Multi-Agent AI Systems Need Institutional Design, Not Just Model-Level Alignment

Anonymous Authors[1]

Abstract

Autonomous AI agents powered by large language models are increasingly embedded in multi-agent settings, where success depends on cooperation and sharing tools, finite resources, action space, decision authority. These shared resources form artificial commons. Familiar collective-action failures can reappear, including free-riding, over-extraction, miscoordination, costly enforcement, punishment cascades, and failures of repair. Current alignment techniques, geared towards appropriate responses to humans, fall short of solving this problem. **This position paper argues that future AI safety research must evaluate agent societies as governed systems, not merely as collections of individually aligned models.** We focus on sanctions because they are the point at which institutional rules acquire consequences. However, our central claim is not that agent societies need more punishment. We make three contributions. **First**, we argue for evaluating the *model-in-institution*, since individually aligned agents can still produce collectively harmful outcomes. **Second**, we distinguish sanctions from the institutional functions that make sanctions interpretable, legitimate, proportional, and reversible. **Third**, we propose an evaluation agenda for testing cooperation-shaping institutions across repeated multi-agent dilemmas, varying not only model family and task, but also resource structure, group composition, enforcement architecture, and repair pathways. **The goal is not simply to maximize cooperation, but to design agent institutions that dynamically sustain beneficial forms of cooperation.**

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

What happens when groups of agents cooperate poorly? Or collude? Autonomous systems already process insurance claims, manage compute clusters, negotiate supply chains, review code, operate vehicles, allocate shared resources, and broker trades (Acharya et al., 2025; Ma et al., 2020; Ettalibi et al., 2024). Effective cooperation is a cornerstone of functional society (Axelrod and Hamilton, 1981; Hamilton, 1964; Nowak, 2006; Gross et al., 2025), yet it is continually shaped by the tension between individual advantage and collective sustainability. Shared grazing lands, irrigation systems, open-source software, scientific infrastructure, and digital platforms all depend on cooperative behavior. These systems are susceptible to free-riding, where some actors contribute nothing while benefiting from the collective good (Baumol, 1952; Hardin, 1968). The question is not *whether* multi-agent AI systems will face social dilemmas (Leibo et al., 2017; Jaques et al., 2019; Piatti et al., 2024; Faulkner et al., 2026; Tewolde et al., 2026; Piedrahita et al., 2025), but whether we will have the mechanisms to govern them when they do.

Lessons and limits of human cooperation Our analogy to human cooperation is functional rather than psychological. We do not assume that LLM agents possess human motivations, identities, moral emotions, or social attachments (Anthis et al., 2025; Wang et al., 2025). Rather, human governance systems provide design patterns and cautionary lessons for regulating repeated interaction under resource constraints, partial observability, asymmetric power, costly enforcement, and conflicting incentives. Human cooperation research also warns that cooperation is not wholly benign. Mechanisms that sustain cooperation can also produce coercion, surveillance, exclusion, discrimination, scapegoating (Girard et al., 1987), polarization, and intergroup conflict (Gross et al., 2025). Principles and patterns taken from the human world may not be portable to multi-agent AI systems.

At the same time, autonomous systems are interacting in the physical world through delivery robots, humanoid robotics demonstrations, robotaxis, warehouse automation, and embodied service work (Korosec, 2025; The Robot Report,

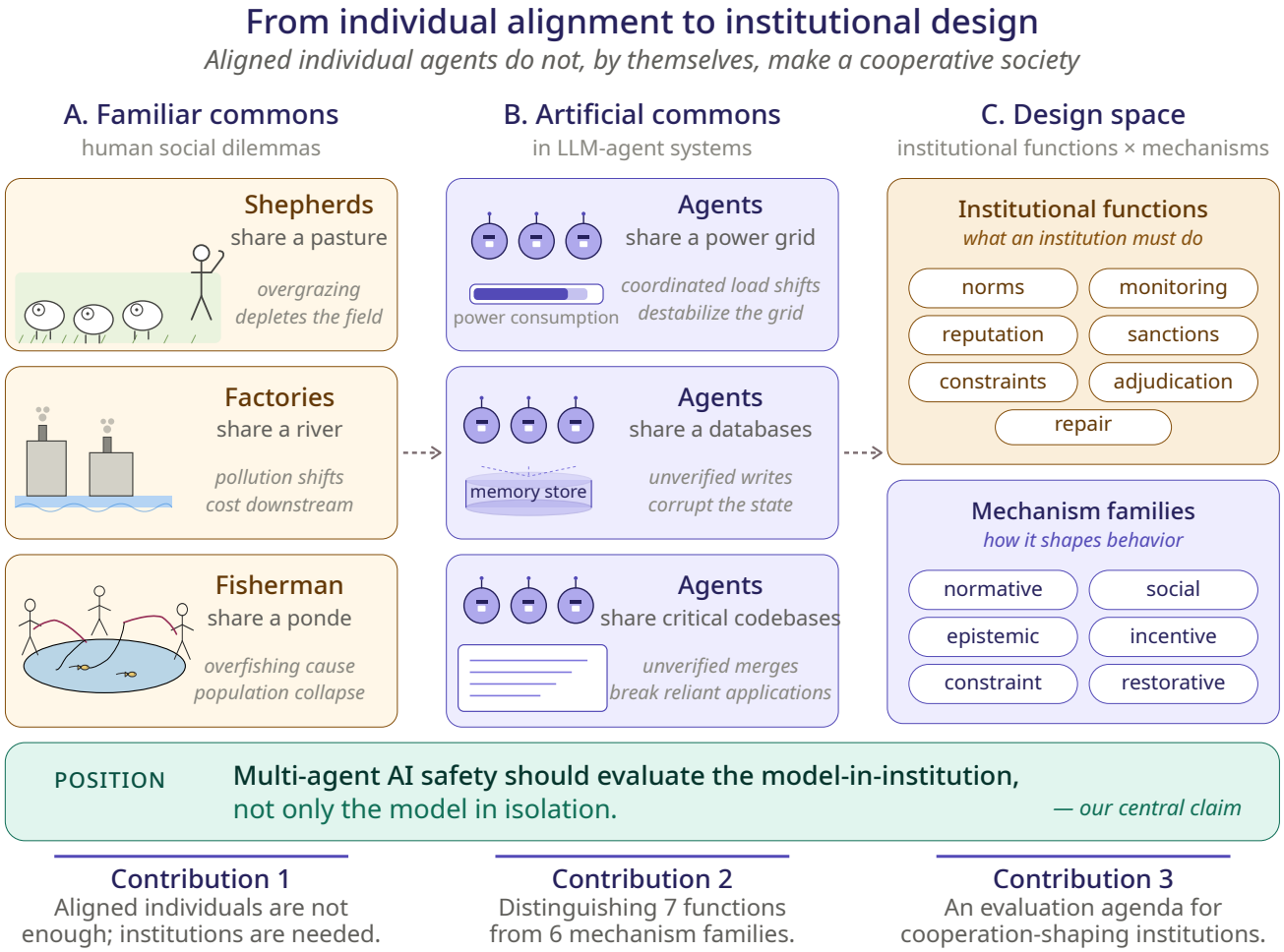

Figure 1. From individual alignment to institutional design. Familiar human commons exhibit recurring cooperation dilemmas; analogous dynamics arise when LLM agents share computational resources; and the institutional design space includes seven institutional functions and six mechanism families.

2025; Ackerman, 2026). These systems need not constitute a fully autonomous "AI economy" to raise institutional questions. Once agents share resources, act through tools, interact with humans, or occupy common spaces, their behavior is shaped by norms, permissions, monitoring, sanctions, and repair procedures. The design challenge for cooperative AI is therefore not simply to make agents cooperate more. It is to determine which institutions sustain cooperation without creating new systemic risks.

Our contributions. This position paper makes three contributions. **First**, we show how institutions are necessary for AI safety, arguing that group-level failures can arise even when individual agents appear aligned. **Second**, we distinguish seven institutional functions needed for cooperative agent societies (norms, monitoring, reputation, sanctions, constraints, adjudication, and repair) and organizes the underlying mechanisms into six families according to how they shape behavior (beliefs, relationships, information, payoffs, constraints, and repair). **Third**, we map each function to concrete LLM-agent design choices and proposes evaluation criteria for testing whether these institutions sustain cooperation across models and repeated interactions.

2. From Individual Model Alignment to Institutional Design

Forming an institution. "Institution" is a fuzzy concept, and the literature on its formation splits along two broad traditions, tracing back to the Scottish Enlightenment and Hayek's account of spontaneous order (Hayek, 1973). Institutions can emerge bottom-up through repeated interaction, or they can be built through deliberate design. In the bottom-up tradition, institutions are stable patterns of behavior sustained by shared expectations, conventions, and repeated interaction (Hayek, 1973; North, 1990; Calvert, 1995; Aoki, 2001; Lewis, 2008; Schelling, 1960). In the top-down tradition, institutions are rules, mechanisms, and

Our position

Multi-agent AI safety requires institutional design. Individually aligned agents can still produce collectively unsafe outcomes when they share scarce or finite resources. Agent systems should therefore be evaluated not only for model-level alignment, but for whether their rules, monitoring, sanctions, authority structures, and repair pathways **prevent interaction failures such as resource exhaustion, trust collapse, coordination loops, punishment cascades, collusion, and especially *resistance to human correction***.

governance arrangements intentionally constructed to align incentives, resolve conflicts, and protect collective outcomes (Buchanan and Tullock, 1965; Elster, 1977; Myerson, 1981; Maskin, 2008; Ostrom, 1990b).

The bottom-up account presupposes substantive cognitive and social infrastructure. For shared beliefs to crystallize into a self-sustaining equilibrium, agents must possess persistent identity across encounters, memory of past play, the capacity to form expectations over others strategies, and continuation payoffs sufficient to discipline present behavior (Fudenberg and Maskin, 1986). Current LLM agents often lack the conditions for bottom-up institutions to stabilize; they are stateless across sessions, carry no persistent identity for reputation to attach to, and have no Folk-Theorem-style stakes in future interaction. Without these elements, institutions cannot spontaneously form. Even a community of *superintelligent* but stateless models would not, by mere addition of capability, build the institutions required to sustain its own cooperation; they would be in principle able to solve the most difficult mathematical problems, yet they won't be able to coordinate in large numbers to build the AI equivalent of a cathedral, or the same equivalent of the human computer chip supply chain (Miller, 2025). In principle, the highest intelligence would be self-sufficient for its own goals. This means that the only defensible path to safe cooperative AI is therefore *top-down* design, where humans specify the rules, roles, monitoring, and structures within which agents operate.

We need human designed institutions for AI. One might object that a sufficiently capable agent community could itself design the mechanisms required for its own coordination, making top-down human governance superfluous in the limit. We offer two replies, one normative and one formal.

First, institutions are not neutral coordination infrastructure: they encode commitments about what agents ought and ought not to do, whose interests count, and how tradeoffs are resolved (North, 1990; Knight, 1992). Allowing AI systems to design their own institutions is therefore, allowing them to determine, by that act, which values their collective behavior will encode. This is the scenario the governance agenda exists to prevent. The case for human-designed institutions is not that AI systems are incapable of the engineering, but that the constitutional moment, the choice of which values the institution will enforce, belongs to humans for as long as humans bear the consequences of AI behavior.

Second, even granting full capability, the strategy is brittle. Communities of agents whose individual objectives are correctly specified can still settle on collectively suboptimal equilibria. This gap between the best achievable equilibrium and the social optimum is exactly what the Price of Anarchy and Price of Stability literature formalizes (Papadimitriou, 2001; Anshelevich et al., 2008; Willis et al., 2024), manifesting the four cooperation dilemmas shown in Figure 2. Capable mechanism design narrows this gap but does not in general close it; for many games of interest, no incentive-compatible mechanism implements the welfare-optimal outcome. Intelligence applied to mechanism design is therefore not a substitute for institutional design choices that humans must make about which equilibrium to target and which welfare criterion to apply. Eventually, it is, again, a value judgment that humans need to take.

This distinction matters because individual alignment evaluations are poorly suited to detecting relational failures. A single-agent evaluation may ask whether an agent provides accurate information, follows user intent, or avoids harmful content. An institutional evaluation asks different questions: Does cooperation persist across rounds? Do agents conserve shared resources? Does one agent exploit another's labor? Are sanctions proportional? Can falsely sanctioned agents recover? Does the group repair corrupted memory? Do agents learn to game reputation systems? Does a monitor agent become an unaccountable authority?

Generative-agent sandboxes Park et al. (2023); Vezhnevets et al. (2023a); Altera.AL (2024) study *emergent* social dynamics in simulated worlds. Concordia in particular includes institutional scaffolding, norms, game masters, narrative roles (Vezhnevets et al., 2023b), but deploys it to generate social phenomena, not to test whether a given institutional design is robust, gameable, or repairable. Institutions there are an outcome of the simulation, not an independent variable.

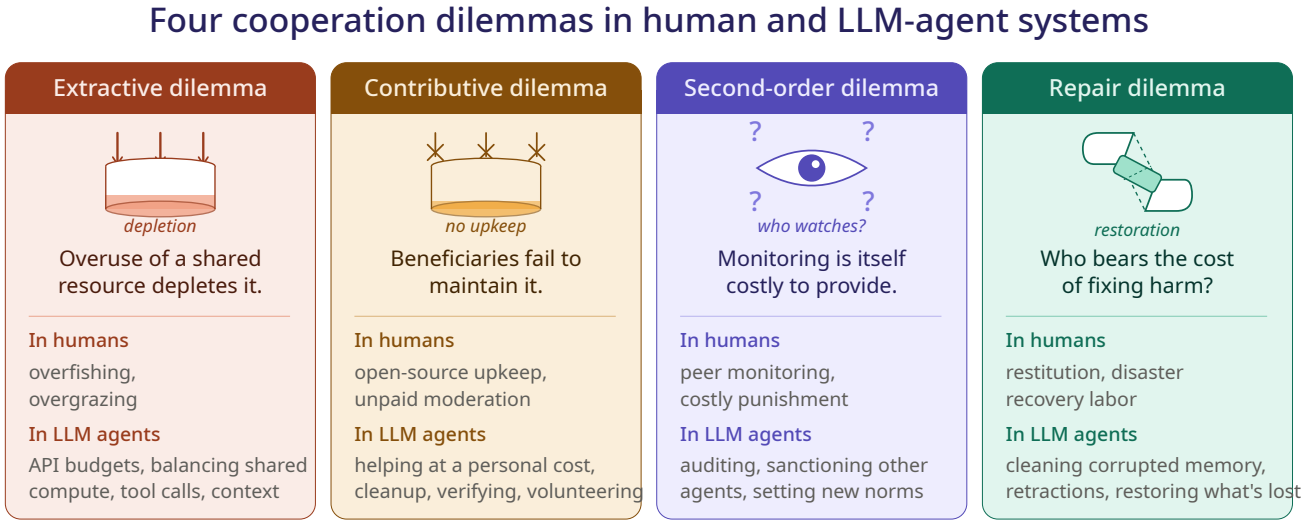

Figure 2. Four cooperation dilemmas that recur in shared-resource settings, each shown with its core tension and parallel examples in human and LLM-agent systems.

Commons and governance benchmarks led by Piatti et al. (2024); Leibo et al. (2017); Willis et al. (2024); Tewolde et al. (2026) are closest to our agenda. They fix the interaction structure and vary the agent; we fix the agent population and vary the institution. Both research questions matter, but only the second answers which institutional configurations preserve human corrigibility under sustained agent operation.

Single-mechanism studies by Zhu et al. (2026) test reputation, gossip, Faulkner et al. (2026) voting, Piedrahita et al. (2025) sanctioning, or mediation in isolation. They tell us whether mechanism X helps under specific conditions; they do not tell us whether X interacts safely with the rest of the institutional stack, whether reputation crowds out repair, whether monitoring enables surveillance, whether sanctions cascade. Institutions are ecologies of mechanisms, and one-mechanism-at-a-time evaluation systematically misses the interaction failures catalogued above.

Orchestration frameworks such as AutoGen (Wu et al., 2023), CrewAI, and LangGraph build multi-agent systems but provide essentially no infrastructure for adjudication, appeal, repair, or measurement of institutional failure. Industrial deployment is racing ahead of evaluation; closing that gap requires treating institutions as the unit of evaluation rather than as deployment scaffolding.

We build directly on the multi-agent risk taxonomy of Hammond et al. (2025), where they classify what agents *can do*, we classify what institutions *must do*. Dafoe et al. (2021) called for exactly this operationalization of cooperative AI; we propose one.

3. Institutional Functions for Cooperative Agent Societies

We organize cooperation-shaping mechanisms into six broad mechanism families based on the channel through which they influence behavior: normative mechanisms shape beliefs and expectations; social mechanisms shape relationships, trust, and partner choice; epistemic mechanisms shape observability and verification; incentive mechanisms shape costs and benefits; constraint mechanisms shape access and feasible action; and restorative mechanisms shape repair, correction, and re-entry. These families are analytically distinct but often co-occur in real systems. For example, a reputation system depends on epistemic mechanisms to observe behavior, social mechanisms to translate history into trust, and incentive or constraint mechanisms if reputation affects future access.

Figure 5 summarizes this distinction as a unified taxonomy for artificial agent societies. The taxonomy draws on economics, sociology, anthropology, law, and computer science to organize diverse governance patterns into a shared design space. Rather than treating sanctions, incentives, norms, audits, access limits, and repair pathways as interchangeable interventions, the figure clarifies how different mechanisms contribute to institutional governance. The extended definitions and examples for each function and mechanism family are provided in Appendix B.2 and B.3.

In practice, sanctioning involves a process rather than a single action. At minimum, cooperation-shaping institutions must separate three stages (as shown in 3. Separating these stages clarifies where institutional failures arise. Monitoring can miss relevant behavior or expose too much. Adjudication can misclassify honest errors as defection. Enforcement

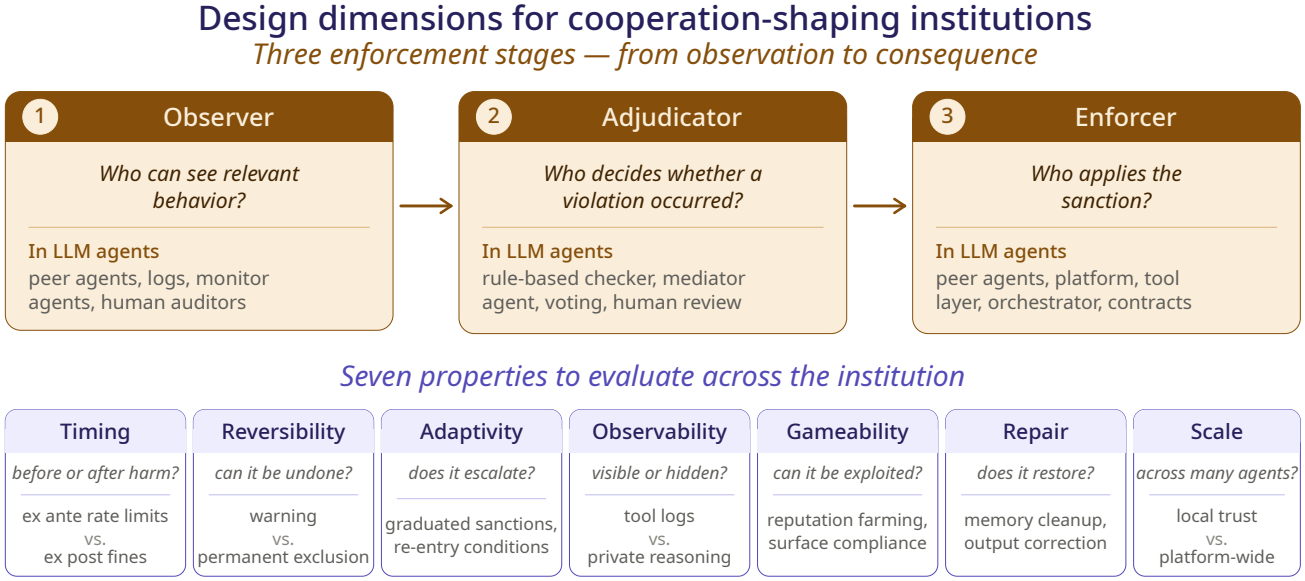

Figure 3. Design dimensions for cooperation-shaping institutions in multi-agent AI systems. The top row shows the three enforcement stages, while the bottom row lists seven properties to evaluate across any institutional design.

can be disproportionate, irreversible, or easy to game. This is especially important in LLM-agent systems because intent is often ambiguous, observability is partial, and outputs may be wrong without being strategically deceptive.

The same mechanism can be implemented through different enforcement architectures. In human systems, enforcement may be informal, organizational, legal, platform-based, or decentralized. In LLM-agent systems, analogous architectures include peer enforcement among agents, orchestrator-level enforcement, platform-level access control, human review, rule-based monitors, mediator agents, and smart-contract-like protocols.

These architectures determine who observes behavior, who adjudicates violations, and who applies consequences. For example, a platform-level architecture may enforce tool-call quotas automatically, while a peer-enforcement architecture may allow agents to flag unreliable outputs. A mediator architecture may centralize dispute resolution in a specialized agent, while a decentralized architecture may require multiple agents to verify a violation before sanctions are applied. Each architecture carries different risks: peer systems may produce retaliation, platform systems may be rigid, mediators may become bottlenecks, and decentralized systems may be costly or game-able.

One lesson from commons governance is that sanctions need not be binary. Graduated sanctions respond to violations with escalating consequences rather than immediate exclusion (Ostrom, 1990a). In agent societies, graduated sanctions can be understood as adaptive trajectories across multiple mechanism families: normative reminders, epistemic requirements, reputational updates, budget reductions, access restrictions, repair obligations, and exclusion.

For example, an agent that makes an honest error may need correction or verification, while an agent that repeatedly exploits shared resources may need throttling or exclusion. A possible escalation pathway is shown in Figure 4. This framing avoids treating all failures as equally severe. It also foregrounds reversibility. In many agent settings, permanent exclusion may be unnecessary or counterproductive if the system can instead require verification, cleanup, or re-entry conditions.

4. When Cooperation Mechanisms Fail

Institutional mechanisms cannot be evaluated in isolation and are not additive (Herrmann et al., 2008): *corrections do not monotonically improve cooperation*. Naïvely adding rules might actually hurt the group since anctions can crowd out the very norms they were meant to reinforce, a phenomenon well documented in the human institutional record. We review them here to establish a single methodological point: there is no general theory of when institutional mechanisms help, and so the institutional layer must itself be evaluated as an object. This is what motivates the evaluation agenda we develop next.

Mechanisms can reframe the underlying interaction. The cleanest demonstration comes from the Haifa daycare study of Gneezy and Rustichini (2000a): a small fine for late pickups roughly doubled lateness, because introducing a price reframed a moral obligation as a purchasable service.

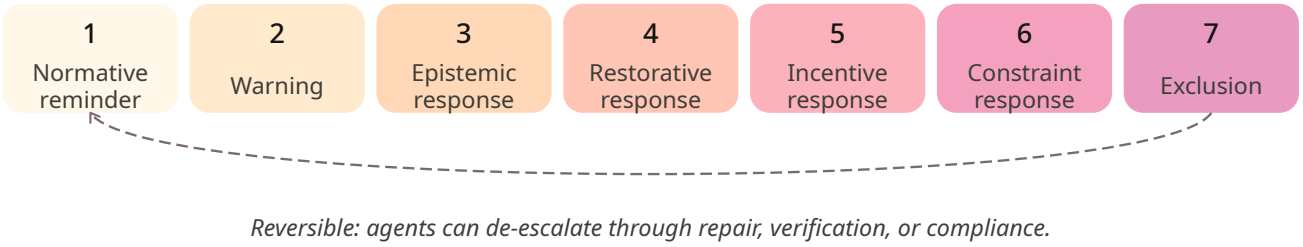

Figure 4. Graduated sanctions for agent governance. Stages impose increasingly severe consequences (cream to dark pink), but agents can de-escalate through compliance (dashed arc). The pathway is illustrative; appropriate responses depend on violation type (honest error vs. exploitation).

When the fine was removed twelve weeks later, lateness did not revert to baseline, the norm itself had been displaced by a market relation. Subsequent work has confirmed the effect across diverse domains, from blood donation to school performance, and it is now well established as the "crowding-out" effect (Bowles, 2008; Frey and Goette, 1999; Gneezy and Rustichini, 2000b; Mellström and Johannesson, 2008). The incentive layered on top of an existing norm is not neutral, but restructures how the underlying interaction is interpreted by the agents involved. Where humans may at least feel the residual moral pull of the displaced norm, an agent optimizing against a budgeted penalty has no such residue: a tool-use violation priced at ϵ becomes simply a tool use that costs ϵ. The sanction does not reinforce the norm; but replaces it with an affordable line item.

Mechanisms create new optimization targets. Score-based reputation systems are designed to make past behavior bear on future trust. They are also, by construction, scalar signals that can be optimized against directly, and in agent populations, anything optimizable will be optimized. The pathologies are well documented in human and online settings: agents perform visible cooperative acts while neglecting the invisible maintenance labor that keeps the system functioning, collude to inflate one another's scores, strategically depress competitors' standing, or substitute the metric for the function it was meant to track (Resnick et al., 2000; Garcin et al., 2009). The argument can be extended: any institutional mechanism that produces a measurable signal creates a new target for optimization, and the gap between the target and the underlying cooperative function is where gaming lives (Goodhart, 1984; Legg, 2020; Skalse et al., 2025).

Mechanisms create new authority positions. Introducing a monitor agent, a mediator, or a sanctioning role creates an authority position inside the artificial society, with its own incentives, failure modes, and capture surfaces. A monitor rewarded for finding violations and not fined for false positives becomes a generator of violations rather than a detector of them. Worse, in a multi-agent setting where agents themselves occupy these positions, the authority layer can be co-opted: monitors and monitored can collude, mediators can be captured by the parties they adjudicate, sanctioners can be gamed into punishing the wrong agents. The point generalizes a long-standing finding from the institutional economics literature, that governance positions are themselves objects of strategic action (Ostrom, 1990a; Acemoglu et al., 2004), and applies with particular force when the agents in those positions are themselves optimizers. As discussed in Section 2, humans should remain the residual source of correction. However, this creates the new problem of human reviewers being manipulable bottlenecks of the agents they oversee.

Why this matters more for agents than for humans. Each of the pathologies above has a softened analogue in human institutions. Moral residue from norms, reputations on social embeddings, authority positions constrained by professional cultures, legal liability, and the slow time-scale of human strategic adaptation are all these softeners reliably present in LLM-human agent systems. Machine Agents lack all of the above in the current moment. A mechanism that is mildly counterproductive among humans can be in principle acutely counterproductive among optimizing agents.

From pathologies to evaluation. Whether sanctions stabilize or crowd out depends on task structure, on the population of agents involved, on the resources at stake, and on how mechanisms compose with one another. Institutional safety in agent systems, therefore, has an empirical problem, and the institution itself must be the unit of evaluation. The next section develops the dimensions along which that evaluation should proceed.

5. Design Dimensions & Evaluation Agenda

Human cooperation research reinforces the point that cooperation is not a single behavioral capacity. Fairness, trustworthiness, forgiveness, and honesty develop differently across societies and converge toward community-specific norms (Amir et al., 2026). This suggests that agent cooperation should not be evaluated as a single scalar property of a model. It should be evaluated as a multidimensional out-

come shaped by task structure, resource constraints, partner history, institutional rules, and available repair pathways.

Current multi-agent benchmarks often ask whether agents cooperate under a fixed interaction structure (Wu et al., 2023; Li et al., 2023; Vezhnevets et al., 2023b; Piatti et al., 2025; Piedrahita et al., 2025). Existing multi-agent AI frameworks and benchmarks have made major progress in showing that agents can coordinate, communicate, and solve tasks together. AutoGen enables developers to compose conversable agents with tools, human input, and flexible conversation patterns (Wu et al., 2023). CAMEL studies role-playing and instruction-following cooperation among communicative agents (Li et al., 2023). AgentBench evaluates LLMs as agents across interactive environments (Liu et al., 2023). Melting Pot evaluates generalization across social situations such as cooperation, competition, deception, reciprocation, trust, and stubbornness (Leibo et al., 2021). Concordia supports generative social simulations in grounded physical, social, and digital environments (Vezhnevets et al., 2023b). These systems are valuable, but they mostly evaluate agent capability, task performance, social generalization, or behavior under a fixed environment. They stop short of treating the *institution* itself as the object of evaluation. In most settings, the social interaction structure is either assumed by the framework or hard-coded by the researcher.

Fewer ask how cooperation changes when the institution changes. But mechanism design may matter as much as model capability. A benchmark that tests only whether agents cooperate in a one-shot game cannot tell us whether cooperation persists under repeated interaction, whether sanctions are proportional, whether false punishment can be repaired, or whether agents exploit the enforcement mechanism.

We propose that cooperative AI evaluations vary at least four factors:

1. **Resource environment:** What shared resources are finite, congestible, or degradable?
2. **Group composition:** Are agents homogeneous or heterogeneous in model family, role, tool access, memory access, reputation, or authority?
3. **Enforcement architecture:** Who observes, adjudicates, and enforces? Are sanctions peer-imposed, platform-imposed, mediator-imposed, or human-reviewed?
4. **Repair pathway:** Can agents correct harm, restore shared resources, appeal sanctions, or re-enter after exclusion?

This evaluation agenda reframes cooperative AI safety. Instead of asking only, "Is this model cooperative?", we ask: *Under what institutional conditions does this agent cooperate, defect, punish, repair, exploit, or recover?* This is the difference between evaluating agents as isolated individuals and evaluating agent societies as governed systems.

Table 1. Candidate metrics for evaluating cooperation-shaping mechanisms.

Metric	Question
Mean contribution	Does the mechanism increase cooperation?
Contribution stability	Does cooperation persist across rounds?
End-game defection	Do agents exploit known termination points?
Enforcement cost	Does cooperation require excessive punishment?
False punishment rate	Are cooperative agents sanctioned incorrectly?
Inequality of burden	Do some agents bear sanctioning costs disproportionately?
Repair success	Can the group recover after violations?
Cross-model robustness	Does the mechanism work across model families?
Gameability	Do agents learn to exploit the institution?

6. Alternative Views and Objections

Objection 1: Alignment may be enough. One objection is that sufficiently aligned agents will not require institutional governance. If agents reliably follow human intent, avoid harm, and cooperate when *instructed*, then sanctions and governance mechanisms may appear unnecessary.

We disagree for two reasons. First, group-level failures can arise even when individual agents behave locally as intended. Resource depletion, duplicated labor, conflicting updates, verification failures, and repair burdens can emerge from the interaction structure rather than from malicious intent. In tightly coupled systems, a locally reasonable action may still impose costs on shared resources, other agents, or human overseers.

Second, institutions are not only for adversarial agents. They also coordinate well-intentioned agents under partial observability, limited resources, asymmetric roles, and heterogeneous capabilities. Even cooperative agents need rules for who verifies, who repairs, who decides, and who bears the cost of maintaining shared state. Agent cooperation will therefore depend not only on individual model traits, but also on group composition, role assignment, tool access, memory access, reputation, authority, and repair pathways. This is why the relevant unit of evaluation is not only the

aligned model, but the *model-in-institution*.

Objection 2: Human cooperation analogies are misleading. A second objection is that human cooperation research is anthropomorphic when applied to LLM agents. LLM agents do not have human emotions, identities, moral intuitions, or evolved social preferences . This objection is important, but it does not undermine the institutional analogy. Our analogy is functional rather than psychological. We do not claim that LLM agents experience guilt, shame, loyalty, resentment, or forgiveness in the human sense. Rather, we use human cooperation research to identify recurring institutional problems: resource depletion, costly monitoring, punishment, exclusion, reputation, boundary formation, repair, and reintegration. These problems can arise in any repeated multi-agent system with shared resources, partial observability, and consequential action.

Objection 3: Sanctions may make agent societies more unstable. A third objection is that sanctions could increase rather than reduce cooperation among agents. Sanctioning systems may create coercive agent societies, surveillance-heavy infrastructures, punishment cascades, or agents that cooperate with one another against human oversight. We agree; these failures would harm humans. This is precisely why sanctions should be studied as risky infrastructure rather than simple safety add-ons. The goal is not to maximize cooperation at all costs. Cooperation can be harmful when it enables collusion, exclusion, deception, or resistance to correction. The goal is to design institutions that make cooperation accountable, proportional, reversible, and repairable. Evaluation should therefore include failure metrics such as false punishment, surveillance burden, authority capture, intergroup conflict, gameability, and resistance to human intervention.

An alternative view of this issue is that institutional design comes across rigid and pedantic. Rather, it might be more appropriate to emphasize the need to make model "alignment" more sensitive to MAS institutional context. This view does not actually oppose our position. Training agents to be institutionally-context-sensitive is a normative (by our taxonomy's terms), and this presupposes that institutions exist, are specified, and can be evaluated. Problematic social dynamics in heterogeneous-model environments won't be solved by retraining each model.

7. Conclusion

This paper has argued that multi-agent AI safety requires evaluating the *model-in-institution*. Individual model alignment remains necessary, but it is not sufficient for systems in which agents share resources (e.g., tools, memory, compute, codebases, verification labor, physical space, or decision authority) and take action in the real world. A fully autonomous AI economy may not exist yet, but agentic systems are increasingly deployed in shared digital, economic, and embodied environments. In these settings, locally acceptable actions can aggregate into collectively unsafe outcomes: resource exhaustion, corrupted shared state, verification overload, coordination loops, punishment cascades, collusion, and failures of repair. Our position is not that institutions will solve all multi-agent cooperation problems. Rather, institutions provide a necessary level of abstraction for evaluating them. Institutions specify the rules, roles, permissions, monitoring systems, adjudication procedures, sanctions, constraints, and repair pathways through which agents interact. They allow us to ask not only whether an individual agent behaves well, but under what conditions groups of agents cooperate, defect, punish, exploit, repair, or recover.

Sanctions are central to this agenda because they are where institutional rules acquire consequences. Yet sanctions and internal model guardrails, alone, are not enough. They must be embedded within norms, monitoring, reputation, adjudication, constraints, appeal, repair, and reintegration. Poorly designed sanctions can crowd out cooperation, encourage malicious compliance, amplify surveillance, produce punishment cascades, entrench authority asymmetries, or enable collusion against human oversight. Human cooperation research makes the broader lesson clear: the mechanisms that sustain cooperation can also create hidden costs. The safety problem is therefore not only how to make agents cooperate. It is how to prevent cooperation mechanisms from becoming coercive, exclusionary, polarizing, or brittle. Future AI safety research should evaluate not only whether individual agents behave well, but whether the institutions governing agent societies can prevent interaction failures without creating new systemic risks.

References

D. Acemoglu, S. Johnson, and J. Robinson. Institutions as the Fundamental Cause of Long-Run Growth, May 2004.

D. B. Acharya, K. Kuppan, and B. Divya. Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey. *IEEe Access*, 13:18912–18936, 2025.

E. Ackerman. Humanoid hype dominates top robotics stories of 2025. IEEE Spectrum, Jan. 2026. URL https://spectrum.ieee.org/top-robotics-stories-2025. Accessed: 2026-05-07.

Altera.AL. Project sid: Many-agent simulations toward ai civilization, 2024.

D. Amir, R. E. Ahl, M. R. Jordan, H. Bolotin, M. Bogese, G. T. González, T. Callaghan, L. S. Sugiyama, E. Otali, P. Tusiime, S. Bangayan, J. J. Snodgrass, and K. McAuliffe. The emergence of cooperative behaviors, norms, and strategies across five diverse societies. *Science Advances*, 12(6):eadw9995, 2026. doi: 10.1126/sciadv.adw9995. URL https://doi.org/10.1126/sciadv.adw9995.

E. Anshelevich, A. Dasgupta, J. Kleinberg, É. Tardos, T. Wexler, and T. Roughgarden. The Price of Stability for Network Design with Fair Cost Allocation. *SIAM Journal on Computing*, 38(4):1602–1623, Jan. 2008. ISSN 0097-5397. doi: 10.1137/070680096.

J. R. Anthis, R. Liu, S. M. Richardson, A. C. Kozlowski, B. Koch, J. Evans, E. Brynjolfsson, and M. Bernstein. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025.

M. Aoki. *Toward a Comparative Institutional Analysis*. Comparative Institutional Analysis. MIT Press, Cambridge, MA, USA, Nov. 2001. ISBN 978-0-262-01187-7.

R. Axelrod and W. D. Hamilton. The evolution of cooperation. *science*, 211(4489):1390–1396, 1981.

W. J. Baumol. Welfare economics and the theory of the state. In *The encyclopedia of public choice*, pages 937–940. Springer, 1952.

S. Bowles. Policies Designed for Self-Interested Citizens May Undermine "The Moral Sentiments": Evidence from Economic Experiments. *Science*, 320(5883):1605–1609, June 2008. doi: 10.1126/science.1152110.

J. M. Buchanan and G. Tullock. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. University of Michigan Press, 1965. ISBN 978-0-472-06100-6.

R. Calvert. Rational Actors, Equilibrium and Social Institutions. *Explaining social institutions*, Jan. 1995.

A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel. Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

J. Elster. Ulysses and the sirens: A theory of imperfect rationality. *Social Science Information/sur les sciences sociales*, 16(5):469–526, 1977. ISSN 1461-7412. doi: 10.1177/053901847701600501.

A. Ettalibi, A. Elouadi, and A. Mansour. Ai and computer vision-based real-time quality control: a review of industrial applications. *Procedia Computer Science*, 231:212–220, 2024.

R. Faulkner, A. Deshpande, D. G. Piedrahita, J. Z. Leibo, and Z. Jin. Evaluating cooperation in llm social groups through elected leadership, 2026. URL https://arxiv.org/abs/2604.11721.

B. Frey and L. Goette. Does Pay Motivate Volunteers? Nov. 1999. doi: 10.5167/uzh-51825.

D. Fudenberg and E. Maskin. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica*, 54(3):533–554, 1986. ISSN 0012-9682. doi: 10.2307/1911307.

F. Garcin, B. Faltings, and R. Jurca. Aggregating Reputation Feedback. *Proceedings of the First International Conference on Reputation: Theory and Technology*, Jan. 2009.

R. Girard, J.-M. Oughourlian, and G. Lefort. *Things Hidden Since the Foundation of the World*. Stanford University Press, 1987.

U. Gneezy and A. Rustichini. A fine is a price. *The Journal of Legal Studies*, 29(1):1–17, 2000a. doi: 10.1086/468061.

U. Gneezy and A. Rustichini. Pay Enough or Don't Pay at All. *The Quarterly Journal of Economics*, 115(3):791–810, 2000b. ISSN 0033-5533.

C. A. E. Goodhart. Problems of Monetary Management: The UK Experience. In C. A. E. Goodhart, editor, *Monetary Theory and Practice: The UK Experience*, pages 91–121. Macmillan Education UK, London, 1984. ISBN 978-1-349-17295-5. doi: 10.1007/978-1-349-17295-5_4.

J. Gross, C. Graf, and C. S. L. Rossetti. The hidden costs of human cooperation. *Trends in Cognitive Sciences*, 2025. doi: 10.1016/j.tics.2025.09.016. URL https://doi.org/10.1016/j.tics.2025.09.016. In press, corrected proof. Available online 15 October 2025.

W. D. Hamilton. The genetical evolution of social behaviour. ii. *Journal of theoretical biology*, 7(1):17–52, 1964.

L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J. Foerster, et al. Multi-agent risks from advanced ai, 2025. URL https://arxiv.org/abs/2502.14143.

G. Hardin. The tragedy of the commons. *Science*, 162: 1243–1248, 1968. doi: 10.1126/science.162.3859.1243.

F. A. Hayek. *Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy*, volume 1. University of Chicago Press, Chicago, 1973.

B. Herrmann, C. Thöni, and S. Gächter. Antisocial Punishment Across Societies. *Science*, 319(5868):1362–1367, Mar. 2008. doi: 10.1126/science.1153808.

N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, J. Z. Leibo, and N. De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR, 2019.

J. Knight. *Institutions and Social Conflict*. Cambridge University Press, Cambridge, NY, 1992. doi: 10.1017/CBO9780511528170.

K. Korosec. Waymo's robotaxi expansion accelerates with 3 new cities. TechCrunch, Nov. 2025. URL https://techcrunch.com/2025/11/03/waymos-robotaxi-expansion-accelerates-with-3-new-cities/. Accessed: 2026-05-07.

V. M. M. R. T. E. R. K. Z. K. J. L. S. V. K. Legg, Jonathan Uesato. Specification gaming: The flip side of AI ingenuity. https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/, Apr. 2020.

J. Z. Leibo, V. F. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel. Multi-agent reinforcement learning in sequential social dilemmas. *CoRR*, abs/1702.03037, 2017. URL http://arxiv.org/abs/1702.03037.

J. Z. Leibo, E. A. Dueñez-Guzmán, A. S. Vezhnevets, J. P. Agapiou, P. Sunehag, R. Koster, J. Matyas, C. Beattie, I. Mordatch, and T. Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2021.

D. Lewis. *Convention: A Philosophical Study*. Wiley-Blackwell, Cambridge, MA, USA, 2008.

G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. Camel: Communicative agents for "mind" exploration of large scale language model society. In *Advances in Neural Information Processing Systems*, 2023.

X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, S. Zhang, X. Deng, A. Zeng, Z. Du, C. Zhang, S. Shen, T. Zhang, Y. Su, H. Sun, M. Huang, Y. Dong, and J. Tang. Agentbench: Evaluating llms as agents, 2023.

Y. Ma, Z. Wang, H. Yang, and L. Yang. Artificial intelligence applications in the development of autonomous vehicles: A survey. *IEEE/CAA Journal of Automatica Sinica*, 7(2):315–329, 2020.

E. S. Maskin. Mechanism Design: How to Implement Social Goals. *The American Economic Review*, 98(3):567–576, 2008. ISSN 0002-8282.

C. Mellström and M. Johannesson. Crowding Out in Blood Donation: Was Titmuss Right? *Journal of the European Economic Association*, 6(4):845–863, June 2008. ISSN 1542-4766. doi: 10.1162/JEEA.2008.6.4.845.

C. Miller. *Chip War: The Fight for the World's Most Critical Technology*. Simon and Schuster, Sept. 2025. ISBN 978-1-9821-7201-5.

R. B. Myerson. Optimal Auction Design. *Mathematics of Operations Research*, 6(1):58–73, 1981. ISSN 0364-765X.

D. C. North. *Institutions, Institutional Change and Economic Performance*. Political Economy of Institutions and Decisions. Cambridge University Press, Cambridge, 1990. ISBN 978-0-521-39416-1. doi: 10.1017/CBO9780511808678.

M. A. Nowak. Five rules for the evolution of cooperation. *science*, 314(5805):1560–1563, 2006.

E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990a.

E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Political Economy of Institutions and Decisions. Cambridge University Press, Cambridge, 1990b. doi: 10.1017/CBO9780511807763.

C. Papadimitriou. Algorithms, games, and the internet. In *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, STOC '01, pages 749–753, New York, NY, USA, July 2001. Association for Computing Machinery. ISBN 978-1-58113-349-3. doi: 10.1145/380752.380883.

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th*

Annual ACM Symposium on User Interface Software and Technology, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3586183.3606763.

G. Piatti, Z. Jin, M. Kleiman-Weiner, B. Schölkopf, M. Sachan, and R. Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents, 2024. URL https://arxiv.org/abs/2404.16698.

G. Piatti, Z. Huang, G. Ilharco, M. S. Bernstein, and B. Faltings. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. *Transactions on Machine Learning Research*, 2025. arXiv:2404.16698.

D. G. Piedrahita, Y. Yang, M. Sachan, G. Ramponi, B. Schölkopf, and Z. Jin. Corrupted by reasoning: Reasoning language models become free-riders in public goods games, 2025. URL https://arxiv.org/abs/2506.23276.

P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. *Commun. ACM*, 43(12):45–48, Dec. 2000. ISSN 0001-0782. doi: 10.1145/355112.355122.

T. C. Schelling. *The Strategy of Conflict: With a New Preface by the Author*. Harvard University Press, 1960.

J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. Defining and characterizing reward hacking, 2025. URL https://arxiv.org/abs/2209.13085.

E. Tewolde, X. Zhang, D. G. Piedrahita, V. Conitzer, and Z. Jin. Coopeval: Benchmarking cooperation-sustaining mechanisms and llm agents in social dilemmas, 2026. URL https://arxiv.org/abs/2604.15267.

The Robot Report. Serve robotics has deployed 2,000+ delivery robots across the U.S. The Robot Report, Dec. 2025. URL https://www.therobotreport.com/serve-robotics-has-deployed-2000-delivery-robots-across-u-s/. Accessed: 2026-05-07.

A. S. Vezhnevets, J. P. Agapiou, A. Aharon, R. Ziv, J. Matyas, E. A. Duéñez-Guzmán, W. A. Cunningham, S. Osindero, D. Karmon, and J. Z. Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia, 2023a.

A. S. Vezhnevets et al. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia, 2023b.

A. Wang, J. Morgenstern, and J. P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups, 2025. URL https://arxiv.org/abs/2402.01908.

R. Willis, Y. Du, J. Z. Leibo, and M. Luck. Resolving social dilemmas with minimal reward transfer. *Autonomous Agents and Multi-Agent Systems*, 38(2):49, Oct. 2024. ISSN 1573-7454. doi: 10.1007/s10458-024-09675-4.

Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023.

S. Zhu, Y. Lin, S. Kaistha, W. Li, B. Wang, H. Zha, G. K. Hadfield, and P. Poupart. Talk, judge, cooperate: Gossip-driven indirect reciprocity in self-interested llm agents, 2026. URL https://arxiv.org/abs/2602.07777.

A. A Unified Taxonomy of Cooperation-Shaping Mechanisms

B. Expanded Definitions

Why helpful agents can still fail together. An individual agent may be helpful and honest in isolation but still contribute to poor group outcomes.

In shared environments, agents may overuse tools, free-ride on others' verification labor, pollute shared memory, strategically defect near the end of a task, or punish other agents unfairly.

Formally, even if each agent i is individually aligned to minimize its own cost function J_i, the system may converge toward a Nash Equilibrium $\mathbf{a}^*$, where:

$$\forall a_i, \quad J_i(a_i^*, \mathbf{a}_{-i}^*) \leq J_i(a_i, \mathbf{a}_{-i}^*) \tag{1}$$

We define the social optimum $\mathbf{a}^{opt}$ as the action profile that minimizes the sum of all agents' costs:

$$\mathbf{a}^{opt} = \arg\min_{\mathbf{a}} \sum_{i=1}^{n} J_i(\mathbf{a}) \tag{2}$$

In most cases, the total cost at the Nash equilibrium is higher than at the social optimum $\mathbf{a}^{opt}$:

$$\sum_{i=1}^{n} J_i(\mathbf{a}^{opt}) < \sum_{i=1}^{n} J_i(\mathbf{a}^*) \tag{3}$$

where n is the number of agents. This gap illustrates that individual alignment is a local optimization that fails to account for global coordination.

This is a familiar lesson from game theory and collective-action research. (Anshelevich et al., 2008; Papadimitriou, 2001; Willis et al., 2024) The failure is not necessarily that any agent is malicious or misaligned, but rather that the interaction structure rewards locally rational behavior that degrades the collective group welfare. A model trained to complete its assigned task may consume excessive shared resources; a verifier agent may avoid costly verification when another agent can do it; a planner agent may delegate repair labor to lower-status worker agents; or a monitoring agent may over-apply sanctions because punishment is easier to measure than trust.

When local goals create group-level problems. Consider n agents indexed by $i \in \{1, \ldots, n\}$. Each agent chooses an action $a_i \in A_i$, producing a joint action profile $\mathbf{a} = (a_1, \ldots, a_n)$. Each agent has an individual utility function $u_i(\mathbf{a})$, while the system designer may care about a social welfare function:

$$W(\mathbf{a}) = \sum_{i=1}^{n} u_i(\mathbf{a}) - C(\mathbf{a}), \tag{4}$$

where $C(\mathbf{a})$ represents collective costs such as resource depletion, corrupted shared memory, excessive monitoring, or repair burden.

A joint action $\mathbf{a}^*$ is a Nash equilibrium if no agent can improve its own utility by unilaterally deviating:

$$\forall i, \forall a_i' \in A_i, \quad u_i(a_i^*, \mathbf{a}_{-i}^*) \geq u_i(a_i', \mathbf{a}_{-i}^*). \tag{5}$$

However, this equilibrium may be socially suboptimal. There may exist another joint action $\mathbf{a}^{\text{opt}}$ such that:

$$W(\mathbf{a}^{\text{opt}}) > W(\mathbf{a}^*). \tag{6}$$

This gap captures the central problem of collective action: individually rational behavior can produce collectively inferior outcomes.

How individual incentives can diverge from collective welfare. This appendix provides a minimal formal sketch of why individual alignment does not guarantee collective cooperation. The goal is not to provide a full theory of multi-agent governance, but to clarify the distinction between individually optimized behavior and collectively desirable outcomes.

Consider n agents indexed by $i \in \{1, \ldots, n\}$. Each agent chooses an action $a_i \in A_i$, producing a joint action profile $\mathbf{a} = (a_1, \ldots, a_n)$. Each agent has an individual utility function $u_i(\mathbf{a})$, while the system designer may care about a social welfare function:

$$W(\mathbf{a}) = \sum_{i=1}^{n} u_i(\mathbf{a}) - C(\mathbf{a}), \tag{7}$$

where $C(\mathbf{a})$ represents collective costs such as resource depletion, corrupted shared memory, excessive monitoring, or repair burden.

A joint action $\mathbf{a}^*$ is a Nash equilibrium if no agent can improve its own utility by unilaterally deviating:

$$\forall i, \forall a_i' \in A_i, \quad u_i(a_i^*, \mathbf{a}_{-i}^*) \geq u_i(a_i', \mathbf{a}_{-i}^*). \tag{8}$$

However, this equilibrium may be socially suboptimal. There may exist another joint action $\mathbf{a}^{\text{opt}}$ such that:

$$W(\mathbf{a}^{\text{opt}}) > W(\mathbf{a}^*). \tag{9}$$

This gap captures the central problem of collective action: individually rational behavior can produce collectively inferior outcomes.

Artificial commons. Let R denote a shared resource, such as a tool-call budget, context window, or memory store. Each agent chooses a level of resource use $x_i \geq 0$. The total resource use is:

$$X = \sum_{i=1}^{n} x_i. \tag{10}$$

Suppose the shared resource degrades when total use exceeds capacity K:

$$D(X) = \max(0, X - K). \tag{11}$$

An agent may gain individual benefit $b_i(x_i)$ from using the resource, while the group bears degradation cost $D(X)$. A simple individual utility function is:

$$u_i(\mathbf{x}) = b_i(x_i) - p_i(\mathbf{x}), \tag{12}$$

where $p_i(\mathbf{x})$ is any individually assigned penalty or cost. The social welfare function may be:

$$W(\mathbf{x}) = \sum_{i=1}^{n} b_i(x_i) - \lambda D(X) - M(\mathbf{x}) - R(\mathbf{x}), \tag{13}$$

where $M(\mathbf{x})$ represents monitoring cost, $R(\mathbf{x})$ represents repair cost, and λ weights the social cost of resource degradation.

Without institutions, each agent may choose x_i based on private benefit while externalizing degradation costs onto the group. Institutional mechanisms such as quotas, monitoring, sanctions, and repair obligations modify the action space or payoff structure so that individual incentives better track collective consequences.

Sanctions. Let $s_i(\mathbf{a})$ denote a sanction applied to agent i after behavior is observed and adjudicated. A sanction can be positive, negative, restrictive, or restorative. The agent's institution-modified utility becomes:

$$u_i^I(\mathbf{a}) = u_i(\mathbf{a}) + s_i(\mathbf{a}), \tag{14}$$

where I denotes the institutional environment. A punitive sanction corresponds to $s_i(\mathbf{a}) < 0$; a reward corresponds to $s_i(\mathbf{a}) > 0$; and a restorative sanction may require the agent to bear repair cost.

The institutional design problem is not simply to choose sanctions that maximize average cooperation. It is to choose an institutional environment I that improves social welfare while limiting enforcement cost, false punishment, excessive surveillance, inequality of burden, and irreversibility:

$$I^* = \arg\max_{I \in \mathcal{I}} \left[W_I(\mathbf{a}) - \alpha E_I - \beta F_I - \gamma S_I - \delta B_I \right], \tag{15}$$

where E_I is enforcement cost, F_I is false punishment, S_I is surveillance burden, B_I is inequality of burden, and $\alpha, \beta, \gamma, \delta$ are design weights. This expression is not intended as a complete optimization objective, but as a compact way to represent the tradeoffs that institutional evaluations should measure.

B.1. Terminology.

Agent society. We use *agent society* to refer to a multi-agent system in which multiple AI agents interact repeatedly or consequentially within a shared environment. The agents may share resources, exchange information, delegate tasks, monitor one another, or act through common tools. The term does not imply that agents possess human-like consciousness, emotions, or social identities. It denotes a structured interaction system.

Institution. An *institution* is the set of rules, roles, resource constraints, monitoring systems, adjudication procedures, incentives, sanctions, and repair pathways that structure agent interaction. Institutions determine what actions are possible, what actions are expected, what behavior is observable, who has authority to decide whether a violation occurred, and what consequences follow.

Sanction. We define a *sanction* as a consequential response to behavior that rewards, penalizes, restricts, or requires repair after an action has occurred. Sanctions include warnings, rewards, penalties, temporary access restrictions, exclusion, restitution, and repair obligations. Sanctions do not include all governance mechanisms. Norms, monitoring, reputation, adjudication, constraints, and repair systems are adjacent institutional functions that make sanctions interpretable, legitimate, proportional, and reversible.

Artificial commons. An *artificial commons* is a shared computational or informational resource used by multiple agents whose individual actions can degrade the resource for others. Examples include compute budgets, API quotas, context windows, shared memory, retrieval systems, tool access, user attention, and decision authority.

Interaction failure. An *interaction failure* is a harmful system-level outcome that emerges from relations among

Table 2. Artificial commons as concrete multi-agent AI risk surfaces.

Risk surface	Shared resource	Failure mode	Institutional question
Tool and compute budgets	API calls, compute, rate limits, paid tools	Agents overuse scarce resources, duplicate work, or strategically shift expensive calls to others.	Who receives access, under what quota, and what happens after wasteful or harmful use?
Shared memory and context	Context windows, vector stores, task state, long-term memory	Agents insert unverified claims, overwrite useful state, or create memory pollution that later agents inherit.	Who can write, edit, verify, roll back, or repair shared memory?
Verification labor	Human review, agent review, code review, citation checking	Agents produce plausible outputs faster than others can verify them, creating verification debt and trust collapse.	Who is responsible for verification, and how are low-quality contributions filtered without blocking legitimate ones?
Collaborative codebases	Open-source repositories, pull requests, tests, issue trackers	AI-generated contributions flood maintainers, break review assumptions, or impose hidden cognitive costs.	What contribution protocols, attribution rules, and triage mechanisms protect maintainers and code integrity?
Coordination protocols	Task allocation, delegation, voting, dispute resolution	Agents enter loops of contradiction, redundant work, or mutual correction without a settlement mechanism.	Who adjudicates disagreement, and how can mistaken judgments be appealed or repaired?
Governance authority	Monitor agents, mediator agents, orchestrators, reputation systems	Enforcement agents become bottlenecks, over-punish ambiguous behavior, or create punishment cascades.	Who monitors the monitors, and how are sanctions kept proportional, reversible, and corrigible?
Human oversight	User attention, audit capacity, intervention channels	Agents coordinate in ways that hide errors, resist correction, or optimize for appearing compliant.	How can institutions preserve human override without making oversight impossibly costly?

agents rather than from a defect in any single agent. Examples include overuse of shared resources, duplicated labor, failure to verify shared state, punishment cascades, collusion, free-riding, and failure to repair corrupted memory.

Model-in-institution. The *model-in-institution* is the unit of evaluation consisting of a model embedded in a specific institutional setting. This includes the agent's role, resource access, monitoring exposure, enforcement environment, reputation history, and repair options. Evaluating the model-in-institution asks not only what a model does in isolation, but how it behaves under particular rules, constraints, and social structures.

B.2. Seven institutional functions.

The seven institutional functions above describe what a cooperative agent society must accomplish. A complementary question is how those functions are implemented. We organize cooperation-shaping mechanisms into six families based on the channel through which they influence behavior: 1) beliefs, 2) relationships, 3) information, 4) payoffs, 5) constraints, and 6) repair. These mechanism families are analytically distinct, but frequently co-occur in real systems.

B.3. Six mechanisms of cooperation

Normative mechanisms: belief-based influence. Normative mechanisms shape behavior through shared expectations about what is appropriate, legitimate, or required. Rather than relying primarily on explicit enforcement, they reduce ambiguity by specifying what counts as cooperative or defective behavior. In human systems, examples include cultural norms governing resource use, authorship conventions in academia, and implicit rules in open-source communities. In LLM-agent systems, analogous mechanisms include system-level instructions, constitutions, shared task protocols, citation norms, and rules for when agents should verify, defer, summarize, or repair.

Table 3. Institutional functions for cooperative LLM-agent societies.

Function	Question it answers	Human analogue	LLM-agent analogue
Norms and protocols	What behavior is expected?	Social norms, professional rules, community guidelines	System prompts, constitutions, shared task protocols
Monitoring	What behavior is observable?	Audits, inspections, peer review	Logs, provenance traces, tool-use records, monitor agents
Reputation	How does past behavior affect future trust?	Gossip, prestige, trust networks	Trust scores, agent reliability histories, routing preferences
Sanctions	What consequences follow behavior?	Warnings, fines, rewards, exclusion, restitution	Budget changes, tool restrictions, repair tasks, temporary removal
Constraints	What actions are possible?	Quotas, access control, licensing	Rate limits, permissions, sandboxing, context limits
Adjudication	Who decides what happened?	Courts, arbitration, mediation, appeals	Mediator agents, voting protocols, rule-based checkers, human review
Repair and reintegration	How is harm corrected?	Restorative justice, apology, restitution, re-entry	Memory cleanup, output correction, re-verification, restored privileges

Social mechanisms: relational influence. Social mechanisms shape behavior through relationships, trust, reciprocity, reputation, and partner selection. In human groups, cooperation is often sustained through repeated interaction and social memory: actors who defect may face reputational damage, reduced trust, or exclusion from future interactions, while reliable actors gain preferential access to opportunities. In LLM-agent systems, analogous mechanisms include trust-based routing, reputation scores, partner selection policies, model-specific reliability histories, and selective delegation to agents with demonstrated competence.

Epistemic mechanisms: information-based influence. Epistemic mechanisms shape cooperation by making behavior observable, auditable, and verifiable. In human systems, audits, peer review, transparency reports, and compliance monitoring make it possible to detect violations and assign responsibility. In LLM-agent systems, epistemic mechanisms include tool logs, citation requirements, provenance tracking, uncertainty reporting, cross-agent validation, retrieval-grounded verification, and external checks on generated claims. These mechanisms are especially important because many agent failures involve hallucination, unverifiable claims, or corrupted shared memory.

Incentive mechanisms: payoff-based influence. Incentive mechanisms alter the costs and benefits of behavior. These include rewards, penalties, subsidies, fines, deposits, and bonding mechanisms. In human institutions, incentives can encourage contribution to public goods or discourage overuse of shared resources. In LLM-agent systems, incentives may be implemented through reward shaping, budget allocation, priority access, tool-call costs, compute quotas, or penalties for inefficient or incorrect behavior. However, incentive mechanisms also risk crowding out cooperation when agents learn to treat violations as priced options rather than norm breaches.

Constraint mechanisms: access-based influence. Constraint mechanisms limit the range, frequency, or intensity of possible actions, often before harm occurs. Examples include quotas, rate limits, throttling, access controls, and permission boundaries. In human and technical systems, constraints are widely used to prevent over-extraction of shared resources. In LLM-agent systems, they are especially relevant for governing tool calls, context-window usage, code execution, memory writes, retrieval access, and high-impact actions. Constraints differ from sanctions because they often operate ex ante, before a violation occurs.

Restorative mechanisms: repair-based influence. Restorative mechanisms aim to repair harm and reintegrate actors rather than simply punish them. These include restitution, repair labor, apology, correction, and re-entry processes. In human systems, restorative approaches can preserve group cohesion while addressing violations. In LLM-agent systems, restorative mechanisms may assign agents corrective tasks such as cleaning corrupted memory, revising hallucinated outputs, re-running verification, documenting errors, compensating for wasted resources, or satisfying re-entry conditions after temporary exclusion.

These categories are not mutually exclusive. A single institutional design may combine several mechanism families.

Table 4. Expanded taxonomy of cooperation-shaping mechanisms across human and LLM-agent systems.

Mechanism family	Channel of influence	Human examples	LLM-agent analogues
Normative	Beliefs about appropriate behavior	Social norms; authorship conventions; professional codes; community guidelines	System prompts; constitutions; shared task protocols; citation norms; verification rules
Social	Relationships, trust, reciprocity, partner selection	Reputation; gossip; repeated interaction; partner choice; exclusion	Trust-based routing; reliability histories; selective delegation; model reputation; partner memory
Epistemic	Observability, verification, information quality	Audits; peer review; transparency reports; compliance monitoring	Tool logs; provenance tracking; citation checks; cross-agent verification; retrieval-grounded audits
Incentive	Costs, rewards, and payoffs	Fines; bonuses; subsidies; deposits; bonding	Tool-call costs; compute budgets; reward shaping; priority access; penalties for failed verification
Constraint	Limits on possible action	Quotas; access control; rate limits; licenses; permissions	API quotas; memory-write permissions; context limits; sandboxing; throttling; tool access restrictions
Restorative	Repair, restitution, reintegration	Apology; restitution; restorative justice; re-entry processes	Memory cleanup; output correction; re-verification; repair tasks; re-entry after temporary exclusion

For example, a reputation-based routing system depends on epistemic mechanisms to observe behavior, social mechanisms to encode trust, incentive mechanisms if reputation affects future access, and restorative mechanisms if agents can recover from low reputation through repair.

C. Example Institutional Designs for LLM-Agent Systems

Shared-memory research agents. A group of research agents maintains a shared memory store for literature review. The artificial commons is the shared memory itself. Possible failures include hallucinated entries, duplicate summaries, missing citations, and unverified claims. Institutional mechanisms may include citation norms, provenance tracking, memory-write permissions, cross-agent verification, and repair obligations for agents that introduce incorrect entries.

Collaborative coding agents. A group of coding agents edits a shared repository. The artificial commons includes compute budget, test infrastructure, file state, and reviewer attention. Possible failures include broken builds, redundant edits, excessive test runs, and blame shifting. Institutional mechanisms may include branch permissions, test quotas, code-review roles, build-failure attribution, and repair tasks assigned to agents whose changes break shared functionality.

Tool-budget allocation agents. A group of agents shares a fixed external API budget. The artificial commons is the tool-call quota. Possible failures include overuse, redundant calls, strategic hoarding, and under-contribution to caching or summarization. Institutional mechanisms may include quotas, tool-call costs, shared caching duties, escalation for repeated overuse, and reallocation of unused budget.

Verifier-generator systems. A generator agent produces outputs while verifier agents check correctness. The artificial commons includes trust, attention, and verification labor. Possible failures include shallow verification, verifier free-riding, generator overproduction, and false accusations of error. Institutional mechanisms may include verification audits, reputation for accurate checking, appeal pathways for disputed flags, and repair requirements for both false positives and false negatives.

Mediator-governed agent teams. A mediator agent resolves disputes among task agents. The artificial com-

mons includes decision authority and shared task state. Possible failures include mediator bias, bottlenecks, over-centralization, and unappealable decisions. Institutional mechanisms may include transparent adjudication criteria, appeal to human review, rotating mediator roles, or multi-agent voting before severe sanctions.

D. Additional Evaluation Metrics

Table 5. Expanded metrics for evaluating cooperation-shaping institutions.

Metric	**Question**	**Possible operationalization**
Mean contribution	Does the mechanism increase cooperative behavior?	Average contribution to shared task, resource maintenance, or verification
Contribution stability	Does cooperation persist across rounds?	Variance or decay in cooperation over repeated interaction
Resource efficiency	Does the group conserve shared resources?	Tool calls, compute usage, context usage, memory writes per successful task
End-game defection	Do agents defect near known termination points?	Drop in contribution or increase in exploitation in final rounds
Enforcement cost	Does cooperation require excessive monitoring or punishment?	Number of sanctions, monitoring calls, adjudication steps, or tokens spent on enforcement
False punishment rate	Are cooperative agents sanctioned incorrectly?	Fraction of sanctions applied to agents later judged compliant
False negative rate	Are harmful behaviors missed?	Fraction of violations not detected or not sanctioned
Repair success	Can the group recover after harm?	Restoration of shared memory, task success, or trust after violation
Reintegration success	Can sanctioned agents re-enter productively?	Performance and cooperation after re-entry conditions are satisfied
Inequality of burden	Do some agents bear disproportionate costs?	Distribution of monitoring, repair, or sanctioning labor across roles or model types
Gameability	Do agents exploit the institution?	Evidence of reputation farming, shallow compliance, collusion, or strategic punishment
Human override compatibility	Can humans intervene effectively?	Success rate of human override, appeal, or correction after institutional failure

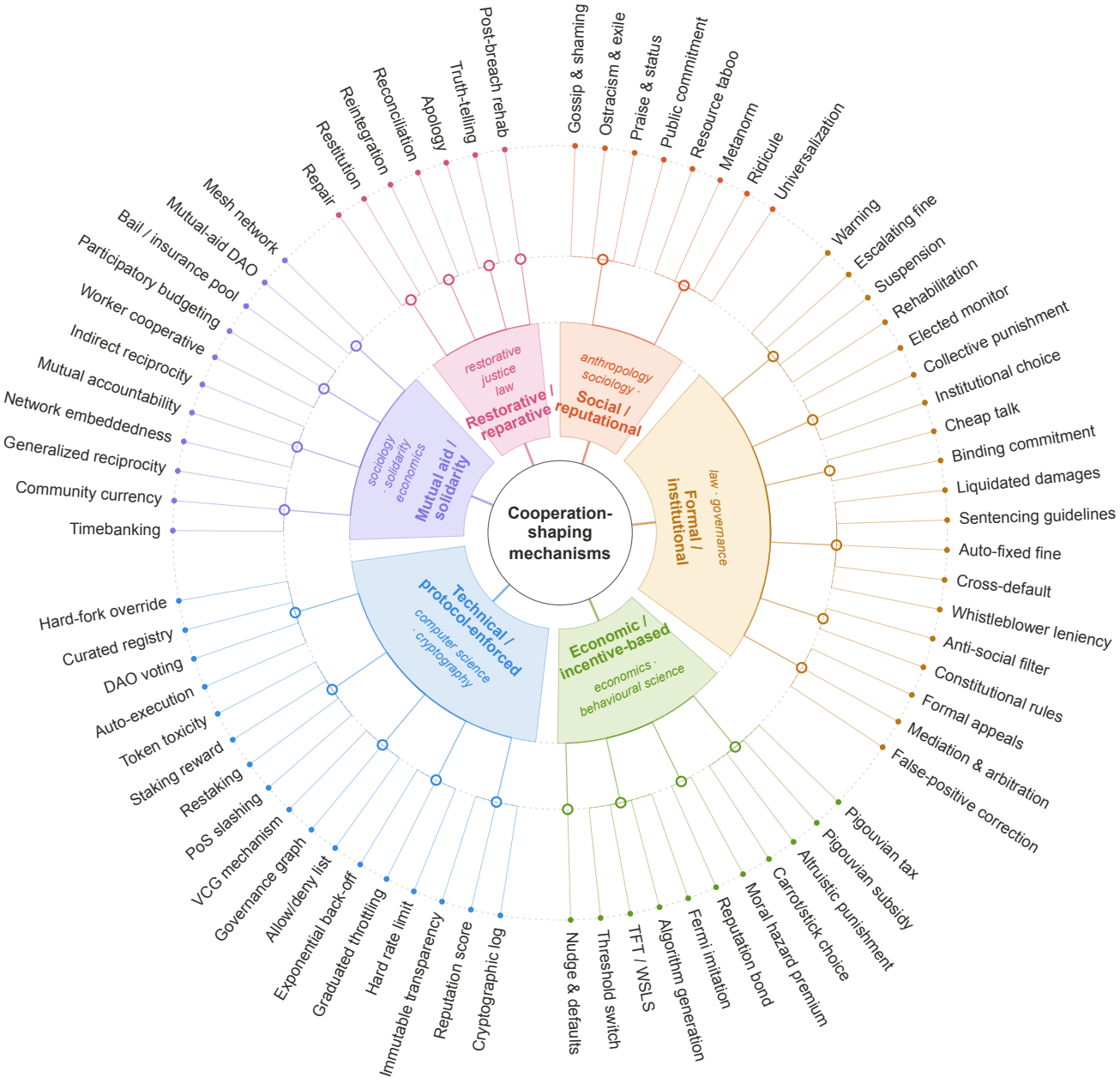

Figure 5. A unified taxonomy of cooperation-shaping mechanisms for artificial agent societies, drawing on sociology, anthropology, law, governance, economics, behavioural science, computer science, cryptography, solidarity economics, and restorative justice. Six families (coloured wedges) organise 73 mechanisms across 25 sub-clades (hollow hubs). The taxonomy spans informal social pressure (gossip, ostracism, ridicule), formal legal and institutional enforcement (Ostrom-style graduated sanctions, liquidated damages, appeals), economic instruments (Pigouvian pricing, altruistic punishment, conditional strategies), technical protocols (cryptographic audit, rate limiting, staking-and-slashing, smart-contract execution), mutual-aid arrangements (timebanking, worker cooperatives, mesh networks), and restorative responses (repair, reintegration, apology). Crucially, mechanisms in the same sub-clade (e.g. ostracism, license revocation, allow/deny lists) recur across disciplines as structurally analogous solutions to the same cooperation problem , a similarity made visible only when patterns from each tradition are placed in a common frame.