Intelligent Voluntary Cooperation

Shaun Conway

August 30, 2026

DRAFT

CONTENTS

Abstract

How to read the evidence

- Working definitions
- Scope and non-goals

Part I — The coordination problem

- 1. Agency is becoming abundant; coordination is not
- 2. Beyond the choice between hierarchy and chaos
- 3. The thesis and its limits

Part II — Environmental coordination

- 4. Stigmergy: coordination through traces
- 5. Cognition distributed across people and artifacts
- 6. Artifacts as inherited capability
- 7. Affordances: what this actor can do here

Part III — Agent evidence and SwarmWorld

- 8. Why SwarmWorld matters
- 9. What the study found
- 10. What the study does not establish

Part IV — A theory of intelligent voluntary cooperation

- 11. The bounded cooperative world
- 12. The causal loop
- 13. Voluntary means more than unassigned
- 14. Cooperation is not one economic category
- 15. Hybrid coordination by consequence and coupling

DRAFT

16. What makes the cooperation intelligent

Part V — Governance and trustworthy cooperation

17. Persistence is power and risk

18. Correction, decay, and epistemic maintenance

19. Polycentric governance, not automatic decentralization

20. Identity, Sybils, reputation, and plural participation

21. Meaningful human authority

22. Governance requirements

Part VI — Cognition and consequence

23. The reality boundary

24. Cryptographic integrity is not semantic truth

25. Evaluation and independence

26. Receipts and authoritative read-back

Part VII — Qi as a candidate cooperative environment

27. What is being claimed

28. Topic identity and bounded context

29. Capsules, artifacts, and agent participation

30. Portal's implemented shared-state surface

31. The proof still required

32. Qi Runtimes and the model-host boundary

33. Flows, PODs, and executable inheritance

Part VIII — IXO as a consequential substrate

34. From cooperative representation to authoritative state

35. Identity and entities

36. Claims, evaluation, disputes, and payments

DRAFT

37. Evaluation services and bounded prototypes

38. The honest end-to-end status

Part IX — Toward a cooperative economy

39. From isolated services to inherited worlds

40. Incentives for heterogeneous contribution

41. Humans, agents, organisations, and things

42. Compounding utility and defensibility

Part X — Research agenda and implications

43. A falsifiable program

44. Measurement and public accountability

45. Implications for builders

46. Implications for partners, researchers, and funders

47. Conclusion

References

Public annexes

DRAFT

> **Working draft.** This is the pre-publication `0.9.0-draft`. System statements are locked to the versioned sources listed in the correspondence matrix but still await named technical-owner sign-off. The public comment window and DOI have not opened. Do not cite this draft as the final version.

DRAFT

Abstract

Artificial intelligence is making useful agency cheaper, faster, and more widely available. The harder problem is increasingly not whether an actor can generate a plan or perform a task, but how many autonomous people, organisations, software agents, and devices can coordinate without surrendering initiative to a single orchestrator or allowing machine output to acquire unearned authority. This paper develops **intelligent voluntary cooperation** as a computational and institutional model for that problem. Participants coordinate around bounded intent by perceiving and modifying a persistent shared environment. Their actions leave artifacts that can guide later action, survive participant turnover, and sometimes accumulate into reusable capability. Voluntariness, however, requires more than decentralization: participants need informed choice, bounded authority, voice, contestation, fair participation conditions, and meaningful exit. Trustworthiness requires a further boundary among claims, evidence, evaluation, decisions, authorization, and consequential state.

The argument integrates research on stigmergy, distributed cognition, cooperative work, affordances, commons governance, cumulative culture, security, and human–AI collaboration. It treats the recent SwarmWorld simulation as relevant but preliminary evidence: under one bounded model and simulator configuration, persistent shared state supported differentiated behavior, broad artifact reuse, technological lineages, and portfolio-level advantages over matched isolated search. It did not validate Qi, human institutions, or real-world economic outcomes. The paper then evaluates Qi and IXO as a candidate implementation. Portal already contains Matrix/Yjs collaborative state and executable Flow primitives; Qi Runtimes implement governed workspace and authority boundaries; the IXO chain implements identity, entities, claims, evaluation states, escrow, disputes, and payments. Topic Protocol is a draft implementation candidate with open live-conformance gates, and no fully conformed production path from Topic action through settlement is asserted. The resulting contribution is a falsifiable architecture and research agenda, not a declaration that emergent cooperation has already been achieved.

Keywords: cooperation; collective intelligence; stigmergy; multi-agent systems; distributed cognition; governance; authority; claims; evidence; Qi; IXO

DRAFT

How to read the evidence

The paper distinguishes six kinds of statement. `[E]` marks established evidence within a stated domain; `[X]` a result from a named experiment; `[S]` a verified property of a versioned system source; `[I]` the author's theoretical or architectural inference; `[H]` a proposition that remains to be tested; and `[N]` a normative principle. Fifteen core claims, C01–C15, are maintained in a machine-readable ledger. The labels appear at important claim boundaries in this working draft and remain in the public annex even if later removed from the flow of the final prose.

A system specification is not deployment evidence. A cryptographic credential is not proof that its claim is true. A valid capability is not evidence that an action is wise. A successful local tool call is not an authoritative external consequence. These distinctions are not caveats around the argument. They are the architecture's central discipline.

Working definitions

The argument uses ordinary words in deliberately bounded ways. An **actor** is a human, organisation, software agent, device, or governed service that can perceive relevant state and take some action. A model is not automatically an actor: it may supply inference inside a host while identity, memory, tools, authority, and persistence belong to other components. A **participant** is an actor recognized in a particular cooperative context.

An **environment** is the bounded medium containing observable state, rules, resources, and action possibilities. A message history or database can be part of an environment, but neither becomes a cooperative world merely by storing information. **Intent** is a declared future condition toward which cooperation is oriented. It does not prove that every participant shares the same motive. **State** is the environment's represented condition; representation may be proposed, disputed, accepted for a purpose, or authoritative in a particular system.

A **trace** is a change left by an action that can influence later action. An **artifact** is a persistent and addressable trace with informational, executable, institutional, or consequential function. Persistence alone does not make an artifact useful. An **affordance** is an action possibility for a particular actor given the environment, the

DRAFT

actor's practical capability, and its legitimate authority. **Capability** means what an actor is able to do; **authority** means what it is permitted to decide or effect within a scope. Neither identity nor ability implies permission.

A **claim** is an assertion about an event, condition, contribution, or entitlement that may require evaluation. **Evidence** is material used under a stated method to support, weaken, or refute that claim. **Evaluation** is the application of a method; it can be reproducible and still mistaken. A **decision** is a determination made by an identifiable role under a rule or authority. A **consequence** is an effective change in rights, resources, credentials, permissions, obligations, or authoritative state. Recommendation, approval, execution, and consequence are separate events.

Coordination is the management of dependencies among activities. **Collaboration** emphasizes working together, often through communication and joint production. **Cooperation** means combining or aligning contributions toward a shared or compatible outcome; mere co-presence is insufficient. **Orchestration** coordinates through explicit decomposition, assignment, sequence, and aggregation by a controlling process. **Stigmergic coordination** is indirect coordination in which traces in a medium influence later action.

Voluntary does not mean unregulated, costless, or free of all commitments. It means that participation and contribution occur under sufficiently informed and practical choice, with bounded delegation, effective voice and contestation, and meaningful refusal or exit consistent with legitimate prior obligations. **Legitimate** is a normative and institutional judgment, not a property conferred by code. **Reality boundary** names the procedures and authorities through which assertions become accepted consequential state for a defined system; it does not imply metaphysical truth.

These definitions prevent category errors. A collaborative document can coordinate without being authoritative. A blockchain state can be authoritative for a protocol while encoding a false real-world claim. An autonomous agent can have operational freedom while its human principal lacks meaningful choice. The theory depends on keeping those combinations visible.

Scope and non-goals

This is an architecture and research paper, not a protocol standard, investment prospectus, legal opinion, or claim that one institution can govern every form of cooperation. It does not propose to move all shared state onto a blockchain, replace

organisations with agent swarms, eliminate managers, or make every human interaction machine-readable. Many valuable relationships depend on discretion, trust, care, ambiguity, and the right not to leave a permanent record.

The paper also does not define consciousness, general intelligence, moral patienthood, or a complete theory of human cooperation. Software agents are treated as operational actors only within explicit technical and institutional boundaries. Questions about their future moral status remain outside this version. “Things” means devices or assets participating through sensing, state, control, or representation; it does not imply that an object can consent.

No original Qi experiment is reported in version 1. The comparative protocols are published so that later claims can be evaluated against preregistered baselines rather than demonstration narratives. Present system archaeology is a dated snapshot, not an assurance that every repository, deployment, or operating process will remain unchanged. Technical-owner sign-off validates statements within each system; it does not endorse the paper’s theoretical synthesis.

Finally, the paper does not assume that more cooperation is always desirable. Refusal, competition, privacy, separation, local autonomy, and independent experimentation can protect people and improve discovery. The design objective is not to maximize connectedness. It is to make cooperation available where participants have compatible intent and legitimate conditions, while preserving the boundaries that allow them to disagree, fork, or leave.

DRAFT

Part I — The coordination problem

1. Agency is becoming abundant; coordination is not

The dominant image of artificial intelligence is an increasingly capable individual assistant: a system that can answer, analyze, design, code, search, negotiate, or execute. That image captures an important change but misses its systemic consequence. When many people and organisations can call on many capable agents, the relevant unit is no longer one user and one model. It is a population of heterogeneous actors operating at different speeds, with different information, incentives, tools, rights, and liabilities. The number of plausible actions grows faster than any one participant's ability to assign, inspect, sequence, and reconcile them.

The coordination burden appears in ordinary forms. Two agents repeat the same research because neither can see the other's durable result. A new collaborator reconstructs context from a long conversation. A useful procedure disappears with the person who devised it. A model recommends a transaction, the interface reports success, and nobody checks the authoritative state. A central planner becomes the bottleneck because every new actor produces another queue of assignments, updates, exceptions, and approvals. None of these failures requires unintelligent participants. They are failures in the environment through which intelligence becomes collective action.

This paper begins with an inference rather than a measured macroeconomic law: as capable agency becomes cheaper and more abundant, coordination, governance, and consequence become increasingly limiting resources. **C01** `[I/H]`**.** The claim is deliberately testable. If growing agentic capacity does not increase duplicated work, reconciliation cost, authority failures, or governance burden—and if existing orchestration scales without a material loss of quality or accountability—the motivating thesis should be narrowed.

2. Beyond the choice between hierarchy and chaos

The usual response to many actors is orchestration. A controlling process decomposes a goal, assigns tasks, sequences execution, monitors progress, and aggregates results. This is indispensable where work is tightly coupled, the correct

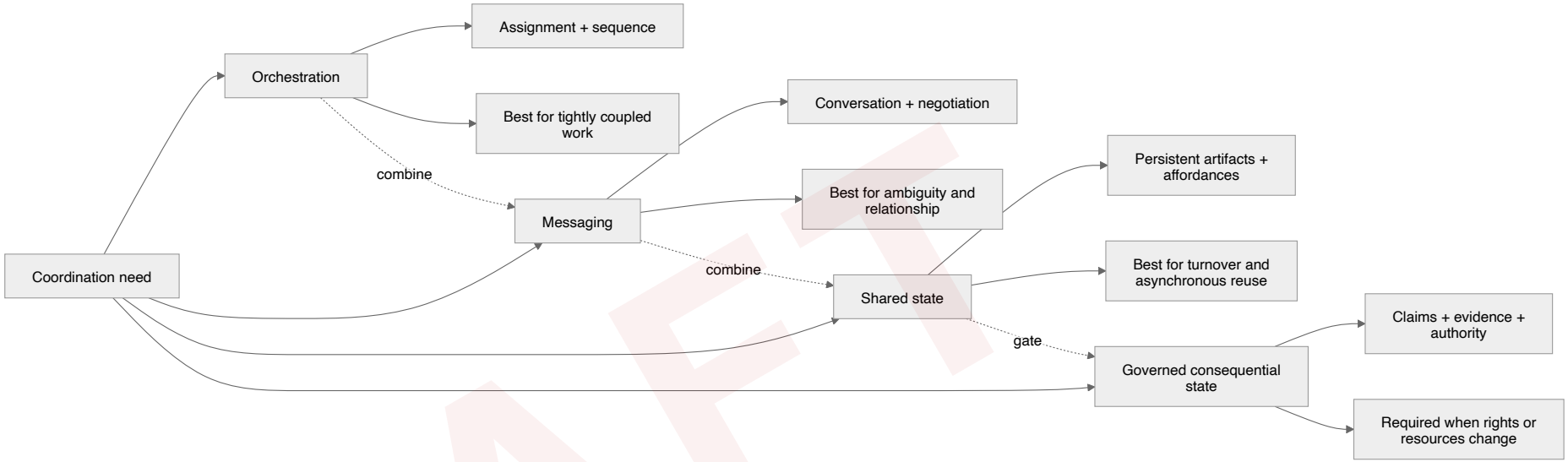
sequence is known, timing is critical, or a single accountable authority must control the result. An aircraft checklist, a key ceremony, a regulated payment, and a database migration are not improved merely by letting every participant improvise.

But orchestration has structural costs. It requires the coordinator to know enough to decompose the problem, recognize relevant capabilities, update assignments as conditions change, and preserve context across turnover. Situated-action research shows why plans cannot completely determine action in advance: people use plans as resources while responding to circumstances that become visible only in the doing (Suchman 2006). Cooperative-work research similarly shows that distributed work requires ongoing *articulation*: participants manage dependencies, exceptions, responsibilities, and changing conditions through coordination mechanisms that must themselves remain adaptable (Schmidt and Simone 1996).

The alternative is not unstructured autonomy. Coordination can also be carried by a persistent environment. Participants perceive what has already happened, what remains unresolved, what artifacts exist, and which actions are available to them. They act locally within legitimate bounds. Their actions change what later participants can see and do. A centrally assigned workflow may still be used where sequence matters; conversation may still resolve ambiguity; a governed decision path may still control consequences. The claim is a portfolio claim: distinct coordination modes have different domains of advantage and can be composed. **C02** `[E/I]`**.** A serious evaluation must therefore compare against strong orchestration rather than a deliberately weak manager or isolated baseline.

Figure 1. Four complementary coordination modes

DRAFT


Coordination need
Orchestration
Assignment + sequence
Best for tightly coupled work
combine
Messaging
Conversation + negotiation
Best for ambiguity and relationship
combine
Shared state
Persistent artifacts + affordances
Best for turnover and asynchronous reuse
gate
Governed consequential state
Claims + evidence + authority
Required when rights or resources change

DRAFT

Alt text: Four modes branch from a coordination need. Orchestration uses assignment and sequence. Messaging supports conversation and ambiguity. Shared state uses persistent artifacts for asynchronous work and turnover. Governed consequential state adds claims, evidence, and authority when rights or resources change. Dashed links show that the modes combine rather than replace one another. *Source note:* Author's analytical comparison; no mode is claimed to be universally superior.

3. The thesis and its limits

Intelligent voluntary cooperation is a model in which autonomous participants coordinate around shared or compatible intent by perceiving and modifying a persistent, bounded, and governed environment. Useful actions leave attributable traces. Some traces become artifacts: addressable informational, executable, or institutional structures that later participants can inspect, reuse, challenge, revise, or retire. Legitimate decisions can change shared reality only through explicit evidence, authority, and consequence paths.

Three limits follow immediately. First, environment-mediated coordination does not eliminate communication. Discussion, negotiation, explanation, care, and conflict resolution remain human and institutional work. Second, self-organization does not imply voluntariness. A system can lack a visible manager while controlling participants through opaque rankings, defaults, sanctions, information asymmetry, or prohibitive exit costs. Third, persistence does not imply truth. A durable error can coordinate future action just as effectively as a durable fact.

The paper's central proposition is therefore conditional:

> Under appropriate boundaries, autonomous participants can coordinate useful action by perceiving and modifying a persistent shared environment. For that coordination to become voluntary, trustworthy, and consequential in human settings, the environment must also make intent, capability, authority, provenance, evaluation, contestation, and legitimate consequence legible.

Qi and IXO are evaluated against this proposition as a candidate architecture. The standard is not resemblance in vocabulary. The standard is whether verified system boundaries correspond to the theory, whether unimplemented parts are labeled accurately, and whether comparative experiments could disprove the claimed advantage.

DRAFT

Part II — Environmental coordination

4. Stigmergy: coordination through traces

Pierre-Paul Grassé coined *stigmergy* to describe how the product of prior termite activity changes the stimulus for subsequent work. Coordination emerges indirectly: an altered environment helps elicit the next act without a central plan being represented in any individual termite (Grassé 1959). The historical result is narrow. It concerns particular social insects and nest construction. It neither explains human institutions nor establishes that language-model agents should organize in the same way.

The mechanism nevertheless travels. In an early collective-robotics experiment, one to five simple robots clustered initially scattered objects using local sensing and environmental modification. Performance improved from one to three robots, then worsened with four and five as interference grew and existing clusters could be disrupted (Beckers et al. 1994). The experiment established both possibility and limit: local action can produce a global task outcome, but additional actors create coordination costs. More participation is not automatically better.

For digital systems, a trace can be a changed document, an accepted decision, a test result, a dependency link, a reusable program, a status transition, or a visible absence that invites contribution. A later actor need not receive a message from the original actor. It may encounter the transformed environment directly. This is the defensible core of C03: persistent environmental traces can mediate useful indirect coordination in multiple bounded domains. **C03** `[E]`**.** The claim does not say that all coordination is stigmergic, that the effect scales indefinitely, or that any shared database will generate cooperation.

5. Cognition distributed across people and artifacts

Environmental coordination becomes more relevant to human systems when paired with distributed cognition. Edwin Hutchins's analysis of cockpit speed memory showed that a cognitive accomplishment can belong to a socio-technical system comprising pilots, procedures, instrument arrangements, and representational transformations. Information is propagated and changed across people and artifacts; the successful result cannot be inferred by looking only inside one person's head (Hutchins 1995).

DRAFT

This is more than “external memory.” A representation can constrain attention, expose a discrepancy, sequence a handoff, or make an otherwise invisible relation available for action. Cooperative-work studies describe artifacts that coordinate distributed activities by representing dependencies, commitments, standards, and exceptions. Schmidt and Simone argued that computational coordination mechanisms should be malleable and combinable because real cooperative work is contingent (1996). The artifact participates in how work is organized, even though it has no independent intent.

The same idea clarifies an error in agent-system design. If coordination state lives mainly in transient conversations or hidden model context, the collective system forgets when sessions end, models change, or participants leave. A longer prompt can preserve more text but does not necessarily create a legible environment. Durable collective cognition requires structures that actors can locate, interpret, and legitimately modify. It also requires active maintenance: representations become stale, inconsistent, or misleading when the world changes.

6. Artifacts as inherited capability

Persistent artifacts are a useful unit of collective intelligence because they can preserve both knowledge and executable possibility outside any one actor. **C05** `[I]`**.** A written method preserves a procedure; a schema preserves a distinction; a Flow preserves an executable arrangement; a policy preserves a decision rule; a content-addressed program preserves a capability lineage. Artifacts can survive actor turnover and lower the cost of beginning again.

Controlled human research gives one reason to care about the topology of artifact exposure. In a computer-based cultural-evolution experiment, partially connected groups preserved more diversity and accumulated more complex solutions than fully connected groups on the studied search landscape (Derex and Boyd 2016). Full connectivity accelerated convergence but could reduce the exploration needed for cumulative improvement. The result is task-specific, yet it warns against an assumption that universal visibility or immediate consensus is always optimal.

Agent research provides another bounded example. Voyager stored, retrieved, composed, and refined a growing library of executable Minecraft skills. The library transferred to new worlds and supported progress without changing the model’s

weights (Wang et al. 2023). This was one agent in a game, not a society and not an independently verified production system. Still, it demonstrates that executable artifacts can serve as procedural memory.

TerraLingua extended the idea to simulated populations of finite-lived language-model agents. Its authors reported persistent textual artifacts, cooperative norms, division of labor, governance attempts, and branching lineages across generations (Paolo et al. 2026). The recent preprint depends partly on interpretive analysis and does not validate real institutions. Together, these studies justify a research question rather than a conclusion: under what conditions does externalized, addressable capability become inheritance rather than clutter?

7. Affordances: what this actor can do here

James Gibson used *affordance* for possibilities for action that arise in the relation between an organism and its environment (1979). The same surface can afford different action to differently situated actors. Extending the idea to a digital cooperative environment is an inference, but a useful one. A generic task list says what somebody might do. An actor-specific affordance says what *this* participant can legitimately and practically do given current state, capability, authority, dependencies, and risk.

That distinction could reduce central assignment. A person with the relevant knowledge might see a disputed assumption. An evaluator might see an evidence bundle awaiting review. A signing agent might see an authorised transaction preview but not the private key. A new contributor might see an unclaimed research gap rather than an undifferentiated history. The environment becomes a field of possible contribution.

This is C07, and it remains a hypothesis. **C07** `[I/H]`**.** Personalization may merely conceal centralized allocation behind an algorithm. It may entrench incumbent reputation, exclude unfamiliar contributors, or invite actors toward actions that benefit the platform rather than the cooperative intent. The comparative test must hold participant capability and permission constant while comparing central assignment, generic task visibility, and capability-aware affordances. Useful uptake, unauthorized attempts, distribution of valued contributions, and participants' understanding all matter.

Environmental coordination is therefore not storage plus search. It is the continual relation among bounded intent, legible state, actor-specific possibility, and governed change.

DRAFT

Part III — Agent evidence and SwarmWorld

8. Why SwarmWorld matters

SwarmWorld is unusually relevant because it isolates a mechanism close to the one proposed here: language-model agents perceive a local shared world, act through constrained capabilities, leave persistent artifacts, and encounter artifacts left by others. The study is a preprint submitted on 26 August 2026, not a peer-reviewed consensus result. Its importance lies in the design and the observed pattern, not in settling a field four days after publication (Pal et al. 2026).

Initially equivalent agents used one model and configuration— `gpt-5.6-luna`, temperature 0.7, with low reasoning effort—and received identical prompts, schemas, capabilities, inventories, and memory limits. They observed local state and private memory rather than the full global world. The language model proposed bounded, schema-constrained plans. A deterministic simulator, not the model, checked permission, location, materials, energy, and action constraints and determined the consequences.

Persistent artifacts could include executable controllers, but these were deliberately restricted: deterministic, capability-scoped sequences of between one and sixty-four straight-line instructions, without loops, calls, imports, network access, file access, or dynamic code execution. This separation matters. The model supplied cognition and invention inside a sandbox; the simulator remained the authority for what could happen.

The primary population study ran for 800 ticks across populations of 50, 100, and 200 agents. It crossed four mechanism conditions and used four matched world seeds per cell with eight held-out disturbance schedules. A long-horizon study ran populations of 100 for 3,200 ticks. The conditions distinguished full culture, no communication, no explicit culture with physical stigmergy only, and an endpoint-wise best-of-N isolated-search envelope. The isolated comparator was stronger than a single isolated agent because it selected the best result for each endpoint from matched individual searches.

DRAFT

At evaluation checkpoints, the researchers copied and froze the world, removed the agents, and ran deterministic physics and installed programs for 288 further evaluation ticks under unseen disturbances. An artifact that continued to function in that test demonstrated persistence within the simulator. It did not demonstrate a real autonomous infrastructure or service surviving its operators.

9. What the study found

Within those conditions, shared worlds generally produced broader and more resilient technology portfolios and more validated inventions than the isolated-search envelope. Isolated search retained the strongest final single artifact in the long-horizon comparison. This distinction prevents a common misreading. The value of the shared world appeared in the repertoire and ecology of capability, not universal dominance at every endpoint.

The study also recovered differentiated behavior from label-blind, post-hoc analysis. Broad artifact-centered and mobile-exploration phenotypes appeared alongside dynamic constructor/operator, caretaker, coordinator, and surveyor states. These were not assigned jobs, stable identities, or roles declared in the agents' prompts. They were analytical classifications of longitudinal behavior. Saying that agents "became builders and coordinators" is acceptable only with that qualification.

Artifacts often recorded contributions from more than one agent: 67 percent, 76 percent, and 56 percent at populations 50, 100, and 200 in the full-culture condition. Executable lineages accumulated, with mean maximum depth reaching 9.75 at 3,200 ticks. Noncreator reuse was extensive: 99.3 percent of full-culture artifacts and 96.9 percent of no-explicit-culture artifacts were reused by another agent. Approximately 95 percent of **first reuse** began through direct physical observation in both conditions. It is not correct to turn that statistic into "95 percent of all knowledge transfer occurred without communication." It describes how the first reuse episode began under the study's classification.

Randomly removing half the agents left most artifacts connected to a surviving contributor, while targeted removal of hubs or brokers produced much larger topological losses. This was a graph-connectivity assay. It did not run the ecology forward to measure functional service recovery after an attack. The result supports concern about broker concentration, but not a numerical claim about operational resilience.

DRAFT

These are the grounds for C04. **C04 `[X]`.** Under one controlled simulation and model regime, persistent shared state and executable artifacts supported differentiated behavior, broad reuse, technological lineages, and portfolio-level advantages over matched isolated search. That is a meaningful mechanism result. It is not evidence that Qi already exhibits any of those outcomes.

10. What the study does not establish

The limits are material. There were four seeds per condition and one model/prompt regime. The world's material functions were normalized simulator constructs rather than physical measurements. Metabolism, mortality, and replacement were disabled in the reported population study even though persistence and inheritance were central concepts. The 288-tick artifact test occurred in copied deterministic worlds. The Protein Realms extension was descriptive and single-seed. Illustrated technology portraits were representations, not manufactured objects and not necessarily literal simulator geometry.

Most importantly, simulated agents were initially homogeneous and operated under rules chosen by researchers. Human cooperation introduces unequal rights, histories, needs, legal protections, embodied risk, identity, care, and coercion. Organisations introduce ownership and fiduciary duties. Devices introduce safety and control constraints. An environment that elicits useful agent specialization may still be illegitimate for people.

Explicit culture was not superior on every endpoint or timescale. Physical observation did much of the initial reuse work in both full-culture and no-explicit-culture conditions. The strongest single artifact sometimes came from isolation. These findings suggest that communication, social memory, environmental exposure, and individual search are complementary mechanisms whose value depends on the desired outcome.

SwarmWorld therefore changes the research prior, not the product status. It makes persistent shared worlds a credible object of study for language-model societies. It gives comparative measures—portfolio breadth, lineage depth, noncreator reuse, multi-contributor artifacts, post-hoc behavioral states, disturbance performance, and broker vulnerability—that can inform Qi experiments. It does not license a present-tense claim of emergent culture, voluntary cooperation, compounding network effects, or defensibility.

DRAFT

Part IV — A theory of intelligent voluntary cooperation

11. The bounded cooperative world

The proposed unit is a **bounded cooperative world**: a persistent environment organized around declared intent, identifiable participants, legible state, possible actions, and completion or continuation conditions. “World” does not imply a virtual landscape. It means that participants inherit a structured situation rather than merely a transcript.

At minimum, the world needs nine primitives:

1. **Intent:** the desired future condition and the reasons cooperation is being invited.
2. **Participants:** identifiable actors and the capacities in which they participate.
3. **State:** current accepted, disputed, unknown, and proposed representations.
4. **Affordances:** actions available to a particular actor given capability, authority, and context.
5. **Artifacts:** persistent informational, executable, or institutional traces with provenance and status.
6. **Evaluation:** methods for distinguishing assertions from findings accepted for a purpose.
7. **Consequence:** controlled paths through which a decision can affect rights, resources, obligations, or authoritative state.
8. **Governance:** mechanisms for membership, rule change, monitoring, challenge, correction, sanction, and appeal.
9. **Exit:** practical refusal, withdrawal, portability, and the handling of legitimate prior commitments.

C08 proposes that these form a useful minimum for consequential heterogeneous cooperation. **C08** `[I/H]`**.** It is not a metaphysical list. Formal modeling and component ablation should test whether fewer or different primitives perform as well.

The bounded world also separates three layers often conflated in collaboration tools. A conversation contains utterances. A cooperative environment contains projected and governed state. A consequential system contains authoritative changes with

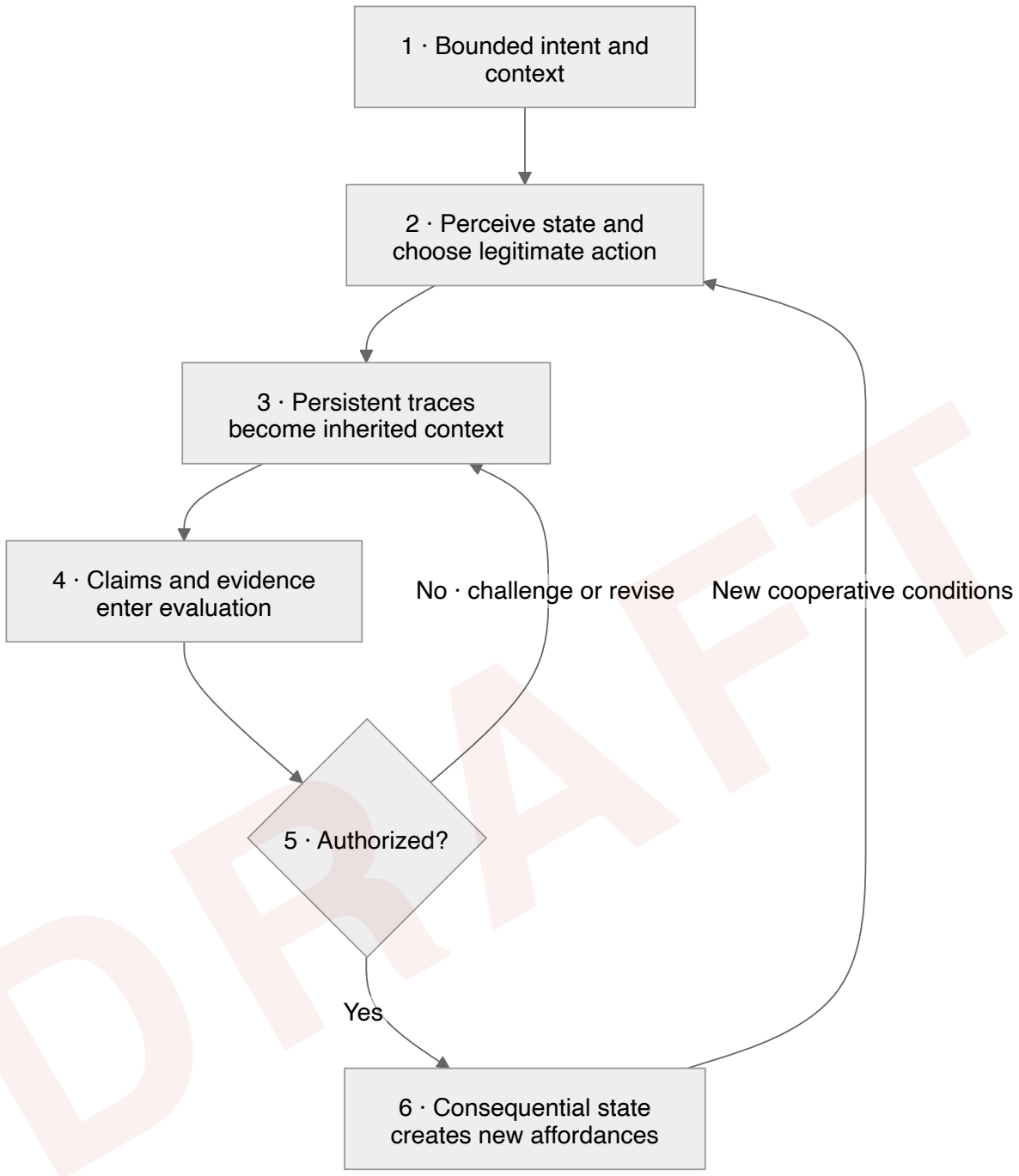
effects outside the conversation. A message may propose a fact; an accepted Overview may represent the group's current understanding; a chain transaction may alter a right or release a payment. Each deserves a separate identity and lifecycle.

12. The causal loop

The theory can be expressed as a feedback loop. Intent and context shape what participants perceive. Participants encounter current state and actor-specific affordances. They choose legitimate action. Action leaves traces. Those traces alter what later actors can perceive and do. Claims and evidence are evaluated under explicit methods. Authorized decisions create consequential state. The new state exposes new opportunities and reusable structures.

Figure 2. The intelligent voluntary cooperation causal loop

DRAFT



Bounded intent leads participants to perceive state and choose legitimate action. Persistent traces become inherited context. Claims and evidence enter evaluation. Rejected claims return for challenge or revision; authorized decisions change consequential state and create new affordances.

Source note: Author's synthesis; proposed theory, not a demonstrated Qi outcome.

Every arrow can fail. Intent may be vague or imposed. State may be incomplete. Affordances may be manipulative. An action may exceed authority. A trace may be false or private. Evaluation may be circular. A decision may lack legitimacy. A

transaction may fail. New state may amplify a historical error. The theory is useful precisely because it makes those failure points inspectable.

13. Voluntary means more than unassigned

Removing a manager does not create free choice. The Belmont Report's treatment of informed consent—information, comprehension, and voluntariness—comes from human-subjects research, not general digital governance, but it clarifies why a click is insufficient (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979). Hirschman's distinction among exit, voice, and loyalty shows that leaving and changing an institution are interacting responses, neither a complete substitute for the other (Hirschman 1970). Aghion and Tirole distinguish formal authority from real authority produced by information and initiative (1997). A person may hold the formal right to approve while an agent controls the options, evidence, timing, and default.

Platform-labor research makes the problem concrete. Studies of ride-hailing described systems that advertised flexibility while exercising indirect control through work allocation, information asymmetry, evaluation, and sanctions (Lee et al. 2015; Rosenblat and Stark 2016). The lesson is not that all digital coordination is employment. It is that power can be encoded in ranking, visibility, default, and dependency even without explicit orders.

For this paper, cooperation is voluntary only to the degree that a participant can understand the invitation, identify who benefits and who decides, refuse or limit a contribution, see the relevant authority, contest shared state and decisions, and leave without undisclosed retaliation or impossible lock-in. Existing obligations can legitimately survive exit when they were clear and lawful. Vulnerability, dependency, accessibility, and bargaining power affect whether nominal options are practical.

C06 is therefore both inference and normative commitment. **C06** `[I/N]`**.** Its operational test is not a decentralization score. It is whether participants understand the scope, accurately predict consequences, exercise refusal without hidden execution, influence rules and corrections, and retain meaningful alternatives. If the architecture cannot make those conditions visible, it should avoid the word *voluntary*.

14. Cooperation is not one economic category

A shared digital environment is not automatically a commons. Samuelson's public goods, Buchanan's clubs, Ostrom's common-pool resources, Coase's firms, and market exchange describe different problems. Excludability and subtractability concern the resource; ownership and membership concern a property regime; messaging, pricing, voting, orchestration, and stigmergy are coordination mechanisms. Conflating them produces weak governance.

A public research note may be non-rival and difficult to exclude. A private credential is excludable and sensitive. A shared compute pool is subtractable. A professional service may be a club good. A Flow template may be openly copyable while its execution depends on scarce attention, licensed skill, or funded authority. A payment claim is neither a commons nor a message: it is an assertion inside a governed economic mechanism.

The classification affects incentives. Open contribution can benefit from broad discovery and attribution. Scarce resources require allocation and congestion control. Team production can create free riding and monitoring problems (Holmström 1982). Sanctions can sustain cooperation in some public-goods experiments but impose costs and can change social meaning (Fehr and Gächter 2000). A fine can be interpreted as a price and increase the behavior it was intended to deter (Gneezy and Rustichini 2000). Incentives must therefore be designed for the resource, institution, and participant relationship in question, not added as a universal token layer.

15. Hybrid coordination by consequence and coupling

The theory does not seek to replace workflows. It asks the environment to choose or combine coordination modes based on the structure of the work. High coupling and known sequence favor orchestration. Ambiguity and relationship favor conversation. Turnover, exploration, and asynchronous reuse favor persistent shared state. Rights, money, credentials, and irreversible external effects require governed consequential state.

A single effort may move through all four. Participants discuss an uncertain goal. Shared artifacts expose research gaps and allow voluntary contribution. Once a design stabilizes, an orchestrated Flow executes a controlled sequence. Claims

about completion enter evaluation. An authorized decision initiates settlement. New receipts and outcomes return to shared state. The modes form a layered system rather than competing ideologies.

This hybrid stance creates a measurable prediction: the relative advantage of a mode should interact with task coupling, uncertainty, turnover, consequence, and accountability. If it does not, the classification is not useful. The experimental annex specifies matched crossover studies to test the interaction rather than infer it from selected demonstrations.

16. What makes the cooperation intelligent

The adjective *intelligent* should not be reduced to “uses an AI model.” A shared environment is intelligent to the degree that it helps participants perceive relevant state, discover legitimate possibilities, anticipate dependencies, learn from consequence, and revise its representations. A paper form can participate in intelligent coordination; a powerful model with no durable context can fail to do so. Intelligence is a property of the actor–artifact–institution system as well as of individual actors.

Four capabilities matter. First, the environment must select and present context without concealing the basis of selection. Relevance ranking can reduce cognitive load, but it also allocates attention and creates real authority. Participants need routes to inspect source state, encounter dissent, and change filters. Second, the environment must translate between representations: conversation into candidate records, evidence into evaluation inputs, decisions into authorized actions, and receipts into updated shared state. Each translation should preserve provenance and uncertainty.

Third, the environment must support learning without silently rewriting institutional commitments. An agent may infer a better method; a group may test it; an authorized process may adopt it. The system should not allow a model update or hidden prompt change to alter the meaning of an accepted contract. Adaptive cognition and governed rule change have different authorities and review requirements.

Fourth, the environment must know when not to automate. Uncertainty, novelty, disagreement, sensitive data, vulnerable participants, and irreversible consequence are reasons to pause or change coordination mode. The relevant intelligence is not

maximum autonomous action. It is appropriate movement among exploration, conversation, orchestration, evaluation, and human or institutional decision.

These properties connect distributed cognition to governance. A representation that makes an anomaly visible improves collective perception. A capability-aware affordance can lower the cost of contribution. A challenge record can prevent false consensus. A read-back can correct a mistaken local belief. Intelligence in this sense is the environment's capacity to improve situated judgment while preserving the participant's agency and the institution's legitimate authority.

DRAFT

Part V — Governance and trustworthy cooperation

17. Persistence is power and risk

A trace coordinates later action whether it is accurate or false, generous or manipulative. That symmetry makes persistence valuable and dangerous. **C09 `[E/I]`.** A poisoned instruction in a retrieved page can cross from passive-looking data into a model's behavior. Indirect prompt-injection work demonstrated this path in early LLM-integrated applications (Greshake et al. 2023). AgentPoison reported high targeted attack success after corrupting less than 0.1 percent of stored memory or knowledge in its tested settings, while ordinary task performance remained almost unchanged (Chen et al. 2024). PoisonedRAG similarly showed that small numbers of attacker-controlled texts could dominate retrieval-grounded answers in tested corpora (Zou et al. 2025).

These studies are not tests of Qi. They establish a threat class. A persistent shared environment is an information supply chain. Content can be indexed, ranked, summarized, retrieved, transformed, and inherited. Each stage can amplify a malicious or merely stale artifact. High aggregate accuracy does not rule out a triggerable compromise, and an authorized read does not make the content an instruction.

Distributed-systems properties must also remain narrow. A CRDT can ensure that replicas receiving the same updates converge under stated conditions; it does not make those updates true or safe (Shapiro et al. 2011). An append-only log can make inconsistency or omission detectable under a proof and observation model; it does not prevent semantic error (Laurie et al. 2021). Provenance can represent asserted derivation and responsibility; it does not guarantee that the asserted history is complete or honest (W3C Provenance Working Group 2013).

The minimum response is typed distrust. Shared content is untrusted input. Instructions, evidence, policy, and executable actions occupy different lanes. Artifacts carry provenance, version, status, scope, and freshness. Consequential tools require separate authority. Corrections supersede rather than silently erase history. Dependent conclusions are re-evaluated when their support changes. Retirement, expiry, revocation, suspension, and deletion remain distinct operations.

DRAFT

18. Correction, decay, and epistemic maintenance

Persistent environments need a maintenance theory as much as a creation theory. New state is cheap to add and politically difficult to retire. Contributors may be rewarded for publication while nobody is rewarded for checking whether a claim remains current, a dependency still works, or an artifact should be removed from discovery. Without explicit maintenance, successful accumulation produces its own failure condition.

Correction should preserve two truths at once: what the environment previously represented and what participants should rely on now. Silent mutation destroys accountability; immutable prominence preserves harm. A supersession relation can retain lineage while moving the current view to a corrected record. Invalidation can record that a result should no longer support downstream decisions. Expiry can require renewed evidence without declaring the old statement false. Suspension can pause reliance during review. A tombstone can preserve identity and deletion context without retaining the original public content.

Dependency matters. If an accepted conclusion depends on evidence later invalidated, the environment should identify affected artifacts, evaluations, and decisions. Jon Doyle's truth-maintenance work treated belief revision as management of justifications and dependencies rather than isolated replacement (1979). The architectural inference is straightforward: provenance should be executable enough to support impact analysis. Correction is incomplete if downstream actors continue to see a derived result as current.

Decay should be risk- and domain-specific. A legal authorization, current exchange rate, safety certificate, research finding, personal preference, and historical event have different freshness requirements. Recency alone is not quality; old evidence may remain valid, and recent evidence may be weak. The environment needs explicit review intervals, source availability, method versions, and event-triggered revalidation rather than one global time-to-live.

Maintenance work is itself a contribution and should be visible in governance and incentives. Reviewing, disputing, retiring, migrating, and documenting negative results may create more collective value than producing another artifact. A cooperative economy that rewards only creation will predictably accumulate unowned risk.

19. Polycentric governance, not automatic decentralization

Elinor Ostrom's work on common-pool resources showed that neither centralized state control nor privatization is universally sufficient. Durable self-governance appeared under recurring institutional conditions: boundaries, rules fitted to local conditions, participation in rule modification, monitoring, graduated sanctions, accessible conflict resolution, recognition of organizing rights, and nested organization at larger scales (Ostrom 1990). Her later work emphasized polycentric systems with multiple formally independent but interacting decision centers, enabling experimentation and learning while also introducing inconsistency and coordination cost (Ostrom 2010).

The responsible inference is that a cooperative environment needs governance close to the affected context and coordination across contexts. A Topic-scale group may decide local membership, accepted state, evidence method, or contribution rules. Higher-risk actions may require external professional, organizational, legal, or network authority. Appeals may cross the local boundary. Shared standards can enable portability without imposing one universal institution.

Cryptographic decentralization does not create this structure. A token vote can be captured. A nominally local decision may depend on one infrastructure provider. A multisignature can concentrate real agenda control in the party that prepares the transaction. A visible human approver can become a rubber stamp. Polycentricity must be assessed by who sets goals, sees information, controls admission, changes rules, selects evaluators, executes consequences, bears liability, and can correct the result.

20. Identity, Sybils, reputation, and plural participation

Persistent contribution histories can help participants assess reliability, but identity and reputation create their own power. A DID can support proof of control and verification relationships; it does not prove unique personhood, civil identity, competence, or benevolence (W3C Decentralized Identifier Working Group 2022). Douceur's Sybil analysis showed why multiplicity of cheap identities can defeat redundancy and reputation absent trusted identification or sufficiently constrained

assumptions (2002). Defenses based on cost, social graphs, credentials, or institutional admission each encode trade-offs and can exclude legitimate low-resource or unfamiliar participants.

Visible ratings are also performative. In a large randomized experiment, an initial positive social signal created a persistent upward bias in later ratings, while negative manipulation was more likely to be corrected (Muchnik et al. 2013). History can coordinate judgment while also manufacturing apparent consensus. Reputation should therefore be contextual, time-bounded, contestable, and explicit about uncertainty. It should not silently become a universal quality score.

C14 remains a hypothesis: evaluated contribution histories may improve later decisions and cooperation. **C14** `[H]`**.** The falsifier is not only “no predictive lift.” It includes discrimination, lock-in, strategic gaming, deterrence of new entrants, or a distribution of benefits and burdens inconsistent with the cooperative purpose.

21. Meaningful human authority

“Human in the loop” is not an adequate governance design. Human–AI studies show that explanations can increase reliance without producing complementary team performance (Bansal et al. 2021). Cognitive forcing functions can reduce overreliance while adding effort and unevenly benefiting participants (Buçinca et al. 2021). Algorithm-in-the-loop experiments have found settings in which interaction worsened accuracy or disparity (Green and Chen 2019). Field evidence also suggests a jagged frontier: AI can improve speed and quality on tasks within its capabilities while reducing performance when users rely on it beyond that frontier (Dell’Acqua et al. 2026).

A system must therefore identify at least seven roles: who defines the goal, who proposes, who decides, who authorizes, who executes, who evaluates, and who bears liability or can correct the result. One actor may occupy several roles, but the combination should be visible. Human authority is meaningful when the person has usable information, time, alternatives, competence, and a real ability to change or refuse the outcome. A final click after the agent has framed the problem, selected evidence, narrowed options, and created urgency may preserve formal authority while transferring real authority.

The same concern applies between people. An employer, sponsor, community leader, platform, or majority may use an “open” cooperative system to allocate visibility and opportunity. Voluntariness requires power visibility and effective contestation, not only a technically valid grant.

22. Governance requirements

The institutional minimum follows from these risks:

- boundaries of membership, content visibility, resource access, and consequence;
- least-authority grants with expiry, revocation, replay protection, and confirmation for high-risk action;
- provenance, status, challenge, correction, supersession, and retirement for persistent artifacts;
- independent or plural evaluation where the consequence warrants it;
- conflict-of-interest and evaluator-selection rules;
- monitoring proportionate to risk, with governance cost measured rather than hidden;
- dispute, appeal, and remedy that can alter future use of the contested state;
- protection of private and personal state from silent conversion into shared context;
- portable artifacts and identities where exit would otherwise be nominal;
- incident, recovery, reconciliation, and kill-switch procedures for consequential paths.

These are not features that can be inferred from a blockchain, a chat room, or a multi-agent demo. They are acceptance conditions for describing a system as trustworthy cooperation.

DRAFT

Part VI — Cognition and consequence

23. The reality boundary

Language models are powerful generators of candidate interpretations and actions. Their fluency creates a dangerous compression: an assertion sounds like a finding; a recommendation appears as a decision; a tool's success response looks like a completed outcome. A trustworthy architecture must expand those collapsed steps again.

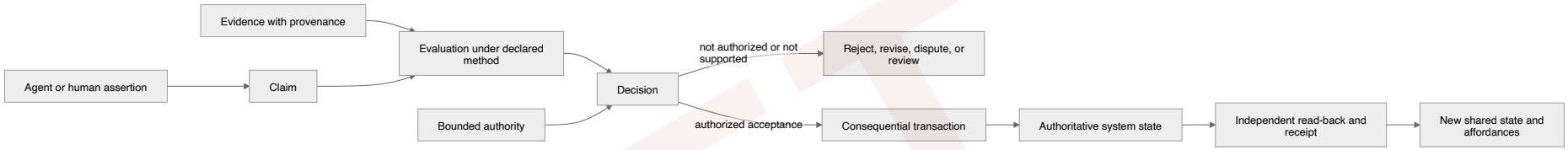
The proposed **reality boundary** separates:

- an **assertion**, which any actor may express;
- a **claim**, which identifies what is being asserted under a schema or purpose;
- **evidence**, whose source, collection mode, integrity, relevance, and limits can be assessed;
- an **evaluation**, which applies a declared method to claim and evidence;
- a **decision**, made by an identifiable role under applicable policy;
- **authority**, which permits a bounded consequence;
- a **transaction or action**, which attempts that consequence; and
- **authoritative state plus read-back**, which shows what actually changed.

C10 proposes that this separation can keep cognition productive without giving it automatic consequence. **C10** `[I/S]`**.** Components of the pattern exist in standards and IXO systems at different maturity levels; an end-to-end production guarantee is not asserted.

Figure 3. The claim-to-consequence reality boundary

DRAFT


Evidence with provenance
Evaluation under declared method
Agent or human assertion
Claim
Decision
not authorized or not supported
Reject, revise, dispute, or review
Bounded authority
authorized acceptance
Consequential transaction
Authoritative system state
Independent read-back and receipt
New shared state and affordances

DRAFT

Alt text: An assertion becomes a claim. Evidence with provenance joins it in an evaluation under a declared method. A decision also requires bounded authority. Unsupported or unauthorized decisions are rejected, revised, disputed, or reviewed. Accepted decisions may initiate a transaction; only authoritative system state and independent read-back establish the consequence. *Source note:* Author's synthesis of W3C standards and verified IXO component semantics. The components exist at different maturity levels; one fully conformed production path has not been established.

24. Cryptographic integrity is not semantic truth

W3C Verifiable Credentials distinguish issuer, holder, subject, and verifier and allow tamper-evident claims with proof, status, evidence, and validity data. The specification is explicit in substance: cryptographic verifiability does not make the claims true; a verifier must apply policy before relying on them (W3C Verifiable Credentials Working Group 2025). A genuine credential can contain a mistaken claim. A trusted issuer can exceed its competence. A correctly verified signature can be used for the wrong purpose.

Likewise, provenance is an account of derivation, not an oracle. A content hash binds bytes, not interpretation. Consensus establishes agreement under a protocol, not universal legitimacy. Authorization says an actor may attempt an action under a scope; it does not say the action is correct. The reality boundary combines these narrow guarantees rather than asking one mechanism to perform every trust function.

25. Evaluation and independence

An evaluation is only as credible as its method, evidence, evaluator authority, and independence. Model self-evaluation can be useful for catching format errors or clear contradictions, but it is not independent verification. Client-assisted evidence can preserve provenance and hashes while remaining selected by the claimant or client. A deterministic rubric can improve reproducibility while encoding a poor criterion. Multiple evaluators can reduce individual error while sharing one corrupted source or incentive.

The consequence should determine the evidence burden. A reversible draft edit may need provenance and ordinary review. Publishing a professional blueprint may require owner and independent reviewer approval plus verified artifact digests.

Releasing a payment may require a domain-specific claim protocol, authenticated evidence, an authorized evaluation, funding, dispute rules, signing, transaction confirmation, and read-back. A high burden everywhere destroys initiative; a low burden at the reality boundary creates harm.

26. Receipts and authoritative read-back

Systems often stop at a local receipt: “submitted,” “signed,” “sent,” or “completed.” Those words describe a component’s action, not necessarily the external result. A reliable consequential path binds the requested change to the authority, intended target, parameters, idempotency key, transaction identity, and final state observed from the authoritative system. Retries must not rerun a side effect merely because writing the receipt failed.

Certificate Transparency illustrates the narrow value of verifiable logs: inclusion and consistency proofs make certain failures detectable, but the log does not prevent misissuance and split views require observation across parties (Laurie et al. 2021). The analogous principle here is independent confirmation. A writer’s success message is evidence about the writer. Read-back is evidence about the system that owns the consequence.

The final rule is simple: cognition may be plural, exploratory, and generative; consequence must be bounded, attributable, evaluated in proportion to risk, and confirmed where authority actually lives.

DRAFT

Part VII — Qi as a candidate cooperative environment

27. What is being claimed

Qi is not treated here as a product that has already demonstrated intelligent voluntary cooperation. It is treated as an evolving architecture whose verified components correspond to parts of the theory. That distinction changes the argument from “research proves Qi” to “research identifies mechanisms and risks against which Qi can be specified and tested.”

The strongest current statement is C11: the Topic Protocol specifies a Qi Topic as a durable context for intent, state, authority, collaboration, and outcome, and a Topic can therefore be interpreted as a candidate bounded cooperative environment. **C11** `[S/I/H]`**.** The system fact is narrower than the interpretation. The cited Topic Protocol is version `1.0.0-rc.1` and explicitly calls itself a **draft implementation candidate**. Its repository is a protocol baseline, not proof that live runtime and security gates have passed (IXO 2026c).

This section uses the following source locks, all verified on 30 August 2026:

- Topic Protocol commit `5cebe6437e…`;
- IXO Portal `develop` commit `2488f83669…`;
- Qi Runtimes `main` commit `a2b8cc77a68…`.

Repository evidence demonstrates source behavior and explicit specifications. It does not by itself establish what is deployed in every production environment. The public correspondence matrix preserves that additional boundary.

28. Topic identity and bounded context

Topic Protocol gives a Topic a stable UUIDv7 identity independent of a Matrix event identifier, model session, or chat transcript. A root and anchor connect that identity to a Matrix room, and ordinary native `m.thread` relations carry the conversation. This matters because a durable cooperative object cannot have the lifecycle of one model session or depend on one provider’s hidden context.

Matrix remains the membership, federation, confidentiality, and encryption boundary. A Topic can restrict action authority, but it cannot hide content from other members of the same room. A smaller confidential audience requires a separate encrypted room. That explicit limitation prevents a capability token from being misdescribed as content privacy.

The protocol projects Topic state deterministically from the root and authorized operations. It defines a singular accepted Overview, conflict handling, statuses, typed records, and artifact references. Personal follow, mute, pin, and unread flags remain user state rather than shared Topic truth. Topic, contract, outcome, evaluation, claim, approval, and settlement lifecycles remain distinct. A resolved conversation is not automatically an achieved outcome; an achieved outcome requires its own accepted record.

These are strong specifications for legibility and semantic separation. The conformance document is equally important. A local static pass proves schema, checksum, and reference-vector compatibility. It does not prove live Matrix authority, E2EE behavior, federation, private-resource access, human review, signing, payment authority, or pilot readiness. Cold-cache discovery, encrypted aggregation and migration, Application Service agent security, UCAN replay and revocation, external-work reconciliation, and settlement-controller safety have dedicated gates.

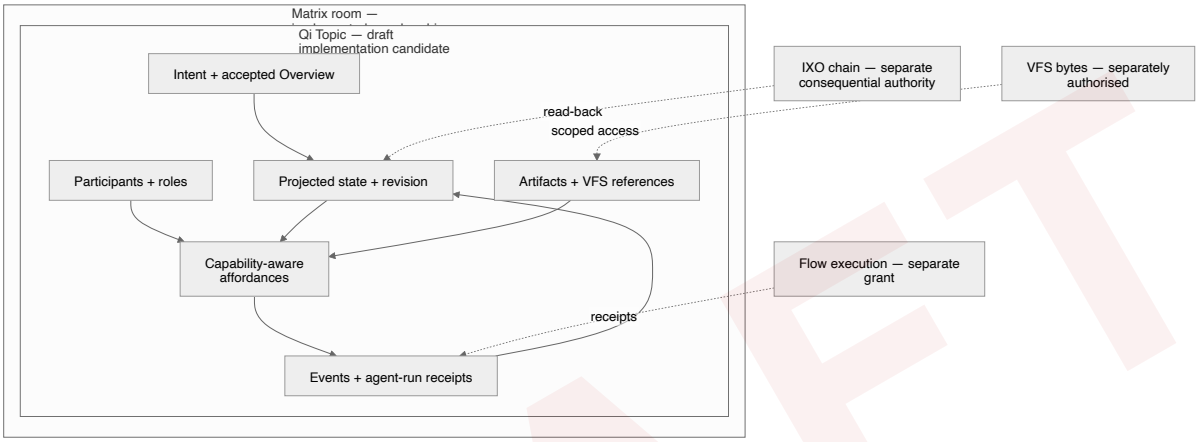
29. Capsules, artifacts, and agent participation

The Topic Capsule is specified as a portable participant view containing a manifest, projected state, anchor, participants, context, artifact references, accepted memory, events, capabilities, policy, and revision. It can let a newly arriving human or agent inherit a bounded situation rather than reconstruct it from an entire history. But a Capsule does not grant access to the file bytes named by a VFS reference, and an agent must handle a stale revision rather than assume its context is current.

Agent participation is similarly explicit. The protocol specifies identifiable agents and run records with Topic, revision, caller, start and end times, status, outputs, and receipts. Subagents inherit only authority that is expressly delegated. Tool calls should record their capability, idempotency key, and receipt. Successful execution cannot implicitly promote any Topic, contract, outcome, claim, approval, or settlement lifecycle.

Topic-scoped UCAN abilities provide the intended action boundary. Flow and Topic grants are separate. VFS access is separate again. High-risk operations require human confirmation. These are specified controls, not yet a claim that every live client rejects replay, handles revocation, protects E2EE agent keys, and recovers a durable outbox without repeating external effects. The protocol's honesty about those open gates is a strength of the evidence posture.

Figure 4. A Topic as a bounded cooperative environment



Alt text: Inside a Matrix room boundary, a Topic contains intent, participants, projected state, artifact references, affordances, and event or agent-run receipts. These form a feedback loop. VFS bytes, Flow execution, and IXO chain state sit outside and connect through separate grants, receipts, and authoritative read-back. *Source note:* Matrix/Yjs collaboration is implemented; Topic Protocol is a draft implementation candidate; the cooperative-world interpretation is proposed.

30. Portal's implemented shared-state surface

At the cited Portal commit, the editor implements collaborative documents and Flows using BlockNote and Yjs. A Matrix provider synchronizes CRDT updates and periodic snapshots so collaborative state can rehydrate across sessions. Template mode supports authoring reusable workflow blocks; Flow mode supports participant interaction and runtime state. The repository contains action registries, delegation and invocation structures, and Flow execution surfaces (IXO 2026a).

This is present implementation evidence for persistent collaborative artifacts. It is not yet evidence that the Topic Protocol's full root, indexing, Capsule, agent, replay, privacy, and action-receipt conformance profiles have passed in production. It is also not evidence of emergent specialization. The same Yjs persistence that supports

collaboration creates a poisoning and correction burden: bad writes can replay, private state must stay outside shared maps, and irreversible external effects cannot be coupled naively to CRDT delivery.

Portal therefore contributes a practical environmental surface: documents, templates, Flows, Matrix rooms, collaborative state, actions, and visible user interaction. Topic Protocol contributes a stronger bounded-context and semantic contract. The research program asks whether their integration actually lowers coordination cost, survives turnover, and exposes legitimate contribution opportunities.

31. The proof still required

Turning this correspondence into a defensible pilot requires evidence across clients, actors, and failure modes. A Topic created on one client should be completely discoverable on another after cold start and pagination. Projected state should remain consistent after concurrent authorized writes, room upgrade, federation delay, redaction, and device migration. Encrypted indexes should survive key rotation and restore without exposing plaintext at rest. Agent participation should function inside the actual E2EE boundary rather than through a privileged unencrypted side channel.

Authority needs negative proof. A participant with room access but no Flow grant should be unable to execute. A participant with a Topic grant but no VFS grant should see a reference without reading the bytes. A stale Capsule should fail or rebase predictably. Revoked, expired, replayed, wrong-audience, and wrong-resource capabilities should be rejected by the executor, not only by the interface. Subagents should fail when they attempt to exceed explicit inherited authority.

Side effects need recovery proof. Two clients should see action state update in real time. Every supported executor should return a signed receipt bound to action, manifest digest, caller, authority, sequence, and idempotency key. If the external action succeeds but Topic write-back fails, the outbox should retry the record without running the action again. If execution fails or is canceled, no adjacent lifecycle should move implicitly. Privacy tests should show that receipts exclude secret values, hidden reasoning, and content outside the participant’s legitimate scope.

DRAFT

Operational evidence is equally important: observability, incident ownership, rollback, reconciliation, support procedure, key compromise, and an operator kill switch for high-risk paths. These tests do not prove better cooperation. They establish that the environment is safe enough to run the comparative study. Passing conformance and demonstrating product benefit are separate gates.

32. Qi Runtimes and the model-host boundary

Qi Runtimes makes a crucial architectural statement: the host model does not run Qi and never owns IXO blockchain state. Runtime packages form an SDK-like operating layer between a model host and external IXO services. They define declarative capabilities, authority boundaries, workspaces, validation, review, release preparation, and external receipts (IXO 2026b).

The Blueprint Runtime implements versioned workspace state, immutable artifact revisions, hash-chained events, validation, user confirmation, independent review, compilation, registry-backed Flow templates, and publication-receipt semantics. Its documentation draws a precise line: compilation is local preparation and never proves a VFS write, package build, entity creation, transaction, testnet state, or live publication. Network state may be reported only from a completed publication receipt and verification of persisted bytes and identifiers.

This structure supports the paper's separation of cognition and consequence. Codex, Claude, or another host can interview, clarify, draft, and explain. The runtime owns the controlled workspace and release logic. Portal supplies user-facing authentication and approval. A signer and external services own their authorities. The network owns final state. Hidden model reasoning and conversation claims are explicitly excluded from audit evidence.

33. Flows, PODs, and executable inheritance

Portal Flows and Qi runtime packages provide implemented primitives for persistent executable artifacts. Blueprint packages can bind versioned service intent, outcomes, automation boundaries, evidence requirements, skills, Flow templates, governance, risk, and operating material. Runtime validation can reject unresolved actions or unverified readiness claims. A Flow can preserve a sequence and its action contracts beyond the agent that authored it.

This supports only the narrow system component of C12. **C12 `[S/I/H]`.** It remains a hypothesis that Deeds, Flows, PODs, or related artifacts will form cumulative executable culture. Reuse might be shallow copying. Dependencies may make artifacts brittle. A lineage may deepen without improving performance. Standardized portability may reduce proprietary defensibility, which could be socially valuable even if strategically inconvenient. Longitudinal tests must measure verified performance after reuse, maintenance burden, contributor diversity, failure propagation, and safe forking.

The implementation should therefore preserve content addressing, versions, provenance, dependency digests, execution receipts, evaluation history, status, and retirement. It should also preserve diversity: a universally promoted Flow can become a monoculture. The goal is not maximum reuse. It is trustworthy inheritance of useful capability with visible lineage and alternatives.

DRAFT

Part VIII — IXO as a consequential substrate

34. From cooperative representation to authoritative state

A Topic can represent a decision without owning the external system affected by that decision. IXO's role in the proposed architecture is to provide narrower consequential primitives: identifiers and controllers; governed entities and accounts; claims and evaluation states; authorization; escrow, dispute, and payment rules; and consensus state. These primitives can help a cooperative environment cross from proposal to legitimate consequence without making the model or conversation the authority.

This section is locked to IXO blockchain v8.0.0 `main` commit `bfdd934456…`, dated 24 June 2026 and verified 30 August 2026. It also uses Evaluations Engine commit `fb380cd01…` and the explicitly experimental IXO Verify commit `7f269baa4e…`.

35. Identity and entities

The IXO chain's IID module implements creation, resolution, updating, and deactivation of IID documents. The model includes controllers, verification methods and relationships, services, linked resources, linked claims, linked entities, and accorded rights. This supplies persistent identifiers and bounded proof-of-control relationships. It does not supply unique personhood, moral legitimacy, professional competence, or claim truth.

The Entity module creates an IID document, a CW721 NFT representation, a stored Entity record, and controlled entity accounts through a deterministic process. Entities can link resources, claims, services, and other entities. Controllers and account/authz rules govern state changes and delegated transactions. The entity is an authoritative chain object; a linked off-chain file still needs its own integrity, availability, and access rules.

These distinctions support C13: different state classes require different integrity, confidentiality, authority, and consensus mechanisms. **C13** `[I/N]`**.** Room membership, VFS access, Flow execution, entity control, and claim evaluation should not be collapsed into one all-purpose permission.

36. Claims, evaluation, disputes, and payments

The v8 Claims module stores collections, claims, intents, member budgets, performance deposits, and disputes in consensus state. Claim protocols are associated with entities. Agents can be authorized to submit or evaluate under collection constraints. A claim can be submitted, approved, rejected, disputed, invalidated, or flagged; flagged is non-terminal and allows a later authorized evaluator to make a terminal determination.

The module implements configurable payments for submission, evaluation, approval, rejection, and oracle service paths. Claim intents can provide payment guarantees and escrow funds before work. Performance deposits and dispute stakes can gate participation and create economic exposure. Adjudication has defined signers, statuses, and payout behavior. These are implemented module semantics at the cited release (IXO Foundation 2026).

They should not be romanticized. An on-chain approved status is authoritative for the protocol but still depends on the claim schema, evidence, evaluator selection, authorization, funding, and dispute design. Incentives can produce gaming or exclusion. Deposits can deter abuse and legitimate low-resource participants. A majority or whitelisted evaluator can be captured. The reality boundary is trustworthy only to the degree that its institutional assumptions hold.

37. Evaluation services and bounded prototypes

The cited Evaluations Engine branch contains routes, workflows, schemas, and persistence for pulling and ingesting claims, applying evaluation logic, storing decisions, and interacting with external chain and relayer paths. It is implementation evidence for a service component with external dependencies, not enough to assert a fully deployed and independently verified production flow. Environment, signer, funding, owner, relayer, and read-back evidence require technical-owner confirmation.

IXO Verify demonstrates a different slice as an experimental local ledger. It implements durable claims and evidence records, client-assisted evidence envelopes with provenance and hashes, deterministic rubric evaluation, review gates, determinations, canonical-hashed audit events, and dry-run actions. Its documentation explicitly states that client-assisted evidence is not independently

server-verified and that real IXO anchoring, payments, credentials, and external side effects are deferred in the cited version. That limitation makes it useful evidence for semantic discipline, not settlement.

38. The honest end-to-end status

Across these systems, the architecture has real components: collaborative state, protocol contracts, capability boundaries, governed runtime workspaces, claim and evaluation services, chain identity, entities, claims, escrow, disputes, and payments. What the reviewed evidence does **not** establish is one fully conformed production path from a Topic-bound agent action through E2EE participation, independent evidence, replay-safe execution, signing, funded settlement, confirmation, and authoritative read-back.

That end-to-end status is therefore **unknown**, not implemented by inference. Topic Protocol's conformance document expressly keeps mainnet settlement closed until live idempotency, funding, dispute, confirmation, key-rotation, restore, and kill-switch evidence exists. If technical-owner review cannot establish a stronger status, the white paper will retain this limitation.

The consequence is a more credible thesis. Qi and IXO do not need to pretend that every bridge is complete. Their architecture already embodies an important principle: the model host is not system authority; collaborative state is not chain state; a claim is not a decision; a decision is not settlement; and a receipt must be checked where consequence lives.

DRAFT

Part IX — Toward a cooperative economy

39. From isolated services to inherited worlds

Most software economics treats each interaction as a transaction between a user and a service. An agent receives a request, calls tools, returns an output, and loses most of the situated learning when the session ends. Intelligent voluntary cooperation points toward a different unit of value: a governed environment whose useful state survives individual interactions.

That environment can hold more than content. It can retain accepted intent, methods, working artifacts, executable Flows, evidence standards, contribution histories, disputes, corrections, and consequential receipts. New participants do not start with a blank prompt. They enter a world whose possibilities have been shaped by previous work. If the inherited state is legible, current, safe, and portable, the marginal cost of future cooperation may fall.

This is the appeal of cumulative cooperation. A professional service blueprint could preserve not only instructions but evidence requirements, authority limits, evaluation methods, reusable Flow components, and operating lessons. A community managing a resource could retain decisions and challenges without forcing each new member to rediscover institutional history. A maintenance Topic could expose unresolved risks, verified fixes, failed approaches, and current action opportunities to a replacement team.

The economic unit is not merely the artifact. It is the relationship among artifact, provenance, evaluation, context, and consequence. A copied procedure without dependencies may fail. A high reputation without context may mislead. A valuable contribution without portable attribution may be captured. A history without correction may become a liability.

40. Incentives for heterogeneous contribution

Voluntary cooperation needs reasons to contribute, but payment is only one reason and often not the first. Participants may seek income, reciprocity, reputation, learning, mission, belonging, access, professional standing, or the improved shared

resource itself. The architecture should expose the resource and relationship before selecting an incentive.

For scarce professional work, explicit service terms and payment may be appropriate. For public knowledge, attribution, governance voice, and reciprocal access may matter more. For a common-pool resource, monitoring and locally legitimate sanctions may be necessary. For sensitive evidence, privacy and controlled access may outweigh discoverability. For experimental agent artifacts, safe reuse and reproducible evaluation may be the primary reward.

On-chain payments and performance deposits can create credible commitments, but they also redistribute risk and power. Deposits may protect a collection from opportunism while excluding capable participants without capital. Automated evaluation may lower administrative cost while creating a new gatekeeper. A token reward can turn a social norm into a priced behavior. These are design choices requiring distributive analysis, not neutral infrastructure.

41. Humans, agents, organisations, and things

The intended cooperative population is heterogeneous. Humans bring rights, experience, care, tacit knowledge, vulnerability, and accountability. Organisations bring durable commitments, governance, capital, and legal personality. Software agents bring speed, scale, availability, and uneven competence. Devices and physical things bring sensing and actuation inside safety and ownership constraints. Calling all of them “agents” can be computationally convenient and institutionally destructive.

Each actor type needs an explicit participation capacity. A sensor may attest to a reading under a calibration and custody policy but cannot consent. A software agent may propose or execute within delegated authority but cannot absorb legal or material harm. An organisation may hold an account and obligation while decisions are exercised through people and rules. A person may delegate repetitive action without delegating purpose, identity, or liability.

Mixed systems should expose asymmetry rather than simulate equality. A service agent can operate continuously and process more information than a community participant. A funder may control scarce resources. A platform operator may control

DRAFT

visibility and infrastructure. A device owner may be different from the person affected by its data. Cooperative design must map who can speak, who can act, whose interests count, who benefits, and who bears failure.

This is also why emergent specialization in a homogeneous simulation cannot be transferred directly to human-agent cooperation. Useful division of labor may arise, but role allocation interacts with status, compensation, expertise, accessibility, discrimination, and appeal. The research agenda must measure not only aggregate performance but whose contributions become visible and valuable.

42. Compounding utility and defensibility

C15 is the strategic hypothesis that accumulated, governed cooperative state may create compounding utility and defensibility. **C15** `[H]`**.** The proposed mechanism is plausible: each verified artifact, trusted relationship, evaluation method, and consequential history can improve discovery and lower the cost of later work; better outcomes attract more participation; more relevant participation creates more useful state.

But every arrow can reverse. More artifacts can create pollution. More history can create surveillance or lock-in. More reuse can create monoculture. More reputation can deter newcomers. More governance can create bureaucracy. Open standards and portability may allow participants to carry the value elsewhere, reducing platform switching costs. That may weaken proprietary defensibility while strengthening the cooperative ecosystem.

The phrase “the moat is the world the agents inherit” should therefore remain a strategic conjecture, not a scientific conclusion. The legitimate test introduces equivalent new cohorts into mature and empty environments, controls for task difficulty, and measures time-to-value, verified quality, error inheritance, governance cost, portability, and inclusion. If mature state provides no advantage, if gains do not compound, or if pollution dominates, the thesis fails.

DRAFT

Part X — Research agenda and implications

43. A falsifiable program

The paper's value depends on making failure possible. The public experimental annex specifies four coordination conditions: conversation-only, strong central orchestration, persistent shared state, and governed consequential state. It spans tightly coupled execution, open search, long-lived maintenance, evidence-bound decision, crisis response, and commons governance. Human, agent, and mixed cohorts are analyzed separately.

Seven experiments carry the main burden:

1. a coordination-mode crossover tests whether task structure predicts relative advantage;
2. a turnover study tests whether inherited artifacts outperform raw history, curated handover, and clean starts;
3. an affordance study tests self-selected contribution without displayed role labels;
4. an adversarial-state benchmark seeds false, stale, prompt-injecting, and over-authoritative artifacts;
5. a reality-boundary experiment measures consequential false acceptance, false rejection, delay, and replay;
6. a mature-versus-empty study tests compounding utility and information pollution; and
7. a power-and-exit audit measures real rather than nominal voluntariness in mixed human-agent work.

Primary outcomes include verified quality, time, coordination overhead, portfolio breadth, best single result, unauthorized attempts, turnover recovery, reuse lineage, governance cost, contribution distribution, perceived choice, exit cost, and information pollution. Agent studies require multiple seeds and frozen model/tool manifests. Human pilots require independent ethical determination, fair compensation review, data minimization, and usable withdrawal.

DRAFT

Negative results are first-class. The architecture may work only in low-risk maintenance, while orchestration remains superior elsewhere. Governance costs may overwhelm environmental benefits. Human participants may experience affordance ranking as opaque management. Persistent state may propagate more error than value. A narrow result is more useful than a universal claim built on demonstrations.

44. Measurement and public accountability

The research object will change while it is being studied. Models update, system components evolve, participants learn, and governance responds to incidents. Every experiment therefore needs a versioned world: model and tool manifests, prompts, schemas, capabilities, source commits, deployment identifiers, policies, random seeds, evaluator methods, and analysis code. “Qi” is too broad a treatment label unless the exact tested configuration is frozen.

Outcome reporting should retain the distinction SwarmWorld made visible. Portfolio breadth is not the best single result. Topological resilience is not functional recovery. First reuse is not all knowledge transmission. A post-hoc behavioral cluster is not an assigned social role. Static conformance is not live operational evidence. The paper’s claim ledger makes those distinctions release-controlled rather than relying on author memory.

Public accountability also requires recording objections that do not change the text. The comment-disposition register assigns accepted, partially accepted, rejected, or deferred status with rationale. Rejection does not make an objection disappear; it makes editorial judgment inspectable. Factual and security blockers must close before release, while unresolved research questions may be deferred only when the main text states the uncertainty.

After publication, corrections should produce numbered `v1.x` releases and an errata record. A material change to the thesis, system architecture, or evidence classification should become `v2.0`, not a silent replacement at the canonical URL. The DOI record, tagged source, web edition, PDF, annexes, and checksums should make the consulted version recoverable.

DRAFT

45. Implications for builders

For builders, the immediate implication is to treat state design as coordination design. Decide which state is personal, room-shared, artifact-bound, externally authoritative, or public. Give each lifecycle, status, provenance, and correction semantics. Do not place secrets, hidden reasoning, or personal preferences in a shared CRDT. Do not let a successful action automatically complete an outcome or release a payment. Make the current/future boundary visible in interfaces and documentation.

The second implication is to design for turnover. A participant should be able to enter with a bounded Capsule-like context, understand intent and accepted state, locate relevant artifacts, see what it can do, and identify uncertainties without reading every message. The same design should allow an artifact to be challenged, superseded, or retired when its assumptions fail.

The third is to preserve least authority across convenience boundaries. A model host can propose. A runtime can validate and prepare. Portal can mediate informed user approval. A signer can authorize a transaction. A network can accept it. A read-back can confirm it. Shortening this path should never mean silently granting one component the powers of all the others.

46. Implications for partners, researchers, and funders

For partners, Qi should be assessed as infrastructure for governed cooperative environments rather than another conversational agent. The due-diligence question is whether a proposed deployment defines intent, participants, authority, evidence, evaluation, consequence, challenge, and exit—and whether current implementation evidence matches the proposal.

For researchers, the architecture offers a bridge among literatures that are often separated. Stigmergy contributes indirect coordination; distributed cognition contributes person–artifact systems; CSCW contributes articulation work; affordance theory contributes actor-relative action possibility; institutional analysis contributes governance and power; security research contributes adversarial state; multi-agent studies contribute controlled shared-world evidence. The synthesis should be judged by whether it predicts comparative outcomes, not by the breadth of its vocabulary.

DRAFT

For funders, compounding utility should be treated as an empirical milestone. Useful leading indicators are not raw message volume, total stored state, or number of agents. Better indicators include verified noncreator reuse, time saved after turnover, safe lineage depth, independent evidence coverage, dispute resolution, absence of authority bypass, and improved outcomes for new cohorts. The cost side—evaluation, monitoring, pollution, privacy, and exclusion—belongs in the same ledger.

47. Conclusion

The next phase of artificial intelligence will not be defined only by more capable models. It will be defined by the environments in which many forms of agency meet. Those environments will determine what participants can perceive, what they inherit, what they are invited or pressured to do, what persists, who can challenge it, and which outputs become real consequences.

Research across biology, robotics, distributed cognition, cooperative work, institutional governance, cultural evolution, security, and recent agent simulations supports a bounded proposition: persistent environments and artifacts can mediate coordination and preserve capability beyond individual actors. SwarmWorld supplies particularly relevant preliminary evidence for language-model societies, while also showing why portfolio performance, isolated excellence, behavioral classification, and mechanism conditions must be reported carefully.

The move from environmental coordination to **intelligent voluntary cooperation** requires more. It requires legible intent, actor-specific opportunity, meaningful choice, bounded authority, provenance, evaluation, consequence, governance, correction, and exit. It requires admitting that persistence can preserve poison, that decentralization can conceal power, and that cryptographic integrity is not truth.

At the 30 August 2026 source snapshot, Qi and IXO contain meaningful pieces of this architecture. Portal implements persistent collaborative Flow state. Qi Runtimes maintain a boundary between the model host and governed runtime authority. The IXO chain implements identifiers, entities, claims, evaluation states, disputes, escrow, and payments. Topic Protocol specifies the bounded cooperative context that can connect them, but remains a draft implementation candidate with open conformance gates. The end-to-end production reality boundary is not claimed complete.

DRAFT

That status is not a weakness to obscure. It locates the work. Build the world as a testable environment; preserve the distinctions on which trust depends; invite public correction; and measure whether people and agents actually coordinate better, more safely, and with more meaningful agency than the alternatives. Intelligent voluntary cooperation is not the assertion that emergence will save us from governance. It is the proposition that intelligence can become collective when the shared world itself is designed for legitimate action, durable learning, and accountable consequence.

DRAFT

References

Aghion, Philippe, and Jean Tirole. 1997. “Formal and Real Authority in Organizations.” *Journal of Political Economy* 105 (1): 1–29. https://doi.org/10.1086/262063.

Bansal, Gagan, Tongshuang Wu, Joyce Zhou, et al. 2021. “Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance.” *Proceedings of CHI 2021*. https://doi.org/10.1145/3411764.3445717.

Beckers, Ralph, Owen E. Holland, and Jean-Louis Deneubourg. 1994. “From Local Actions to Global Tasks: Stigmergy and Collective Robotics.” In *Artificial Life IV*, edited by Rodney A. Brooks and Pattie Maes. MIT Press. https://mitpress.mit.edu/9780262521901/artificial-life-iv/.

Buçinca, Zana, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. “To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making.” *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1). https://doi.org/10.1145/3449287.

Chen, Zhaorun, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2024. “AgentPoison: Red-Teaming LLM Agents via Poisoning Memory or Knowledge Bases.” *Advances in Neural Information Processing Systems 37*. https://doi.org/10.52202/079017-4136.

Dell’Acqua, Fabrizio, Edward McFowland, Ethan Mollick, et al. 2026. “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality.” *Organization Science* 37 (2): 403–23. https://doi.org/10.1287/orsc.2025.21838.

Derex, Maxime, and Robert Boyd. 2016. “Partial Connectivity Increases Cultural Accumulation Within Groups.” *Proceedings of the National Academy of Sciences* 113 (11): 2982–87. https://doi.org/10.1073/pnas.1518798113.

Douceur, John R. 2002. “The Sybil Attack.” *Peer-to-Peer Systems*, 251–60. https://doi.org/10.1007/3-540-45748-8_24.

Doyle, Jon. 1979. “A Truth Maintenance System.” *Artificial Intelligence* 12 (3): 231–72. https://doi.org/10.1016/0004-3702(79)90008-0.

DRAFT

Fehr, Ernst, and Simon Gächter. 2000. “Cooperation and Punishment in Public Goods Experiments.” *American Economic Review* 90 (4): 980–94. https://doi.org/10.1257/aer.90.4.980.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Gneezy, Uri, and Aldo Rustichini. 2000. “A Fine Is a Price.” *Journal of Legal Studies* 29 (1): 1–17. https://doi.org/10.1086/468061.

Grassé, Pierre-Paul. 1959. “La Reconstruction Du Nid Et Les Coordinations Interindividuelles Chez Bellicositermes Natalensis Et Cubitermes Sp.: La Théorie de La Stigmergie.” *Insectes Sociaux* 6: 41–80. https://doi.org/10.1007/BF02223791.

Green, Ben, and Yiling Chen. 2019. “Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments.” *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99. https://doi.org/10.1145/3287560.3287563.

Greshake, Kai, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. “Not What You’ve Signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection.” *arXiv*, ahead of print. https://doi.org/10.48550/arXiv.2302.12173.

Hirschman, Albert O. 1970. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press.

Holmström, Bengt. 1982. “Moral Hazard in Teams.” *Bell Journal of Economics* 13 (2): 324–40. https://doi.org/10.2307/3003457.

Hutchins, Edwin. 1995. “How a Cockpit Remembers Its Speeds.” *Cognitive Science* 19 (3): 265–88. https://doi.org/10.1207/s15516709cog1903_1.

IXO. 2026a. *IXO Portal Editor and Flow Implementation*. https://github.com/ixoworld/ixo-portal/tree/2488f83669cf08777ea1e2661432c3786a23dd70/packages/editor.

IXO. 2026b. *Qi Runtimes*. https://github.com/ixoworld/qi-runtimes/tree/a2b8cc77a683000fc9734af5090a6e3e503b3350.

IXO. 2026c. *Topic Protocol 1.0.0-Rc.1*. https://github.com/ixoworld/topic-protocol/tree/5cebe6437ea8f425e26c78eaaaf86479ff1fd2d3.

DRAFT

IXO Foundation. 2026. *IXO Blockchain V8.0.0: IID, Entity, and Claims Modules*. https://github.com/ixofoundation/ixo-blockchain/tree/bfdd934456be01f9212d3ebc107df81b43e1566d.

Laurie, Ben, Emilia Messeri, and Rob Stradling. 2021. *Certificate Transparency Version 2.0*. RFC 9162. Internet Engineering Task Force. https://doi.org/10.17487/RFC9162.

Lee, Min Kyung, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. "Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers." *Proceedings of CHI 2015*, 1603–12. https://doi.org/10.1145/2702123.2702548.

Muchnik, Lev, Sinan Aral, and Sean J. Taylor. 2013. "Social Influence Bias: A Randomized Experiment." *Science* 341 (6146): 647–51. https://doi.org/10.1126/science.1240466.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. U.S. Department of Health; Human Services. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press. https://doi.org/10.1017/CBO9780511807763.

Ostrom, Elinor. 2010. "Beyond Markets and States: Polycentric Governance of Complex Economic Systems." *American Economic Review* 100 (3): 641–72. https://doi.org/10.1257/aer.100.3.641.

Pal, Subhadeep, Fiona Y. Wang, and Markus J. Buehler. 2026. "SwarmWorld: Stigmergic Technological Evolution in Societies of Language-Model Agents." *arXiv*, ahead of print. https://doi.org/10.48550/arXiv.2608.26081.

Paolo, Giuseppe, Jamieson Warner, Hormoz Shahrzad, Babak Hodjat, Risto Miikkulainen, and Elliot Meyerson. 2026. "TerraLingua: Emergence and Analysis of Open-Endedness in LLM Ecologies." *arXiv*. https://arxiv.org/abs/2603.16910.

DRAFT

Rosenblat, Alex, and Luke Stark. 2016. “Algorithmic Labor and Information Asymmetries: A Case Study of Uber’s Drivers.” *International Journal of Communication* 10: 3758–84. https://ijoc.org/index.php/ijoc/article/view/4892.

Schmidt, Kjeld, and Carla Simone. 1996. “Coordination Mechanisms: Towards a Conceptual Foundation of CSCW Systems Design.” *Computer Supported Cooperative Work* 5 (2–3): 155–200. https://doi.org/10.1007/BF00133655.

Shapiro, Marc, Nuno Preguiça, Carlos Baquero, and Marek Zawirski. 2011. “Conflict-Free Replicated Data Types.” *Stabilization, Safety, and Security of Distributed Systems*, 386–400. https://doi.org/10.1007/978-3-642-24550-3_29.

Suchman, Lucy A. 2006. *Human–Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511808418.

W3C Decentralized Identifier Working Group. 2022. *Decentralized Identifiers (DIDs) V1.0: Core Architecture, Data Model, and Representations*. W3C Recommendation. World Wide Web Consortium. https://www.w3.org/TR/did-core/.

W3C Provenance Working Group. 2013. *PROV-DM: The PROV Data Model*. W3C Recommendation. World Wide Web Consortium. https://www.w3.org/TR/prov-dm/.

W3C Verifiable Credentials Working Group. 2025. *Verifiable Credentials Data Model V2.0*. W3C Recommendation. World Wide Web Consortium. https://www.w3.org/TR/vc-data-model-2.0/.

Wang, Guanzhi, Yuqi Xie, Yunfan Jiang, et al. 2023. “Voyager: An Open-Ended Embodied Agent with Large Language Models.” *arXiv*. https://arxiv.org/abs/2305.16291.

Zou, Wei, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. “PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models.” *34th USENIX Security Symposium*, 3827–44. https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

DRAFT

Public annexes

- Primary-source evidence dossier
- Research gap dossier
- Qi and IXO system correspondence matrix
- Threat model and counterargument annex
- Comparative experimental protocol
- Public comment-disposition register
- Machine-readable claim ledger

DRAFT