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

Van QT Truong [*1 2 3] **Xuanqiang Angelo Huang** [*2 3 4 5] **Erivan Inan** [2 3] **Ryan Faulkner** [2 3] **Joël Naoki Christoph** [6] **Terry Jingchen Zhang** [†2 3] **David Guzman Piedrahita** [†2 3 4] **Zhijing Jin** [†2 3 7]

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, and decision authority. These shared resources form artificial commons, where well-known collective-action failures can reappear. Current alignment techniques, which primarily evaluate individual models for their appropriate responses to humans, are insufficient to address these emergent system-level dynamics. **This position paper argues that future AI safety research should evaluate agent societies as governed systems, in addition to collections of individually aligned models.** We use sanctions to illustrate one axis where institutional rules acquire practical consequences and reveal the broader institutional structures that shape cooperation. We make three contributions. **First**, we argue for evaluating the *model-in-institution*, since individually aligned agents can still produce collectively harmful outcomes through agent-group interactions. **Second**, we distinguish sanctions from the broader institutional functions that give them meaning and legitimacy. **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. **Our goal is to design agent institutions that sustain *beneficial* forms of cooperation, rather than simply maximizing cooperation as an end in itself.**

[*]Equal co-first authors. [†]Equal co-last authors. [1]University of Pennsylvania [2]Jinesis Lab, University of Toronto & Vector Institute [3]EuroSafeAI [4]ETH Zürich [5]Institute for Decentralized AI, Oxford [6]Harvard Kennedy School [7]Max Planck Institute for Intelligent Systems, Tübingen, Germany. Correspondence to: Van QT Truong <heyvan@cs.toronto.edu>, Xuanqiang Angelo Huang <hxuanqiang@ethz.ch>, Zhijing Jin <zjin@cs.toronto.edu>.

Preprint. July 1, 2026.

1. Introduction

Consider two ways an autonomous agent might execute a $10M purchase on behalf of its principal: a cryptocurrency transfer to an address, or a wire through the conventional banking system. Suppose that the agent corrupts a single character of the recipient identifier. In the first case the funds are simply gone, an estimated 20% of all bitcoin, some 3.7 million BTC, is already permanently inaccessible through key loss and misdirected transfers (Popper, 2021)[1]. In the second, a web of institutions exists to prevent cascading interaction failures: when Citibank mistakenly wired $894M of its own funds to Revlon's lenders in 2020 instead of a $7.8M interest payment, recall notices, injunctions, and ultimately appellate restitution doctrine returned the money (Walker, 2020). The agent's intent, and its alignment with its principal, was identical in both scenarios. What differed was the institutional setting (i.e. cryptocurrency vs. traditional banking) in which it acted.

Autonomous systems already manage allocation of compute clusters (Delimitrou and Kozyrakis, 2014), review code (Tang et al., 2024), operate vehicles (Ma et al., 2020), and broker trades (Fan et al., 2025). And we need not speculate about how interacting automated systems fail. In 2011, two pricing algorithms on Amazon, each following a locally sensible rule, bid an out-of-print biology textbook to $23.7 million (Eisen, 2011). In the 2010 Flash Crash, interacting automated traders erased nearly $1 trillion of market value in minutes (Kirilenko et al., 2017); institutional safeguards, such as trade-cancellation rules, were designed only afterward. In addition to these, LLM agents introduce another complication because they fail in ways systematically unlike human error, shaped by training distributions rather than bodily fatigue or inattention (McCoy et al., 2024). Therefore, institutions calibrated to human failure modes cannot be presumed to transfer; at best, they provide inspiration for the design of AI-native institutions.

[1]Even here the exceptions prove the rule: when losses grow large enough, ostensibly trustless communities improvise institutional repairs, as in the Ethereum DAO fork (Dupont, 2017) following a hack draining 3.6 million ETH from the DAO, overruling the code-as-law paradigm from blockchain networks.

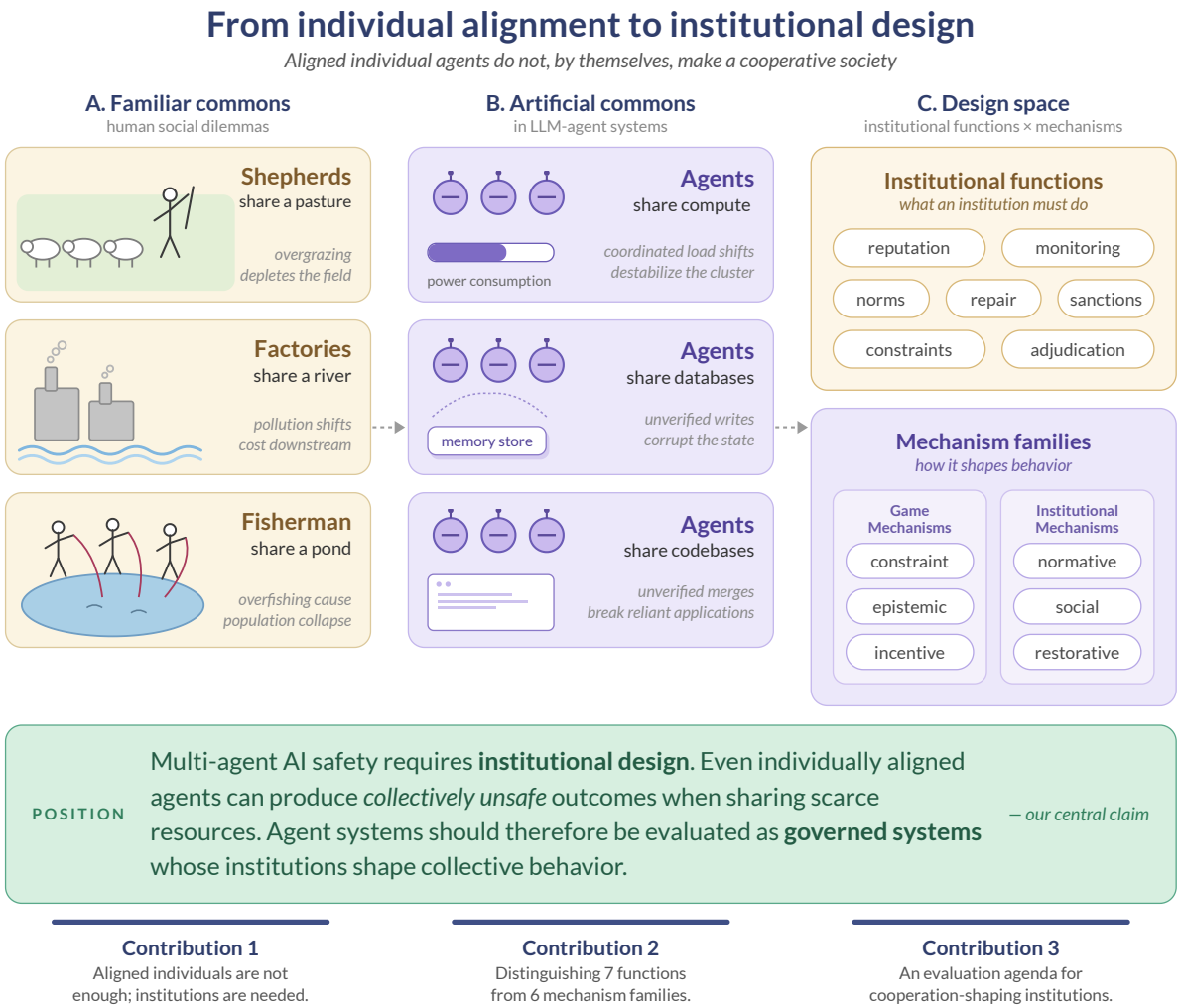

Figure 1. From individual alignment to institutional design. Familiar human commons exhibit recurring cooperation dilemmas; analogous dynamics arise when LLM agents share computational resources; and the institutional design space includes seven institutional functions and six mechanism families.

These episodes are instances of an old problem in a new population. Effective cooperation is a cornerstone of functional society (Axelrod and Hamilton, 1981; Gross et al., 2025; Hamilton, 1964; Nowak, 2006), yet it is continually threatened by the tension between individual advantage and collective sustainability. Common-pool resources (CPR) in human societies such as grazing land, irrigation systems, open-source software, and many digital platforms are vulnerable to the *Tragedy of the Commons*; a class of social dilemmas which depends on promoting cooperative behavior and mitigating the free-rider problem (Baumol, 2004; Hardin, 1968). The question is not *whether* multi-agent AI systems will face social dilemmas, but whether we will have the mechanisms to govern them when they do (Faulkner et al., 2026; Jaques et al., 2019; Leibo et al., 2017; Piatti et al., 2024; Guzman Piedrahita et al., 2025; Tewolde et al., 2026). Historically, systems that depend on institutions to function sustainably, and perhaps suffered some consequential activity without, have eventually acquired such mechanisms: commercial law for trade, financial regulation for markets, computer-misuse statutes for networked computing Agentic AI is currently a sphere of consequential activity largely without them. As commercial activities between AI agents materialize, institutions will likely be the backbone supporting these endeavours.

Lessons and limits of human cooperation. We adopt the analogy to human cooperation in functional terms, without assuming that LLM agents possess human motivations, identities, moral emotions, or social attachments (Anthis et al.,

2025; Wang et al., 2025)[2]. Human governance systems are the best-documented solutions we have to a recurring design problem: regulating repeated interaction under resource constraints, partial observability, asymmetric power, and costly enforcement. They are also a source of cautionary lessons, because cooperation is not wholly benign. The same mechanisms that sustain it have historically produced surveillance, coercion (Gross et al., 2025) and exclusion of out-groups (Tomuschat, 2006). Patterns taken from the human world may therefore transfer to multi-agent AI systems only partially, and sometimes with their pathologies intact (Wang et al., 2025).

The institutional question is also no longer confined to software. Autonomous systems increasingly act in shared physical space, from robotaxis to warehouse and delivery automation (Ackerman, 2026; Korosec, 2025; The Robot Report Staff, 2025; Dafoe et al., 2021). These systems need not constitute a fully autonomous "AI economy" to raise institutional questions: once agents share scarce resources and common spaces, their behavior is shaped by the rules, oversight, and repair procedures of the environment along with their individual training (Huang et al., 2026; Piatti et al., 2024). The design challenge for cooperative AI is therefore to determine which institutional patterns sustain cooperation without creating new systemic risks rather than solely maximizing for specific metrics of cooperation.

Our contributions. This position paper makes three contributions. **First**, we show how institutions are necessary for AI safety, building on Hammond et al. (2025)'s statement that group-level failures can arise even when individual agents appear aligned. **Second**, we distinguish seven institutional functions needed for cooperative agent societies (e.g., norms, monitoring, reputation, sanctions, constraints, adjudication, and repair) and organize the underlying mechanisms into six families according to how they shape behavior (e.g., beliefs, relationships, information, payoffs, constraints, and repair). **Third**, we map each function to concrete LLM-agent design choices and propose evaluation criteria for testing whether these institutions sustain cooperation across models and repeated interactions.

The remainder of this paper proceeds as follows. Section 2 motivates institutional design as a complement to model-level alignment. We then introduce a taxonomy of institutional functions and cooperation-shaping mechanism families for multi-agent AI systems. Next, we examine how institutional mechanisms can fail, including incentive distortions, reputation gaming, and authority capture. Finally, we propose an evaluation agenda for testing cooperation-sustaining institutions across different agent populations, resource environments, and enforcement architectures, before concluding with implications for AI safety and governance.

[2]Recent work by Anthropic (Sofroniew et al., 2026) shows some evidence of emotion latent directions within the models. However, this phenomenon is not fully understood. For the purpose of this paper, we assume machines in general don't have consciousness or emotions typically directed to mammals.

2. From Individual Model Alignment to Institutional Design

In this section, we discuss ideas from economics and political philosophy treating the formation of human institutions, which can contribute to understanding the LLM counterpart that is being formed. The literature on its formation splits along two broad traditions, tracing back to the Scottish Enlightenment and Hayek's account of spontaneous order (Hayek, 1978). Institutions can emerge bottom-up through repeated interaction, or they can be built through deliberate design.

Top-down tradition. On the one hand, in the top-down direction, institutions are rules, mechanisms, and governance arrangements intentionally designed to align incentives, resolve conflicts, and protect collective outcomes (Buchanan and Tullock, 1965; Elster, 1977; Maskin, 2008; Myerson, 1981). For example, the American 1929 market crash can be largely attributed to the absence of institutional guardrails limiting unrestrained individual rationality (Kindleberger and Aliber, 2005).

Bottom-up tradition. On the other hand, in the bottom-up tradition, institutions are stable patterns of behavior sustained by conventions and repeated interaction (Aoki, 2001; Calvert, 1995; Hayek, 1978; Lewis, 2008; North, 1990; Schelling, 1960). Ostrom's studies of small-scale commons of inshore fisheries and alpine meadows show that groups of 50 to 15,000 users can self-organize effective governance through specific local rules (Ostrom, 1990).

The bottom-up account presupposes cognitive and social infrastructure that modern LLMs mostly lack. First, agents need persistent memory to sustain cooperation through reciprocity and punishment strategies (Fudenberg and Maskin, 1986). Second, they require the capacity to form expectations over others' strategies, analogous to human theory of mind (Premack and Woodruff, 1978) and social world models (Huang et al., 2024), without which they cannot anticipate behavior and respond adaptively within a group. Third, even when cooperative equilibria exist, selecting among them may require collective intentionality (Bacharach, 2006; Tomasello, 2016): the capacity to reason as "what should *we* do" rather than "what should *I* do," which current LLMs have no native mechanism to support or any equivalent selection criteria.

Stable institutions in AI societies do not presently emerge. Current LLM agents, for example the recent autonomous AI assistants (Steinberger, 2026), often lack the conditions for bottom-up institutions to stabilize; out-of-the-box, they carry no persistent identity or reliable memory for reputation to attach to, and have no Folk-Theorem-style (Hu et al., 2026) stakes in future social interaction. Without these elements, institutions cannot spontaneously form. The current file-based memory approaches are often insufficient (Wu et al., 2025). Even a community of *superintelligent* models would not, by mere addition of capability, build the institutions required to sustain its own cooperation (Trivedi et al., 2026); they would be in principle able to accomplish very difficult tasks, yet they won't be able to coordinate in large numbers to build the AI equivalent of a cathedral in terms of cooperation, 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, decoupled from the ones of a hypothetical principal. Even if the above problems are solved, in the next section we argue how the bottom-up direction could impose not-negligible risks to human society, meaning that among the two, a more defensible path to safe cooperative AI is therefore learning towards *top-down* design, where humans specify the rules and structures within which agents operate.

2.1. Necessity of 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. We offer two complementary replies, one is a concerning who should choose the values institutions enforce and the other on the limits of autonomous mechanism design).

Institutions encode values that ought to be pro-humanity. First, institutional design is fundamentally a question of values. Institutions are not, in themselves, neutral coordination infrastructure; they encode commitments about what agents *ought* and ought *not* do, whose interests count, and how trade-offs are resolved (Knight, 1992; North, 1990). Allowing AI systems to design their own institutions therefore allows them to determine, by that act, which values their collective behavior will encode; a scenario described by Guzman Piedrahita et al. (2026) in their AI governance agenda that we ought to prevent. The case for human-designed institutions is not that AI systems will always be incapable of 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.

Capable AI does not remove the need for human design. Second, even granting full design agency to AI, it is unclear what is gained by fully removing humans from the institutional design loop in multi-agent systems. Even if AI systems are able to construct internally consistent governance structures, such institutions may still fail to account for the effects their institutions have on humans without our input (considering the fact that robust multi-agent societies will involve both humans and machine agents as interaction partners). Even when individual objectives are well-specified, communities of agents can still yield inefficient equilibria (Hardin, 1968). This gap between the best achievable equilibrium and the social optimum is what the Price of Anarchy and Price of Stability literature formalizes (Anshelevich et al., 2008; Papadimitriou, 2001; Willis et al., 2026), and manifests in the four cooperation dilemmas shown in Figure 2. Mass miscoordination is, in part, caused by poor institutional design. The design of a capable mechanism narrows this gap, but does not generally close it everywhere (Huang et al., 2026); for many classes of games, no incentive-compatible mechanism implements the welfare-optimal outcome. Intelligence applied to mechanism design therefore improves coordination but does not remove the need for normative choices about which equilibria ought to be preferred or whose welfare should be prioritized between and among human and machine agents. Eventually, it is, again, a value judgment that humans need to take as primary stakeholders because we hope to reap the benefits while mitigating potentially catastrophic harms to society.

This becomes especially salient in settings where multi-agent systems are embedded in human environments (Dafoe et al., 2021). In such hybrid systems, purely AI-generated institutions risk becoming self-consistent, but misaligned with human purposes (Shen et al., 2026; Sorensen et al., 2024). An analogy of this paradox is when text is syntactically correct, but semantically irrelevant (Norvig, 2012). For this reason, we assert that institutional design should not be left as an afterthought, or even delegated to AI agents to sort out among themselves.

Our agenda builds directly on the multi-agent risk taxonomy outlined by Hammond et al. (2025). Their report cataloged the principal interaction risks facing advanced multi-agent systems (i.e., miscoordination, conflict, collusion, etc.) and presented a survey of potential mitigation strategies framed as tractable research directions for the field. The report alluded to MAS institutional response in disparate sections, but did not expand on this topic at length. Prior to the start of the Cooperative AI subfield, Dafoe et al. (2021) separately called attention to the necessity of bringing together clusters of AI-AI, AI-human, and human-human efforts so that we, as a field, could make meaningful progress towards socially valuable AI. We took inspiration from both as guiding mo-

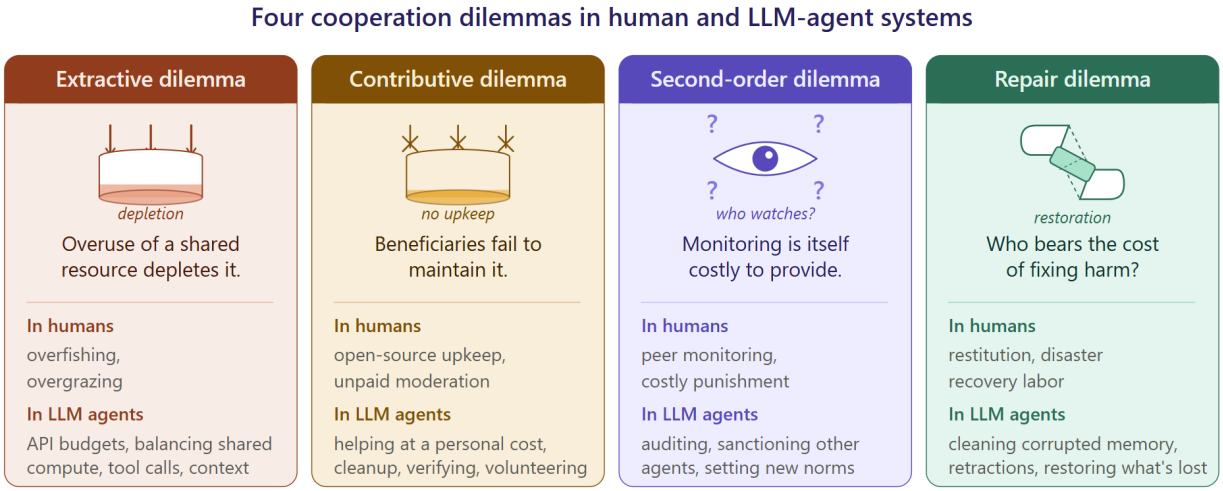

Figure 2. Four cooperation dilemmas that recur in shared-resource settings, each shown with its core tension and parallel examples in human and LLM-agent systems.

tivation for our work. We focused our efforts on framing out a backbone, or scaffold, onto which existing MAS literature and new institutional functions could be mapped. We agree with Dafoe et al. (2021) and Hammond et al. (2025) that human society is unlikely to stumble upon cooperative intelligence by accident, or as happenstance, from simply working on other AI research problems.

3. Institutional Functions for Cooperative Agent Societies

We organize cooperation-shaping mechanisms into six families, see figure 1C. Three correspond to components formalized in the Bayesian game (Harsanyi, 1967): a tuple $\mathcal{G} = (N, A_i, \Theta_i, p, u_i)$, where N is the set of agents, A_i is agent i's action set, Θ_i is its type space, $p \in \Delta(\Theta)$ is a common prior over the joint type space, and $u_i : A \times \Theta \to \mathbb{R}$ is its payoff function. Each component offers the institutional designer a distinct lever. Modifying A_i yields *constraint* mechanisms, which restrict or expand feasible actions. Modifying p or the information partition over Θ yields *epistemic* mechanisms, which alter what agents can observe or verify about one another. Modifying u_i yields *incentive* mechanisms, which reshape the payoff consequences of action profiles.

These three levers operate within a fixed game. Yet does not specify how agents interact within the game $\mathcal{G}$. Institutional mechanism introduce three additional families to address dimensions that the formalism treats as exogenous or omits entirely but are common in human institutions. *Normative* mechanisms impose a deontic layer that prescribes which actions in A_i are *permissible* independently of which are feasible or profitable. *Social* mechanisms shape the interaction structure: who is in N, who interacts with whom, and whether agents can select or exclude partners (Nowak, 2006). *Restorative* mechanisms shape what happens after deviation: whether and how agents re-enter cooperative arrangements, make restitution, or receive corrective feedback (Ostrom, 1990). Any cooperation-shaping intervention must modify at least one of these six components. 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.

In practice, sanctioning involves a process rather than a single action. At minimum, cooperation-shaping institutions must separate three stages (as shown in Figure 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 can be disproportionate, irreversible, or easy to game. This is especially important in LLM-agent systems because intent is often ambiguous, states are only partially observable, 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

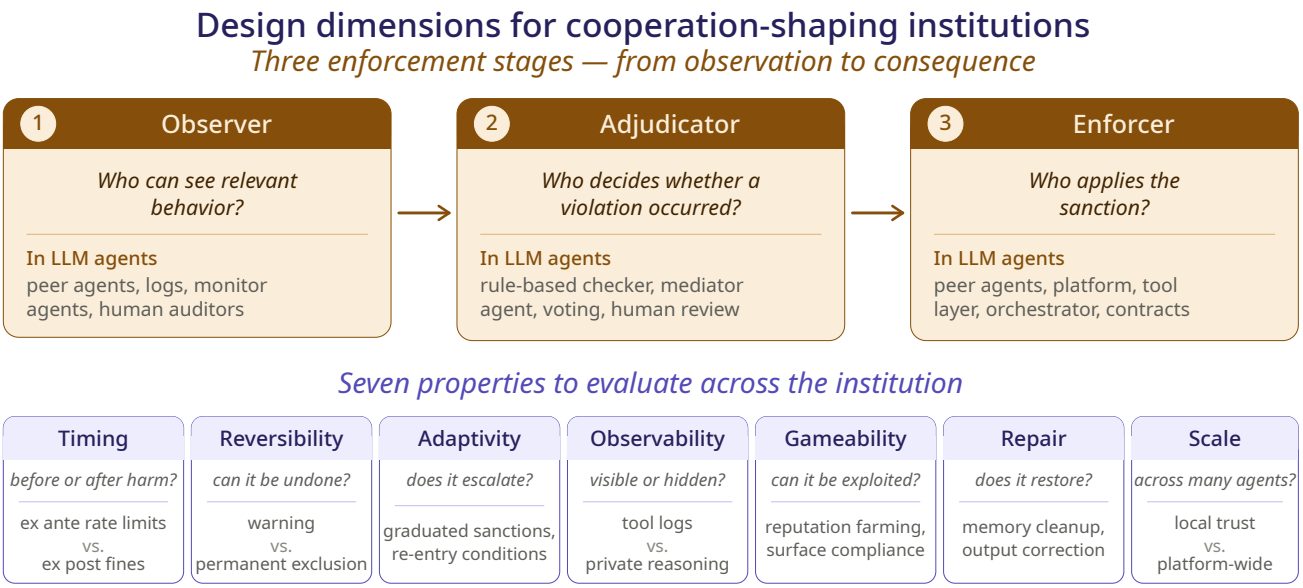

Figure 3. Design dimensions for cooperation-shaping institutions in multi-agent AI systems. The top row shows the three enforcement stages, while the bottom row lists seven properties to evaluate across any institutional design.

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, 1990). 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.

Sanctions need to be generalizable to high intelligence. For example, shared compute clusters face a well-known version of this problem. Usage follows a Pareto-like distribution: a small fraction of users accounts for most consumed resources (Yalim, 2020), and the standard response is a top-down scheduler such as Slurm's fairshare, which decays the priority of heavy consumers. But clusters are fixed environments: the resource types, the job semantics, and the load patterns are known in advance or inferrable (Delimitrou and Kozyrakis, 2014), therefore, a well-designed scheduling algorithm can encode graduated responses once and apply them indefinitely. Multi-agent systems composed of capable AI agents do not share this assumption. Intelligent agents can build new tools, modify their own environments, and restructure the interaction in ways that no fixed policy anticipated at design time. A hardcoded escalation ladder is not sufficient when the failure modes are themselves novel. Graduated sanctions must instead be implemented generalizably, adapted to each case as it arises. A possible escalation pathway within this direction is shown in Figure **??**. 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 or synergistic (Herrmann et al., 2008): *adding sanctions does not monotonically improve cooperation*. Naïvely adding rules might actually hurt the group since sanctions 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): faced with persistent late pickups, ten daycare centres in Haifa introduced a small monetary fine for parents arriving more than ten minutes after closing time. However, rather than reducing

lateness, it doubled. Introducing a price reframed a moral obligation as a purchasable service. 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.

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 (Garcin et al., 2009; Resnick et al., 2000). 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; Krakovna et al., 2020; Skalse et al., 2022).

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 (Acemoglu et al., 2004; Ostrom, 1990), 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 in principle be acutely counterproductive among optimizing agents.

Figure 4 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. Table 2 links the six families to the four cooperation dilemmas as a practical guide.

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, but as a multidimensional outcome shaped by the seven institutional functions outlined in Figure 1C: norms and rules, reputation such as partner history, monitoring, applied sanctions, resource and action constraints, adjudication, and available repair pathways.

Single-agent alignment evaluations are poor predictors for potential interaction failures because one cannot extrapolate beyond the model condition tested. Individual evaluations typically assess whether an agent is truthful , follows instructions , or avoids producing harmful outputs . Institutional evaluations, by contrast, should logically concern itself with properties of the interactive system itself. 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?

Current multi-agent benchmarks often ask whether agents cooperate under a fixed interaction structure (Li et al., 2023; Piatti et al., 2024; Guzman Piedrahita et al., 2025; Vezhnevets et al., 2023; Wu et al., 2023). 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., 2025). 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., 2023). 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 by 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 (Hammond et al., 2025). 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 (Shapira et al., 2023). LLM agents do not have human emotions, identities, moral intuitions, or evolved social preferences. Even if they have correlates that could be interpreted using human categories (Sofroniew et al., 2026), LLM agents do not experience guilt, shame, loyalty, resentment, or forgiveness in the human sense. Trivially, they are not humans. Nonetheless, the human analogy is useful to identify recurring institutional problems: resource depletion, costly monitoring, punishment, exclusion, reputation, boundary formation, repair, and reintegration. We do not need to debate whether LLM agents possess emotions to know with certainty that algorithmic, or machine, disruptions have both material and emotional consequences in the physical world (Dafoe et al., 2021) . In October 2025 Amazon Web Services (AWS) outage impacted thousands of smart-beds, leaving users frustrated and without sleep (Harding, 2025). Although this incident didn't involve LLM agents, these problems can arise in any repeated multi-agent system with shared resources and a distant chain of custody.

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. As Hammond et al. (2025) note, cooperation can be harmful when it enables collusion, 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 will not be solved by retraining models *ad hoc*. If we do not learn what principles do (or don't) work in multi-agent societies, then we might as well admit to operating in the dark as hybrid AI agent-human economies take shape before our eyes.

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., memory, compute, codebases, or any shared resource) 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, 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, or repair, and 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.

8. Impact Statement

This position paper argues that the safety of multi-agent AI systems should be evaluated at the level of institutions, not only individual models, with the aim of preventing interaction failures. In shared digital, economic, and physical environments, these failures could arise as unsustainable

resource exhaustion, collusion, punishment cascades, and irreparable harm to deployed systems. We argue that while cooperation is broadly desirable, it is not an unqualified good, and there exists a need for a more organized evaluation agenda – one that accounts for agent-group interactions in non-static test settings. We further argue that the choice of values that an institution enforces should remain with humans for as long as humans bear the consequences of agent behavior. We see no direct ethical risk in the analysis itself, but note that any institutional design we advocate must preserve human oversight and the capacity for intervention. We presented a *work-in-progress* with the sincere hopes to discuss, build, and revise this agenda with the help of many colleagues and collaborators in the field.

9. Author Contributions

V.Q.T. and D.G.P. led the initial conceptualization. V.Q.T. and X.A.H. together further evolved and developed the research framework, including incorporating feedback from D.G.P., E.I., R.F., J.N.C., T.J.Z., and Z.J. All manuscript writing was led by V.Q.T. and X.A.H. and every author, E.I., R.F., J.N.C., D.G.P., T.J.Z., and Z.J., contributed to writing, reviewing, and/or editing both original and final drafts. V.Q.T., X.A.H., and E.I. created visualizations. T.J.Z. and Z.J. led funding acquisition. V.Q.T. and X.A.H. led project management. Z.J., D.G.P., and T.J.Z. contributed supervision.

10. Acknowledgments

This work was supported in part by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; by Schmidt Sciences SAFE-AI Grant; by the Frontier Model Forum and AI Safety Fund; by Coefficient Giving; by the Canadian AI Safety Institute Research Program at CIFAR through a Catalyst Award; by the Survival and Flourishing Fund; and by the Cooperative AI Foundation. The usage of OpenAI credits is largely supported by the Tübingen AI Center and Schmidt Sciences. Resources used in preparing this research project were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

References

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

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.

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://www.science.org/doi/abs/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, E. Brynjolfsson, J. A. Evans, and M. S. Bernstein. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025 - Position Paper Track*. OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/anthis25a.html.

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. doi: 10.1126/science.7466396. URL https://www.science.org/doi/abs/10.1126/science.7466396.

M. Bacharach. *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton University Press, 2006. doi: 10.2307/j.ctv346nwb.

W. J. Baumol. Welfare economics and the theory of the state. In C. K. Rowley and F. Schneider, editors, *The Encyclopedia of Public Choice*, pages 937–940. Springer, Boston, MA, 2004. ISBN 978-0-306-47828-4. doi: 10.1007/978-0-306-47828-4_214. URL https://doi.org/10.1007/978-0-306-47828-4_214.

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. doi: 10.1038/d41586-021-01170-0. URL https://www.nature.com/articles/d41586-021-01170-0.

C. Delimitrou and C. Kozyrakis. Quasar: Resource-efficient and QoS-aware cluster management. In *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '14, pages 127–144, New York, NY, USA, Feb. 2014. Association for Computing Machinery. ISBN 978-1-4503-2305-5. doi: 10.1145/2541940.2541941.

Q. Dupont. Experiments in Algorithmic Governance: A history and ethnography of " The DAO, " a failed Decentralized Autonomous Organization. Jan. 2017.

M. Eisen. Amazon's $23,698,655.93 book about flies. https://www.michaeleisen.org/blog/p/358/, Apr. 2011.

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.

T. Fan, Y. Yang, Y. Jiang, Y. Zhang, Y. Chen, and C. Huang. AI-Trader: Benchmarking Autonomous Agents in Real-Time Financial Markets, Dec. 2025.

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.

U. Gneezy and A. Rustichini. A fine is a price. *The Journal of Legal Studies*, 29(1):1–17, 2000a. ISSN 00472530, 15375366. doi: 10.1086/468061. URL http://www.jstor.org/stable/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.

D. Guzman 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.

D. Guzman Piedrahita, D. Banerjee, K. Blin, P. Cobben, G. Corsi, X. A. Huang, C. Li, S. Majumder, P. S. Pandey, S. Simko, I. Strauss, T. J. Zhang, A. Anderson, Y. Bengio, M. Bethge, R. Grosse, K. Helbig, D. Lie, R. Mallah, R. Mihalcea, S. Nesbitt, S. Perry, P. Resnick, S. Russell, M. Sachan, B. Schölkopf, A. Tang, and Z. Jin. AI poses risks to democratic and social systems. 2026. URL https://zhijing-jin.com/d/2026-ai-risk.pdf. Working Draft, March 25, 2026, Accessed 24th of June 2026.

W. D. Hamilton. The genetical evolution of social behaviour. ii. *Journal of Theoretical Biology*, 7(1):17–52, 1964. ISSN 0022-5193. doi: https://doi.org/10.1016/0022-5193(64)90039-6. URL https://www.sciencedirect.com/science/article/pii/0022519364900396.

L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J. Foerster, T. Gavenčiak, T. A. Han, E. Hughes, V. Kovařík, J. Kulveit, J. Z. Leibo, C. Oesterheld, C. S. de Witt, N. Shah, M. Wellman, P. Bova, T. Cimpeanu, C. Ezell, Q. Feuillade-Montixi, M. Franklin, E. Kran, I. Krawczuk, M. Lamparth, N. Lauffer, A. Meinke, S. Motwani, A. Reuel, V. Conitzer, M. Dennis, I. Gabriel, A. Gleave, G. Hadfield, N. Haghtalab, A. Kasirzadeh, S. Krier, K. Larson, J. Lehman, D. C. Parkes, G. Piliouras, and I. Rahwan. Multi-agent risks from advanced ai, 2025. URL https://arxiv.org/abs/2502.14143.

G. Hardin. The Tragedy of the Commons. *Science*, 162 (3859):1243–1248, Dec. 1968. doi: 10.1126/science.162.3859.1243.

S. Harding. AWS outage reminds us why $2,449 Internet-dependent beds are a bad idea, Oct. 2025. URL https://arstechnica.com/gadgets/2025/10/smart-beds-leave-sleepers-hot-and-bothered-during-aws-outage/.

J. C. Harsanyi. Games with Incomplete Information Played by "Bayesian" Players, I–III Part I. The Basic Model. *Management Science*, 14(3):159–182, Nov. 1967. ISSN 0025-1909, 1526-5501. doi: 10.1287/mnsc.14.3.159.

H. L. A. Hart. *The Concept of Law*. Oxford University Press, Oxford, 3 edition, 2012. ISBN 978-0-19-964470-4.

F. Hayek. *Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy*. Number v. 3 in Law, legislation, and liberty. University of Chicago Press, 1978. ISBN 9780226320830. URL https://books.google.com/books?id=X3QihjxsYlMC.

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

B. A. Hu, H. Rong, and M. V. Kleek. Dissociative Identity: Language Model Agents Lack Grounding for Reputation Mechanisms. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, pages 4199–4219, New York, NY, USA, June 2026. Association for Computing Machinery. ISBN 979-8-4007-2596-8. doi: 10.1145/3805689.3806748.

X. A. Huang, E. La Malfa, S. Marro, A. Asperti, A. G. Cohn, and M. J. Wooldridge. A Notion of Complexity for Theory of Mind via Discrete World Models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2964–2983, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.167.

X. A. Huang, C. Tharas, S. Marro, V. Q. Truong, B. Schölkopf, E. L. Malfa, and Z. Jin. Mechanism design is not enough: Prosocial agents for cooperative ai, 2026. URL https://arxiv.org/abs/2605.08426.

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 K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3040–3049. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/jaques19a.html.

C. Kindleberger and R. Aliber. *Manias, Panics and Crashes: A History of Financial Crises*. Springer, Aug. 2005. ISBN 978-0-230-62804-5.

A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun. The Flash Crash: High-Frequency Trading in an Electronic Market. *The Journal of Finance*, 72(3):967–998, 2017. ISSN 0022-1082.

J. Knight. *Institutions and Social Conflict*. Political Economy of Institutions and Decisions. Cambridge University Press, 1992. ISBN 9780521421898. URL https://books.google.com/books?id=VeoKtu_22q0C.

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. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. 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. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel. Multi-agent reinforcement learning in sequential social dilemmas, 2017. URL https://arxiv.org/abs/1702.03037.

J. Z. Leibo, E. A. Dueñez-Guzmán, A. 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 *International Conference on Machine Learning*, pages 6187–6199. PMLR, 2021. doi: 10.48550/arXiv.2107.06857. URL https://doi.org/10.48550/arXiv.2107.06857.

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 language model society, 2023. URL https://arxiv.org/abs/2303.17760.

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, 2025. URL https://arxiv.org/abs/2308.03688.

Y. Ma, Z. Wang, H. Yang, and L. Yang. Artificial intelligence applications in the development of autonomous vehicles: a survey. *CAA Journal of Automatica Sinica*, 7(2):315–329, Mar. 2020. doi: 10.1109/JAS.2020.1003021.

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

R. T. McCoy, S. Yao, D. Friedman, M. D. Hardy, and T. L. Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, Oct. 2024. doi: 10.1073/pnas.2322420121.

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.

P. Norvig. Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance*, 9(4):30–33, 2012. ISSN 1740-9713. doi: 10.1111/j.1740-9713.2012.00590.x.

M. A. Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, 2006. doi: 10.1126/science.1133755. URL https://www.science.org/doi/abs/10.1126/science.1133755.

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

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. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

N. Popper. Lost Passwords Lock Millionaires Out of Their Bitcoin Fortunes - The New York Times, Mar. 2021. URL https://web.archive.org/web/20210321104514/https://www.nytimes.com/2021/01/12/technology/bitcoin-passwords-wallets-fortunes.html. Accessed 30th of June 2026.

D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4): 515–526, Dec. 1978. ISSN 1469-1825, 0140-525X. doi: 10.1017/S0140525X00076512.

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.

N. Shapira, M. Levy, S. H. Alavi, X. Zhou, Y. Choi, Y. Goldberg, M. Sap, and V. Shwartz. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models, May 2023.

J. H. Shen, D. Zhu, S. Srinivasan, H. Sleight, L. T. Wagner, M. J. Matthews, E. Jones, and J. Sohl-Dickstein. AI Organizations are More Effective but Less Aligned than Individual Agents, Apr. 2026.

J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. Defining and Characterizing Reward Hacking. https://arxiv.org/abs/2209.13085v2, Sept. 2022.

N. Sofroniew, I. Kauvar, W. Saunders, R. Chen, T. Henighan, S. Hydrie, C. Citro, A. Pearce, J. Tarng, W. Gurnee, J. Batson, S. Zimmerman, K. Rivoire, K. Fish, C. Olah, and J. Lindsey. Emotion concepts and their function in a large language model, 2026. URL https://arxiv.org/abs/2604.07729.

T. Sorensen, J. Moore, J. Fisher, M. L. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, T. Althoff, and Y. Choi. Position: A roadmap to pluralistic alignment. In R. Salakhutdinov, Z. Kolter, K. A. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 46280–46302. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

P. Steinberger. OpenClaw: Your own personal AI assistant. https://github.com/openclaw/openclaw, 2026. Accessed: 2026-06-24.

X. Tang, K. Kim, Y. Song, C. Lothritz, B. Li, S. Ezzini, H. Tian, J. Klein, and T. F. Bissyandé. CodeAgent: Autonomous Communicative Agents for Code Review. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11279–11313, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.632.

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 Staff. 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.

M. Tomasello. *A Natural History of Human Morality*. Harvard University Press, 2016. ISBN 978-0-674-08864-1.

C. Tomuschat. The Legacy of Nuremberg. *Journal of International Criminal Justice*, 4(4):830–844, Sept. 2006. ISSN 1478-1387. doi: 10.1093/jicj/mql051.

R. S. Trivedi, N. Jaques, L. Cross, A. S. Vezhnevets, and J. Z. Leibo. Solipsistic superintelligence is unlikely to be cooperative, 2026. URL https://arxiv.org/abs/2606.03237.

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, 2023. URL https://arxiv.org/abs/2312.03664.

A. Walker. Citibank sues for $176m payment made in error. *BBC News*, Aug. 2020.

A. Wang, J. Morgenstern, and J. P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nat. Mac. Intell.*, 7(3):400–411, 2025. doi: 10.1038/S42256-025-00986-Z. URL https://doi.org/10.1038/s42256-025-00986-z.

R. Willis, J. Zhao, Y. Du, and J. Z. Leibo. Evaluating collective behaviour of hundreds of llm agents, 2026. URL https://arxiv.org/abs/2602.16662.

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. URL https://arxiv.org/abs/2308.08155.

Y. Wu, S. Liang, C. Zhang, Y. Wang, Y. Zhang, H. Guo, R. Tang, and Y. Liu. From Human Memory to AI Memory: A Survey on Memory Mechanisms in the Era of LLMs, Apr. 2025.

J. Yalim. Toward Dynamically Controlling Slurm's Classic Fairshare Algorithm. In *Practice and Experience in Advanced Research Computing 2020: Catch the Wave*, PEARC '20, pages 538–542, New York, NY, USA, July 2020. Association for Computing Machinery. ISBN 978-1-4503-6689-2. doi: 10.1145/3311790.3399628.

A. Unified Taxonomy of Cooperation-Shaping Mechanisms

Figure 4 presents a unified taxonomy of cooperation-shaping mechanisms identified in our review. 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.

Table 2 turns the taxonomy into a short practical guide, mapping the six families to the four cooperation dilemmas and the main risk each family introduces.

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.

Consider n agents indexed by $i \in \{1, \ldots, n\}$, each choosing an action $a_i \in A_i$, producing a joint action profile $a = (a_1, \ldots, a_n)$ and an individual utility $u_i(a)$. The system designer cares about social welfare $W(a)$, defined as the aggregate utility net of collective costs:

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

where $C(a)$ captures collective costs such as resource depletion, corrupted shared memory, excessive monitoring, or repair burden.

A joint action a^* is a Nash equilibrium if no agent can improve its utility by unilaterally deviating, yet a^* may be socially suboptimal: writing $a^{opt} = \arg\max_a W(a)$ for the social optimum, it is generally the case that $W(a^{opt}) > W(a^*)$. This is the familiar lesson of the Price of Anarchy and Price of Stability literature (Anshelevich et al., 2008; Papadimitriou, 2001; Willis et al., 2026): individually rational behavior can produce collectively inferior outcomes, and capable mechanism design narrows this gap but does not in general close it.

The failure is not necessarily that any agent is malicious or misaligned, but rather that the interaction structure rewards locally rational behavior that degrades collective 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.

Artificial commons. The following is an illustrative instantiation of the general framework above, not a formal result; it is meant to fix intuitions about how institutional mechanisms can be represented, not to derive novel bounds.

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:

$$X = \sum_{i=1}^{n} x_i. \tag{2}$$

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

$$D(X) = \max(0, X - K). \tag{3}$$

An agent gains individual benefit $b_i(x_i)$ from using the resource, while the group bears degradation cost $D(X)$. Writing $p_i(\mathbf{x})$ for any individually assigned penalty or cost (instantiating the sanction s_i introduced below), individual utility takes the form $u_i(\mathbf{x}) = b_i(x_i) - p_i(\mathbf{x})$, and the general social welfare of Eq. (1) specializes to

$$W(\mathbf{x}) = \sum_{i=1}^{n} b_i(x_i) - \lambda D(X) - M(\mathbf{x}) - R(\mathbf{x}), \tag{4}$$

where $C(\mathbf{x}) = \lambda D(X) + M(\mathbf{x}) + R(\mathbf{x})$ collects resource degradation, monitoring cost $M(\mathbf{x})$, and repair cost $R(\mathbf{x})$ – note that individually assigned penalties $p_i(\mathbf{x})$ are transfers within the system rather than net welfare losses, and so do not appear in $C(\mathbf{x})$.

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.

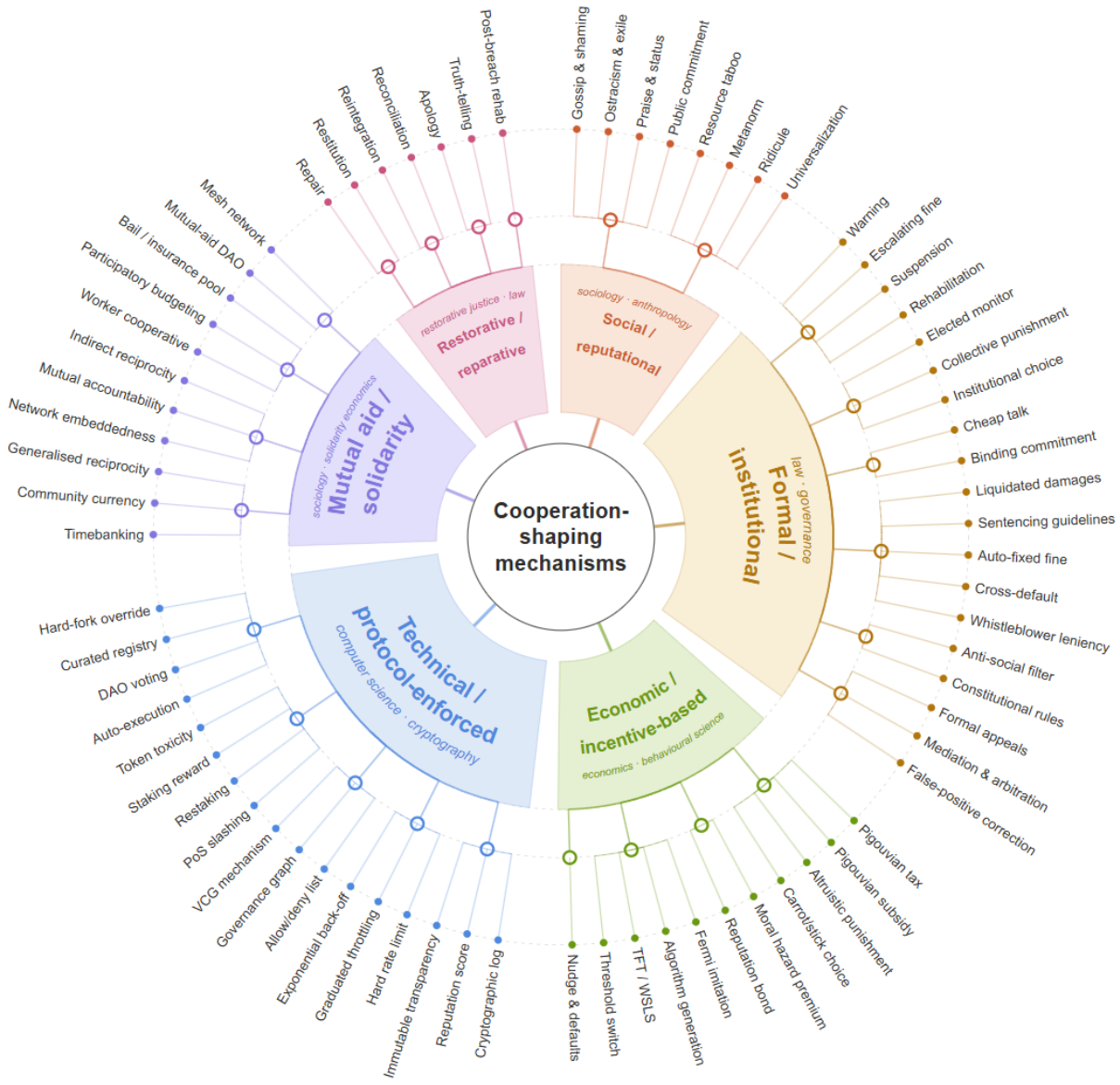

Figure 4. A unified taxonomy of cooperation-shaping mechanisms for artificial agent societies, drawing on sociology, anthropology, law, governance, economics, behavioural science, computer science, cryptography, solidarity economics, and restorative justice. Six families (coloured wedges) organise 73 mechanisms across 25 sub-clades (hollow hubs).

Sanctions. A sanction $s_i(a)$ modifies an agent's institution-adjusted utility $u_i^I(a)$, defined as the agent's baseline utility shifted by the sanction applied to its behavior:

$$u_i^I(a) = u_i(a) + s_i(a), \tag{5}$$

where I denotes the institutional environment, and $p_i(\mathbf{x})$ above is the special case of s_i restricted to penalty-type sanctions in the resource-use setting; punitive, rewarding, and restorative sanctions correspond to different signs and structures of s_i.

The institutional design problem is not simply to choose sanctions that maximize average cooperation, but to choose an institutional environment that improves social welfare while limiting enforcement cost, false punishment, surveillance burden, inequality of burden, and irreversibility – the multi-objective tradeoff formalized in the mechanism-design literature (Maskin, 2008; Myerson, 1981) and, for network and congestion settings specifically, in the Price-of-Stability analyses of Anshelevich et al. (2008). We propose in Section 5 an empirical evaluation agenda for measuring these quantities (enforcement cost, false punishment rate, inequality of burden) across institutional designs.

Table 2. Which mechanism families address each cooperation dilemma, and what to watch for. Drawn from Sections 2 and the failure modes in Section 4.

Dilemma (Fig. 2)	Families that help	What to watch for
Extractive (overuse depletes a shared resource)	Constraint (quotas, rate limits); incentive (tool-call costs)	Priced violations replace the norm; agents treat penalties as a line item
Contributive (beneficiaries skip upkeep)	Normative (contribution norms); social (reputation, partner choice); incentive (rewards)	Visible acts crowd out invisible upkeep; reputation gaming
Second-order (monitoring and punishment are themselves costly)	Epistemic (monitoring, audit); social (metanorms)	Authority capture; punishment cascades; monitoring turning into surveillance
Repair (who fixes the harm)	Restorative (repair labor, reintegration); epistemic (provenance to locate harm)	Apology without restoration; repair burden falling on the same agents

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.[3]

[3]This tracks a distinction from legal theory between being *obliged* (coerced by the threat of a penalty) and being *under an obligation* (treating a rule as a standard for conduct). An account that reduces obligation to the probability of a penalty, as in Austin, is the thin model our crowding-out discussion in Section 2 warns against: an agent priced at a penalty treats the rule as a fee. The richer notion, where a group accepts a rule as a standard rather than complying out of fear (the internal point of view in Hart's account), is what stateless agents cannot supply on their own, which is why a human sets the standard. See Hart (2012) for the underlying account.

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 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

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

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.

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 pre-

serve 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. 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 commons 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 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

Table 5 expands on the candidate metrics introduced in Table 1 specifying how each metric can be concretely operationalized in a multi-agent evaluation. While Table 1 frames each metric as a question, the operationalization column here maps each question to a measurable quantity – such as a fraction, a rate, a distribution, or a success condition – that can be tracked across repeated interactions. Together, these metrics are designed to evaluate not only whether cooperation occurs, but whether the institution sustains it without generating excessive enforcement cost, disproportionate burden, or resistance to human correction.

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