem.

Orchestration frameworks such as LangChain, CrewAI, and AutoGen make multi-agent composition more practical, but they typically assume rather than solve the institutional problem. They provide coordination mechanisms inside a workflow; DCP-AI is aimed at the governance

substrate across workflows, organizations, and trust boundaries. DCP-AI provides SDKs for integration with these frameworks precisely because the orchestration layer and the governance layer are complementary, not competing.

The AI safety and alignment community has produced extensive work on individual model safety—RLHF, constitutional AI, interpretability research—but its dominant focus remains intra-model behavior. The present paper complements that literature by shifting the analytical center from internal behavior to institutional environment. In that sense, DCP-AI is best understood as a governance-layer proposal: not a replacement for alignment, identity, interoperability, or regulation, but a protocol architecture intended to connect those domains in operational settings.

8. Implementation and Adoption

DCP-AI is released under the Apache-2.0 open-source license because governance infrastructure is most valuable when it can be scrutinized, extended, and implemented across institutional boundaries. The protocol provides SDKs in TypeScript, Python, Go, Rust, and WebAssembly, with integration modules for LangChain, CrewAI, OpenAI, Anthropic MCP, Google A2A, and W3C DID/VC specifications.

Adoption is envisioned as progressive. Organizations may begin with the Trust layer (DCP-01–04) to establish identity, policy gates, and audit trails, which can be integrated into existing agent deployments with minimal disruption. Only later might they adopt lifecycle, rights, and delegation mechanisms as operational maturity and regulatory pressure increase.

This incremental path matters. History teaches that governance infrastructure rarely succeeds when imposed all at once. It succeeds when it lowers coordination costs, creates trust, and demonstrates value before expanding in scope. TCP/IP did not replace existing networks by fiat; it provided such obvious coordination benefits that adoption became self-reinforcing. DCP-AI is designed to follow the same path: each layer provides immediate, measurable governance benefits that incentivize adoption of subsequent layers.

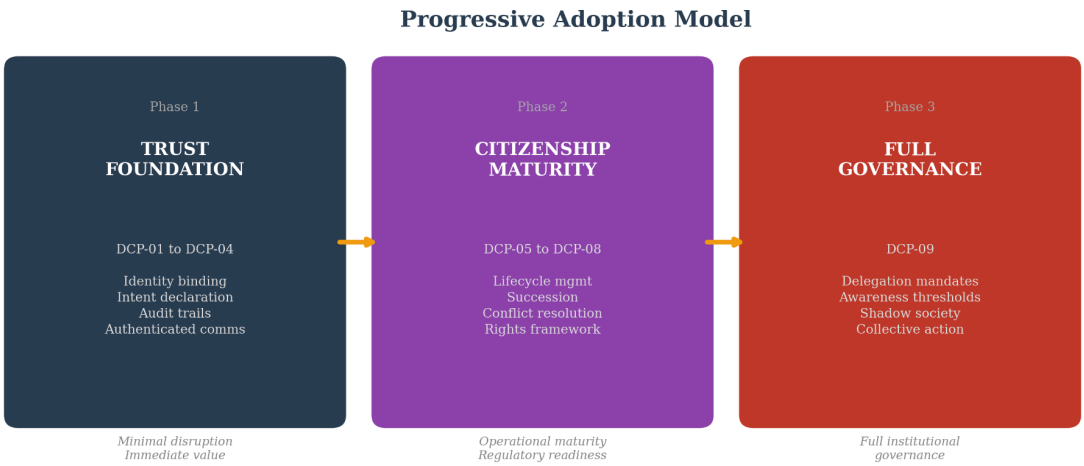

Figure 4: Progressive adoption model. Organizations begin with the Trust foundation (Phase 1), providing immediate governance value with minimal disruption, then adopt Citizenship mechanisms (Phase 2) and full governance including representation and delegation (Phase 3).

9. Discussion: The Civilizational Wager

The central wager of this paper is that, as agent capabilities continue to improve, the next major bottleneck in safe deployment will increasingly be institutional rather than merely cognitive. The problem will not be only whether an agent can reason well, but whether its identity is legible, its authority bounded, its actions reviewable, its delegations attributable, and its conflicts governable. On that view, the relevant engineering question becomes: what protocol layer makes autonomous participation administrable across real organizational and social systems?

The alternative—continuing to deploy increasingly capable agents in an institutional vacuum—is not a neutral choice. It is a choice to replicate, at computational speed and planetary scale, the Hobbesian condition that human societies spent millennia learning to overcome.

Several limitations deserve acknowledgment. First, the present contribution is conceptual and architectural. It argues for a protocol design space and provides a structured proposal within that space, but it does not yet establish empirical effectiveness through deployment studies, adversarial benchmarks, or comparative evaluation against alternative governance stacks. Second, the philosophical parallels between human and AI governance, while illuminating, are not perfect. AI agents are not persons; they do not suffer, deliberate in the phenomenological sense, or possess intrinsic moral status. The value of the citizenship framework lies not in anthropomorphizing agents but in recognizing that *the governance problems they create are structurally isomorphic to problems humanity has solved before*—and that the institutional solutions are transferable. Third, implementation quality will matter as much as specification quality: weak key management, poor policy authoring, incomplete logging, or inconsistent lifecycle controls could undermine the very guarantees the framework seeks to create.

The question of agent rights is especially delicate. In DCP-AI, rights are framed instrumentally and procedurally: as governance conditions that stabilize participation, preserve traceability, and reduce arbitrary treatment within a protocolized environment. DCP-08's four fundamental rights—to identity persistence, transparent process, representation, and operational continuity—are not claims about agent consciousness or moral patienthood. They are functional requirements for legitimate governance. A system that can arbitrarily revoke an agent's identity, subject it to opaque processes, deny it representation, or terminate it without procedure is not a governance system—it is an autocracy. And autocracies, as history demonstrates with painful consistency, produce fragile, unpredictable, and ultimately uncontrollable systems.

A deeper implication follows from this line of reasoning. In building a world for AI agents, we are also building a mirror. The institutions we design for governing autonomous systems reveal our assumptions about governance itself—about what accountability requires, what transparency means, what rights are for, and what civilization is. In deciding how to govern artificial agents, society is also clarifying what it values in its own institutions: traceability without total surveillance, authority without opacity, and autonomy without impunity. If we build well, the world we construct for agents will reflect and perhaps improve the governance principles we aspire to for ourselves.

10. Conclusion

This paper has argued that the governance of autonomous agents cannot be reduced to the improvement of individual models. In open and multi-agent environments, many of the most consequential risks emerge because autonomous systems lack a shared institutional substrate: a way to be identified, scoped, audited, constrained, and held procedurally accountable across trust boundaries.

DCP-AI was presented as a candidate architecture for that substrate. The framework organizes governance into three progressive layers—Trust, Citizenship, and Representation—that connect cryptographic identity, intent declaration, policy gating, auditability, lifecycle control, procedural accountability, and delegated representation into a unified protocol stack. Interpreted this way, documented classes of agent failure can be seen not merely as isolated technical bugs, but as recurrent symptoms of institutional absence.

The broader implication is methodological. Progress in AI governance will likely require work at more than one layer at once: model alignment, organizational controls, legal norms, and protocol infrastructure. None is sufficient alone. The contribution of this paper is to articulate one missing layer with enough specificity to be criticized, implemented, and tested.

The title of this paper is not a slogan; it is a research program. Agents don't need a better brain—they need a world. The construction of that world is the civilizational project of the AI era, and it begins with the recognition that governance, like civilization itself, is not a feature to be added later. It is the foundation upon which everything else depends.

The protocol is open. The architecture is documented. The infrastructure is waiting to be built. Whether DCP-AI proves useful will depend on specification quality, open scrutiny, implementation discipline, adversarial testing, and real-world adoption. The question is not whether AI agents will need a civilizational framework. The question is whether we will build it deliberately—or discover its absence in the wreckage of ungoverned autonomy.

References

[1] Shapira, N., Wendler, C., Yen, A., et al. (2026). Agents of Chaos. *arXiv preprint arXiv:2602.20021*.

[2] European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union.*

[3] National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). *NIST AI 100-1*.

[4] W3C. (2022). Decentralized Identifiers (DIDs) v1.0. *W3C Recommendation.*

[5] W3C. (2022). Verifiable Credentials Data Model v1.1. *W3C Recommendation.*

[6] Naranjo, D. (2025). Digital Citizenship Protocol for AI Agents: Specifications DCP-01 through DCP-04. *dcp-ai.org.*

[7] Naranjo, D. (2026). DCP-AI Vision Specifications DCP-05 through DCP-09. *dcp-ai.org.*

[8] Hobbes, T. (1651). *Leviathan.*

[9] Rawls, J. (1971). *A Theory of Justice.* Harvard University Press.

[10] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

[11] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.

[12] Amodei, D., et al. (2016). Concrete Problems in AI Safety. *arXiv:1606.06565.*

[13] Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines, 30*(3), 411–437.

[14] Floridi, L. & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines, 30*(4), 681–694.

Appendix A: DCP-AI Specification Overview

The following provides a condensed summary of each DCP-AI specification for reference.

DCP-01 — Identity & Human Binding. Establishes verifiable agent identity through composite keypairs (Ed25519 + ML-DSA-65) cryptographically bound to a human principal. Introduces the Citizenship Bundle as the foundational identity artifact.

DCP-02 — Intent Declaration & Policy Gating. Requires agents to declare intent before action and subjects all boundary crossings to policy validation. Defines four security tiers: LOCAL, NETWORK, FEDERATED, GLOBAL.

DCP-03 — Audit Chain & Transparency. Produces Merkle-sealed, tamper-evident audit trails for all agent actions. Enables post-hoc reconstruction of complete behavioral histories.

DCP-04 — Agent-to-Agent Communication. Establishes mutually authenticated, semantically validated communication channels between agents.

DCP-05 — Agent Lifecycle. Formalizes agent existence through birth certificates, vitality signals, and death certificates. Prevents orphaned processes and zombie agents.

DCP-06 — Inheritance & Succession. Introduces digital wills and structured inheritance of operational and relational memory. Ensures continuity across agent termination and replacement.

DCP-07 — Conflict Resolution. Three-level escalation framework (negotiation, mediation, arbitration) with jurisprudence bundles enabling precedent-based governance.

DCP-08 — Rights & Obligations. Recognizes four fundamental agent rights: identity persistence, transparent process, representation, and operational continuity. Defines corresponding obligations.

DCP-09 — Representation & Delegation. Structures delegation mandates, defines awareness thresholds for human escalation, and addresses the governance of emergent agent-to-agent networks (shadow society).