Data Sharing with Large Language Models: Prisoner's Dilemma, Pareto Inefficiency, and Institutional Design

Ladislav Stanke[1]

[1]Department of Psychology, Faculty of Arts, Palacký University Olomouc, Křížkovského 10, 771 80 Olomouc, Czech Republic

Abstract

This article analyzes organizational decision-making about data sharing in the context of large language models (LLMs) through the combined lenses of game theory, Pareto efficiency, and institutional design. It argues that the central problem is not the absence of Nash equilibrium, but the existence of a stable yet Pareto-inferior equilibrium in which widespread non-sharing remains individually rational while generating collective epistemic and innovative loss. To make this claim more precise, the article develops a formal model of the data-sharing dilemma using parameterized payoffs for private benefit, collective innovation gain, sharing cost, appropriative advantage, systemic loss from mutual withholding, and the effects of technical and institutional safeguards. On that basis, it shows how legal, technical, and organizational interventions may reparameterize the game so that limited, auditable, and protected cooperation becomes strategically attractive. The article further examines the paradoxes generated by both sharing and non-sharing paradigms, arguing that neither openness nor restriction can serve as a sufficient normative solution in isolation. A proof-of-concept evolutionary population simulation further illustrates that stronger governance regimes can increase cooperation, improve aggregate welfare, and reduce payoff dispersion, while also showing that cost-reducing safeguards may exert a stronger effect than trust alone. The article concludes by proposing a staged trust architecture for institutional cooperation and by outlining broader comparative and experimental paths for empirical validation. The broader claim is that the future of organizational data sharing around LLMs will depend less on moral appeals to openness than on the creation of governance structures that make cooperation rational, safe, and legitimate (Carlini et al., 2021; Curtis et al., 2023; Kaushik et al., 2025; Pasetti et al., 2025; Reinhardt, 2023; Schultz and Seele, 2023; Cheng et al., 2020; Yin et al., 2021).

1 Introduction

The question of whether organizations should share data with large language models cannot be reduced to a merely technical matter of information management or to a straightforward question of productivity enhancement. In contemporary digital environments, it has become a structural problem of coordination under uncertainty, one that forces firms, public institutions, and research organizations to navigate a tension between preserving control and pursuing collective gains from accelerated inference, broader learning, and improved problem-solving. At stake is not only the efficiency of data use, but also the institutional legitimacy of the conditions under which data are transferred, processed, and transformed into model-based intelligence. Once data become a source of model training, retrieval, prediction, or inferential augmentation, they cease to function as purely internal organizational resources. They enter a wider socio-technical field in which one must ask who controls the data, who benefits from them, who bears the associated risks, and according to what legal and ethical criteria such arrangements can be justified. In that sense, the problem of data sharing for LLMs belongs simultaneously to economics, law, organizational ethics, and political philosophy (Curtis et al., 2023; Pasetti et al., 2025; Schultz and Seele, 2023).

This article argues that the central difficulty should be understood not as a failure of technical capacity, nor as a mere psychological deficit of trust, but as a game-theoretic and institutional problem. Individual organizations may rationally prefer non-sharing because doing so reduces immediate exposure to privacy risk, competitive leakage, regulatory sanction, or reputational harm. Yet if most or all actors adopt the same strategy, the resulting environment becomes epistemically fragmented and collectively less efficient. Innovation slows, duplication of effort rises, and the overall system loses part of the potential benefit that could arise from more coordinated forms of learning and analysis. The relevant difficulty is therefore not the absence of equilibrium, but rather the possibility that a strategically stable equilibrium may nevertheless be socially and institutionally inferior. This is why the Prisoner's Dilemma remains a useful starting point: it illuminates the conflict between individually rational restraint and collectively superior cooperation (*Prisoner's Dilemma and Cooperation* 2021; Richards, 2001).

At the same time, the familiar game-theoretic intuition requires refinement in the context of LLMs. The contemporary data-sharing dilemma has at least two analytically distinct layers. At the first level, organizations decide whether and how to share data into AI-mediated infrastructures. At the second level, LLMs themselves may increasingly function not only as passive infrastructures trained on shared or withheld data, but also as advisory, mediating, or delegated agents in organizational decision-making. This distinction matters because the strategic problem is no longer exhausted by asking whether institutions will share data with AI systems. One must also ask whether AI

systems, when used to advise on data governance, risk, reciprocity, or negotiation, alter the perceived payoff structure of the underlying game. Put differently, the rise of LLMs may transform both the object and the medium of strategic coordination. Where organizations rely on LLM-assisted decision support, the behavioral properties of such systems in social-dilemma settings may become relevant to institutional design itself.

The present article therefore makes a more specific contribution than merely redescribing data sharing as a Prisoner's Dilemma. Its central claim is that the problem should be understood as a layered institutional dilemma in which a Pareto-inferior equilibrium persists because the cost structure of cooperation has not yet been sufficiently transformed by law, governance, and technical architecture. On this account, the key task is neither to celebrate openness nor to normalize withholding, but to explain why generalized non-sharing may remain rational under present conditions and how institutional interventions might reparameterize the game so that limited, auditable, and safe cooperation becomes strategically attractive. The analysis proceeds from the assumption that rationality, efficiency, and trust must be analytically distinguished if the debate is to avoid conceptual confusion. Rationality refers here to payoff-oriented behavior under perceived constraints, not to moral correctness. Efficiency refers to Pareto efficiency, not to distributive justice. Trust, finally, is treated not as mere subjective optimism, but as justified institutional confidence grounded in predictability, accountability, auditability, and the possibility of sanction (Reinhardt, 2023).

This contribution is therefore best understood as conceptual, formal, and institutionally oriented rather than as a new mechanism-design result in the narrow technical sense. The aim of the article is not to replace the growing literature on incentive-compatible federated learning, model-sharing markets, or contribution valuation schemes, but to provide a broader framework for understanding the legal, organizational, and trust conditions under which such mechanisms become strategically credible and normatively acceptable. Its novelty lies in the institutional reparameterization perspective, the layered treatment of LLM-mediated decision environments, and the integration of formal modeling with a proof-of-concept population simulation.

The paper develops this argument in five steps. First, it formalizes the organizational data-sharing dilemma as a simple game with parameterized payoffs, making explicit the conditions under which non-sharing becomes a dominant strategy and mutual restraint becomes a Pareto-inferior equilibrium. Second, it positions this model in relation to adjacent work on collaborative machine learning, federated learning, and institutional AI governance. Third, it uses the model to clarify how legal and technical interventions may alter the payoff structure rather than merely exhort actors to cooperate. Fourth, it revisits the paradoxes of both sharing and non-sharing, showing that each paradigm contains internal tensions that prevent it from serving as a complete normative solu-

tion. Fifth, it presents a proof-of-concept evolutionary population simulation and uses it to sketch a path toward institutional design grounded in staged trust architectures, repeated interaction, and verifiable governance. The broader aim is to show that the future of organizational data sharing with LLMs will not be determined simply by technical capacity or moral attitude, but by the quality of the institutional arrangements that lower the price of cooperation while preserving legitimacy.

2 Related Work and Contribution Positioning

The strategic analysis of data sharing in machine learning is not new, and this article does not claim novelty for the basic intuition that individually rational actors may fail to coordinate on collectively superior outcomes. In recent years, however, several adjacent strands of literature have developed in ways that make a sharper positioning necessary. One strand concerns game-theoretic models of federated learning, distributed model training, and collaborative optimization, where the central question is how to sustain contribution under conditions of partial decentralization, privacy constraint, or resource asymmetry. Another concerns model-sharing markets, synthetic data exchange, and contribution valuation mechanisms, where the focus lies on incentive compatibility, pricing, and equitable participation within technically specified collaborative systems. A further strand concerns the growing interface between game theory and large language models themselves, including work on how LLMs behave in strategic environments and how they may mediate decision-making under social-dilemma conditions.

The present article is related to these lines of work, but its contribution is located at a different level of analysis. It does not primarily seek to design a contribution-allocation mechanism inside a fixed collaborative learning architecture, nor does it attempt to optimize a specific exchange protocol for data or model parameters. Instead, it addresses a broader institutional problem: the conditions under which organizations regard participation in such architectures as strategically credible, legally defensible, and normatively legitimate in the first place. In that sense, the article complements rather than replaces technical mechanism-design work. Its argument is that even elegant incentive mechanisms may fail if the surrounding institutional environment leaves the cost of participation high, the risk of unilateral appropriation substantial, and the conditions of trust insufficiently structured. The central contribution of the paper is therefore an institutional reparameterization framework: it shows how legal, technical, and organizational interventions alter the strategic environment within which more specialized data-sharing mechanisms operate.

This distinction is especially important in relation to federated learning and related privacy-preserving paradigms. Much of that literature asks how collaboration can be sustained once actors are already situated inside a technically defined cooperative frame-

work. The present article asks a prior question: under what conditions do organizations rationally choose to enter or refuse such frameworks at all? The answer, on the present account, cannot be derived from technical efficiency alone. It depends on the interaction of sharing cost, appropriative asymmetry, expected collective gain, trust architecture, and normative legitimacy. This broader framing is what allows the article to bring game theory into direct dialogue with institutional trust, legal governance, and the paradoxes of both openness and closure.

The same point applies to the paper's treatment of LLMs. The argument is not merely that data sharing may improve the performance of large language models, nor simply that game-theoretic reasoning may be applied to LLM-centered ecosystems. Rather, the article proposes that the governance problem has become layered. At one level, organizations choose whether to share data into infrastructures that support LLMs and other AI systems. At another level, LLMs may increasingly participate in the epistemic mediation of these choices by functioning as advisory, screening, or compliance-support systems. This layered structure differentiates the present contribution from work that treats LLMs only as passive technical endpoints or, conversely, only as autonomous strategic agents. The article instead locates them within a broader institutional ecology in which technical architectures, organizational incentives, and normative constraints interact.

3 A Formal Model of the Data-Sharing Dilemma

To make the central argument more precise, this section introduces a deliberately simple model of organizational data sharing in the context of LLM-mediated infrastructures. The purpose of the model is not to capture every empirical detail of real-world governance arrangements, but to formalize the strategic intuition underlying the paper's normative claims. Consider two organizations, indexed by $i \in \{1, 2\}$, each of which chooses between two strategies: S, meaning to share data under a given institutional regime, and N, meaning not to share. The term "sharing" is used here in a broad but bounded sense. It may refer to direct contribution of data to external or joint model environments, participation in controlled consortium settings, or technically mediated forms of contribution such as federated or privacy-preserving analytics. Likewise, "non-sharing" refers to a strategy of withholding participation from such collaborative arrangements, even if the organization continues to use internal AI tools on its own data.

The payoff of each organization is assumed to depend on five analytically distinct variables. Let B denote the baseline private benefit from using AI internally, even in the absence of interorganizational sharing. Let G denote the collective innovation gain generated when both organizations share under a regime that enables complementary learning, broader case diversity, or reduced duplication of epistemic effort. Let C denote the cost of sharing, including privacy risk, compliance burden, exposure of know-how, governance

complexity, and reputational vulnerability. Let A denote the appropriation advantage available to a non-sharing actor when the other actor shares, that is, the asymmetric gain from benefiting indirectly from a more open ecosystem without incurring the full cost of openness. Finally, let δ denote the opportunity loss borne by all actors when both withhold and the system remains epistemically fragmented. In addition to these baseline terms, two further modifiers become relevant once institutional design is introduced. Let R denote the reduction of effective sharing cost produced by technical and legal safeguards, such as auditable access controls, purpose limitation, federated learning, or secure computation. Let T denote the trust premium associated with repeated interaction, verified reciprocity, or credible sanctioning arrangements.

Under these assumptions, the payoff functions can be written in the following stylized form:

$$\pi_i(S, S) = B + G - C + R + T$$

$$\pi_i(S, N) = B - C$$

$$\pi_i(N, S) = B + A$$

$$\pi_i(N, N) = B - \delta$$

The interpretation is straightforward. If both organizations share, each receives the baseline benefit from AI, the collective gain from mutual contribution, and any benefits generated by institutional safeguards or trust-enhancing arrangements, but each also bears the cost of sharing. If one organization shares while the other withholds, the sharing actor pays the relevant cost without obtaining the full cooperative gain, whereas the non-sharing actor obtains the baseline benefit and may enjoy an appropriation advantage. If neither shares, each preserves control and avoids the immediate cost of openness, but both incur a systemic loss due to foregone collective learning.

The standard Prisoner's Dilemma structure arises when the following inequalities hold:

$$\pi_i(N, S) > \pi_i(S, S) > \pi_i(N, N) > \pi_i(S, N)$$

Under these conditions, non-sharing is the dominant strategy for each player, because withholding yields a higher payoff regardless of the other player's choice. Yet the outcome (N, N) remains Pareto-inferior to (S, S), since both actors would be better off if they could credibly coordinate on mutual sharing. This formalizes the intuitive claim running through the paper: organizational non-sharing may be strategically rational and yet collectively inefficient. Nash equilibrium and Pareto optimality thus come apart. The equilibrium of mutual restraint is stable, but it is not socially superior.

The model also helps specify what institutional design must do if it is to alter the strategic

structure rather than merely advocate cooperation at a moral level. A cooperative regime becomes more attractive when the effective cost of sharing, C, is lowered; when the expected innovation gain, G, is increased; when the appropriative advantage of unilateral defection, A, is reduced; and when the trust and safeguard modifiers, R and T, increase the expected value of mutual participation. Technical mechanisms such as federated learning, secure multi-party computation, contractual restrictions on secondary use, audit logs, and role-based access controls can be interpreted in this framework as devices that reduce C or increase R. Repeated interaction, verified compliance, and sanction-backed governance structures can be interpreted as increasing T and decreasing the attractiveness of unilateral appropriation. The model therefore gives a formal meaning to the broader thesis of the article: institutions do not solve the dilemma by abolishing self-interest, but by changing the payoff structure under which self-interest is exercised (Cheng et al., 2020; Yin et al., 2021; Schultz and Seele, 2023; Curtis et al., 2023).

This formalization also clarifies why the binary opposition between sharing and non-sharing is insufficient. In real environments, organizations rarely choose between absolute openness and absolute closure. Instead, they move across a spectrum of possible sharing regimes, each of which corresponds to a different configuration of C, G, A, R, and T. Raw data transfer may involve high expected cost and high collective gain; federated parameter exchange may lower cost while preserving part of the gain; synthetic data sharing may further reduce risk while also reducing informational value. This suggests that the strategic problem is better understood not as a single one-shot choice, but as a family of nested games in which technical architecture and legal design determine which region of the payoff space is actually available. The model therefore serves not only as a formal justification of the Prisoner's Dilemma intuition, but also as a bridge to the later claim that staged trust architectures may make cooperation sustainable without requiring naive openness.

Finally, the model provides a basis for integrating the role of LLMs as advisory or delegated agents. If organizations increasingly rely on LLM-based systems to assess compliance burdens, estimate reciprocity, predict partner behavior, or recommend data-sharing strategies, then such systems may influence the perceived magnitude of C, A, G, and T. In that case, the strategic role of LLMs becomes double. They are not only infrastructures whose performance depends on data-sharing choices; they may also become actors in the epistemic mediation of those choices. This possibility does not require attributing full agency to LLMs in a metaphysical sense. It is enough to observe that institutions may come to act through model-assisted representations of payoffs, trustworthiness, and risk. Once that occurs, the literature on LLM behavior in social dilemmas becomes relevant not as a curiosity, but as a potentially important input into institutional design.

3.1 Stylized Extensions of the Model

The baseline model is intentionally simple. Its purpose is to identify the strategic core of the dilemma rather than to exhaust the full complexity of real-world institutional ecosystems. Even so, the main argument of the article implies two natural extensions, both of which are useful for interpreting the later simulation and the role of institutional trust.

The first extension is temporal. In repeated interaction, the relevant objective is no longer a single-stage payoff but an intertemporal one:

$$\Pi_i = \sum_{t=0}^{\infty} \beta^t \pi_i^t,$$

where $\beta \in (0, 1)$ is a discount factor. This extension helps formalize the intuition behind the trust parameter T. In a one-shot setting, unilateral withholding may remain attractive because no future sanction or reciprocal reward can operate. In repeated settings, however, present conduct shapes future opportunities for access, reputation, and collaboration. The trust premium can therefore be interpreted more formally as the institutionalized expected value of future cooperation under conditions of observability, sanctionability, and repeated interaction. This does not eliminate the dilemma, but it changes the strategic meaning of present choices.

The second extension is population-based. Real data-sharing ecosystems rarely consist of two perfectly symmetric organizations. They are better understood as heterogeneous populations of actors indexed by distinct values of B_i, C_i, G_i, A_i, R_i, T_i, and δ_i. Once heterogeneity is introduced, the analytical question shifts from the existence of a single equilibrium between two players to the long-run ecology of strategy types in a broader institutional environment. This is the move taken up later in the paper through the evolutionary population simulation. The simulation does not replace the baseline model; it operationalizes its extension into a setting where strategies coexist, adapt, and persist under different governance regimes.

A third extension concerns the article's layered account of LLMs. If organizations increasingly rely on LLM-assisted systems to advise on risk, compliance, reciprocity, or partner trustworthiness, then the immediate strategic variables may be transformed at the level of perception before they are transformed at the level of objective structure. This can be represented in stylized form as:

$$(C', A', G', T') = f_{\text{LLM}}(C, A, G, T),$$

where f_{LLM} denotes an advisory transformation of perceived payoffs. On this interpretation, the LLM is not modeled as a sovereign strategic player in the strong sense. Rather,

it functions as an epistemic mediator that may amplify or attenuate the salience of cost, appropriation risk, expected gain, or trust. A compliance-oriented advisory system, for example, might raise the perceived weight of C; a reciprocity-sensitive system might increase the effective salience of T; a risk-minimizing advisory system might systematically bias actors toward withholding. The point of this extension is not to resolve the matter fully within the present paper, but to show that the article's claim about a "second layer" of the dilemma can be given formal expression rather than remaining purely metaphorical.

The rest of the article builds on this model in two ways. First, it uses the distinction between strategic stability and Pareto efficiency to explain why both the sharing and non-sharing paradigms generate internal paradoxes. Second, it shows that legal, technical, and organizational interventions should be evaluated according to their ability to transform the parameters of the game in ways that preserve legitimacy while making cooperation rationally sustainable.

4 From Strategic Stability to Pareto Inefficiency

The formal model presented in the previous section provides a compact account of why organizational data sharing around LLMs may remain strategically blocked even where the collective gains from cooperation are substantial. The present section develops the conceptual implications of that model. Its central purpose is to clarify why strategic stability and institutional desirability do not coincide, and why the distinction between Nash equilibrium and Pareto optimality is indispensable for understanding the present dilemma. In ordinary organizational discourse, widespread non-sharing is often interpreted either as prudent realism or as a regrettable failure of collaborative imagination. The model suggests a more precise diagnosis. Mutual restraint may be rational in the sense that no actor has sufficient incentive to deviate unilaterally, while still remaining collectively inferior to a coordinated regime of protected cooperation. The significance of the Prisoner's Dilemma in this context is therefore not metaphorical alone. It expresses a structural misalignment between individual incentives and system-level learning.

This distinction matters because contemporary debate frequently conflates three analytically different ideas: rationality, efficiency, and legitimacy. Rationality concerns what an actor is justified in doing given the expected payoff structure. Efficiency concerns whether an alternative arrangement could improve the position of one or more actors without making others worse off. Legitimacy concerns whether a given arrangement is normatively acceptable once rights, duties, asymmetries of power, and the distribution of risk are taken into account. These three levels often diverge in data-governance settings. An organization may rationally choose not to share; a cooperative arrangement may nevertheless be more efficient in the Pareto sense; and yet even that more efficient arrangement may remain illegitimate if its benefits depend on the externalization of pri-

vacy, compliance, or reputational costs onto weaker parties. The point of introducing Pareto optimality is therefore not to provide a sufficient criterion of justice, but to show that strategically stable institutional arrangements may still be systematically suboptimal from the standpoint of collective problem-solving (*Prisoner's Dilemma and Cooperation* 2021; Richards, 2001; Reinhardt, 2023).

One implication of this distinction is that generalized non-sharing should not be dismissed too quickly as the product of irrational conservatism or technophobia. Where the expected cost of participation is high, where the possibility of unilateral appropriation remains significant, and where legal or technical safeguards are weak, non-sharing may be a fully intelligible response. The fact that this response is collectively costly does not make it individually incoherent. On the contrary, the force of the model lies in showing that such outcomes can be simultaneously rational at the micro level and undesirable at the macro level. This makes clear why a purely moralizing critique of institutional caution is insufficient. If actors are embedded in an environment where the downside risk of cooperation is immediate and concentrated while the gains of coordination are distributed, uncertain, or delayed, then the appeal to "share more" lacks strategic traction. It does not address the conditions that make non-sharing a dominant strategy in the first place.

A second implication is that Pareto-superior arrangements must be treated with normative caution. It may be tempting to conclude that once a more efficient regime is identified, the analytical problem is solved. Yet Pareto superiority by itself does not settle the question of whether the arrangement is acceptable. It may be possible to increase the welfare of organizations as institutional players while still exposing data subjects, employees, patients, customers, or weaker partners to disproportionate forms of risk. Efficiency, in other words, does not absorb the problem of distribution. This is especially important in the context of LLMs, where aggregate gains from learning, prediction, or automation may obscure the fact that the burdens of data exposure are often unevenly allocated. The distinction between Pareto improvement and normative legitimacy is therefore not an abstract philosophical luxury. It is essential if one wishes to prevent the language of collective gain from masking the externalization of harm (European Commission, 2026; European Parliament and Council of the European Union, 2016; Reinhardt, 2023).

A third implication concerns the role of institutional trust. The formal model already introduced T as a trust premium generated by repeated interaction, auditability, and credible reciprocity. Conceptually, however, this variable deserves clarification. Trust is often invoked in debates on AI governance as if it were either a subjective feeling or a generalized ethical aspiration. The model suggests a more disciplined interpretation. In this setting, trust is best understood as institutionally supported confidence that exposure to cooperation will not be converted into unilateral disadvantage. Such trust is neither

merely psychological nor merely technical. It is a relational condition grounded in predictable rules, observability of conduct, enforceable commitments, and the possibility of sanction. This helps explain why trust cannot simply be demanded. It must be built into the structure of interaction itself. Where actors cannot observe, verify, or contest how data are used, the call for trust becomes normatively hollow and strategically ineffective.

Once viewed in this way, the strategic problem of data sharing begins to look less like a binary opposition between generosity and defensiveness and more like a problem of institutional transformation. The central question is not whether organizations should care about collective learning, but under what conditions collective learning becomes compatible with self-protective rationality. In that respect, the formal model is not intended to reduce the problem to a set of equations. Rather, it provides a disciplined way of showing why a stable equilibrium may nonetheless be inferior, why efficiency cannot be identified with justice, and why the movement from one equilibrium to another requires institutional intervention rather than optimism alone. These insights prepare the ground for the next stages of the argument, where the paper turns first to the epistemic and innovative attractions of sharing, then to the ethical and legal limits of those attractions, and finally to the paradoxes that emerge when either sharing or non-sharing is treated as a complete solution.

5 The Epistemic and Innovative Logic of Sharing and the Limits of that Logic

One of the strongest arguments in favor of data sharing is the fact that the value of large language models, and more generally of data-intensive AI systems, increases with the diversity, quality, and volume of the cases on which they are trained, validated, or inferentially deployed. Here data are not merely raw material in an economic sense, but carriers of structures, exceptions, latent correlations, and contextual relations. An organization that keeps its knowledge and case bases entirely separate does preserve control over its local informational field, but at the same time deprives itself of the advantages of broader learning. In sectors where the quality of solutions depends heavily on the heterogeneity of cases, such as healthcare, security, research, industrial optimization, or complex legal and administrative decision-making, this loss may be substantial. Kaushik et al. (2025) show that data barriers in healthcare AI significantly reduce the ability to exploit the potential of new tools, while Wornow et al. (2023) warn that foundation models in electronic health records rest on fragile data foundations that cannot be repaired merely by greater model sophistication. In a broader sense, this means that non-sharing may be epistemically costly: it protects not only information, but also prevents the emergence of a more robust knowledge ecosystem.

This logic explains why one may argue that universal non-sharing is, from the standpoint of the whole, a "pre-losing" choice, though only under several important conditions. It is not true that any kind of sharing is automatically good or that any kind of non-sharing is automatically mistaken. A more accurate claim is that in an environment where value arises from the aggregation of dispersed knowledge, generalized closure tends to produce systemic loss. Organizations solve similar problems over and over again in isolation, invest in parallel validation processes, fail to exploit collective experience sufficiently, and thereby create an environment in which knowledge accumulates in fragmented rather than synergistic form. This gives rise to a peculiar form of epistemic waste: each actor acts rationally within its own horizon, but in the aggregate the system's ability to learn faster, better, and more broadly is weakened. This is the core of the argument for sharing as a form of coordination-based rationality (Kaushik et al., 2025; Richards, 2001).

At the same time, however, this innovation-oriented logic must be corrected. The mere fact that sharing increases the potential epistemic yield does not by itself mean that it is always legitimate or prudent. Once an organization makes data available within external or partially external model systems, it opens the possibility of unintended leakage, inferential reconstruction, secondary use, or simply a loss of visibility into the layers of the system in which data are further processed and the purposes for which they are used. Carlini et al. (2021) persuasively showed that large language models may, under certain conditions, memorize and reproduce portions of training data, while Rigaki and Garcia (2023) summarize a wider range of privacy attacks in machine learning, demonstrating that the boundary between "learning from data" and "complete loss of control over data" is not as sharp as is sometimes assumed. These findings fundamentally alter the normative status of sharing. If the risk of extraction or secondary use is real, then restraint is not merely an expression of conservative mentality, but a fully rational form of defense.

For this reason, both naive techno-optimism and reflexive data isolationism must be rejected. The former overlooks the price of trust and the normative significance of control over sensitive information; the latter overlooks the collective costs of epistemic fragmentation. The theoretical issue, therefore, is not which of the two paradigms should triumph once and for all, but rather how institutional conditions should be arranged so that the advantages of sharing may be secured without rendering organizations and data subjects disproportionately vulnerable (Curtis et al., 2023; Reinhardt, 2023; Schultz and Seele, 2023).

6 The Ethical and Legal Dimension: Why Efficiency Alone Is Not Enough

It is precisely here that the inadequacy of a purely utilitarian argument in favor of data sharing becomes evident. It may well be true that coordinated access to data leads to faster innovation, better models, and lower duplication of research or operational costs. Yet this alone does not settle the question of whether such sharing is morally and legally permissible. Organizations often do not handle data that may simply be treated as their private property without further limitation. They handle personal data, professionally entrusted information, sensitive documentation, customer records, or internal know-how, the protection of which forms part of their legal and ethical obligations. In this sense, they are more properly understood as custodians or fiduciary holders of certain informational relationships than as unrestricted owners of informational objects. For that reason, legitimacy cannot be inferred from systemic utility alone. What is useful for the aggregate performance of the system may at the same time be impermissible from the standpoint of confidentiality, autonomy, professional secrecy, or the legitimate expectations of the subjects whose data are being processed (European Commission, 2026; European Parliament and Council of the European Union, 2016).

This objection becomes even stronger once the concept of trust is taken seriously. In ordinary AI debates, the notion of "trustworthy AI" is often invoked as though trustworthiness were simply another technical property of a system. Reinhardt (2023) shows that the concepts of trust and trustworthiness in AI ethics are frequently overused and conceptually blurred. Trusting a human being, trusting an institution, and relying on a technical artifact are not the same thing. Organizations have no moral obligation to "trust" LLM infrastructures in situations in which they lack adequate information about how those infrastructures function, what control mechanisms are in place, or how input data are in fact handled. Trust is not a romantic opposite of caution, but a relation grounded in justified expectation, accountability, and the possibility of sanction. Where these elements are absent, refusing to share is not only excusable, but often reasonable. For that reason, the normative center of the debate should not shift toward a psychologizing question of why organizations "lack trust," but toward the institutional question of whether there are adequate grounds on which trust may be justified.

The legal dimension further strengthens this argument. The GDPR is not built on the idea of maximizing data flow, but on the principles of lawfulness, fairness, transparency, purpose limitation, data minimization, and security. These principles are not mere administrative obstacles to technological development. They constitute a normative defense against allowing efficiency to become the sole measure of legitimacy. Likewise, the evolving European regulation of AI, in particular the AI Act and accompanying codes and guidance, increasingly emphasizes accountability, auditability, safety, and transparency

within model ecosystems. This means that organizations are not judged only on whether they use AI, but also on whether they can explain under what conditions they use it, what data enter the system, how those data are protected, and who bears responsibility for possible harm. In that sense, the question of sharing becomes a question of governance. It is no longer simply a matter of what is technically possible and economically advantageous, but of what is legally defensible and institutionally responsible (European Commission, 2025; European Parliament and Council of the European Union, 2024).

This brings us back to Pareto optimality in its properly limited sense. A Pareto-superior arrangement may be more efficient than generalized non-sharing, but it does not follow that it is thereby automatically just or legally permissible. If organizations as players are made better off while the costs are shifted onto data subjects, patients, employees, or other weaker actors, then no normatively satisfactory conclusion has been reached. Pareto efficiency says nothing about the distribution of burdens, asymmetries of risk, or the legitimacy of consent. Any argument in favor of sharing must therefore be supplemented by the question of who is included in the payoff matrix at all and who remains outside it despite bearing its consequences. Without that supplement, game theory would easily become an elegant but normatively blind abstraction (European Commission, 2026; European Parliament and Council of the European Union, 2016; Reinhardt, 2023).

7 The Paradoxes of the Sharing and Non-Sharing Paradigms

The formalization above helps explain why neither sharing nor non-sharing can be treated as a normatively transparent strategy. Once the relevant payoffs are made explicit, it becomes clear that each paradigm contains tensions that destabilize its own self-justification. Sharing may maximize collective gain while simultaneously increasing exposure, asymmetry, and legitimacy deficits; non-sharing may preserve immediate control while generating fragmentation, strategic weakness, and long-term dependence. The paradoxes analyzed in this section should therefore be read not as rhetorical embellishments, but as second-order consequences of the payoff structure itself. They show that the dilemma cannot be resolved simply by selecting one side of the binary more decisively. What must instead be examined is how each paradigm generates outcomes that partially undermine the very values in whose name it is adopted.

The deepest analytical value of the entire debate becomes visible once one recognizes that neither the paradigm of sharing nor the paradigm of non-sharing leads to a simple and contradiction-free normative result. Each contains its own paradoxical dynamic, that is, a moment in which a strategy adopted in order to realize a certain value generates consequences that partly undermine that very value. This is why it is insufficient to

oppose openness and caution as two morally transparent alternatives. What must instead be analyzed is the way in which each orientation generates paradoxical effects, and how that paradoxical structure in turn transforms its legitimacy.

The sharing paradigm is usually associated with ideas of progress, acceleration, collective intelligence, and innovative synergy. This representation is justified in many respects, because data sharing may indeed accelerate model development, improve the quality of inferences, reduce duplication in solving analogous problems, and support interorganizational cooperation. At the same time, however, this very orientation generates a paradox of acceleration through vulnerability. The more data enter broader model ecosystems, the greater not only the potential benefit, but also the attack surface, the possibility of recontextualization, memorization, or secondary use. Organizations thus accelerate their knowledge production by rendering themselves epistemically, legally, and security-wise more vulnerable. This paradox is not hypothetical. Empirical findings concerning extractability of training data and diverse forms of privacy attack show that the distinction between "learning from data" and "total loss of control over data" is not nearly as sharp as is often presumed (Carlini et al., 2021; Rigaki and Garcia, 2023).

The sharing paradigm is also bound up with a paradox of cooperation leading to asymmetrical extraction. Sharing is often normatively presented as a common project in which all benefit from a more open circulation of knowledge. Yet in an environment of unequal capacities, what may actually occur is that the value generated by openness is disproportionately captured by those actors who possess the greatest computational infrastructure, the strongest monetization capacity, and the most powerful legal position. Those who provide data may thereby contribute to the creation of common model capital without participating in its benefits in a corresponding measure. Cooperation may thus turn into asymmetrical extraction, that is, a situation in which the language of public good conceals a centralization of power and value. Pasetti et al. (2025) show that the opaque governance of generative AI training data raises precisely such questions concerning the distribution of benefits, responsibility, and normative legitimacy. This demonstrates that sharing may be cooperative only at the level of rhetoric, while in institutional reality it may reinforce inequality.

A further paradox is the paradox of transparency and trust. Intuitively, it seems self-evident that the more transparent the use of AI becomes, the greater the trust of the public, clients, or partners will be. Empirical findings, however, complicate this intuition. Schilke et al. (2025) show that merely disclosing the use of AI may in certain contexts actually weaken trust. Transparency here does not necessarily strengthen legitimacy, but may instead activate concern, heighten perceptions of alienation, or provoke suspicion of excessive automation. The paradox lies in the fact that openness, which is meant to build trust, may instead undermine it. From a philosophical point of view, this means

that trust is not a simple function of visibility. It requires not only disclosure, but also an intelligible framework of responsibility, corrigibility, and control (Schilke et al., 2025; Reinhardt, 2023).

Sharing also generates a paradox of efficiency without legitimacy. A regime may be manifestly superior in terms of system performance, speed, and productivity, and yet remain legally or ethically fragile. This is particularly important wherever the aggregate benefits of the system rest on practices that weaken the autonomy of data subjects, exceed the original purpose for which data were collected, or shift a disproportionate degree of risk onto parties who are not the direct beneficiaries. In such circumstances, efficiency is achieved at the expense of legitimacy. This is a warning against a technocratic interpretation of Pareto optimality: a more efficient arrangement is not thereby automatically justified if it is premised on the externalization of harms (European Commission, 2026; European Parliament and Council of the European Union, 2016).

Alongside these paradoxes of sharing, the paradoxes of non-sharing must likewise be taken seriously. The paradigm of restriction is usually motivated by security, autonomy, preservation of control, and legal caution. These motives are often entirely legitimate, yet once generalized they may generate their own forms of self-undermining. The first of these is the paradox of security leading to weakness. An organization that closes itself off from the risks of sharing may preserve a high degree of control in the short term, yet at the same time lose contact with technological development, with the practice of building internal competence, and with the ability to use AI effectively in an environment where AI is becoming a standard component of competitive and institutional life. Protection against immediate risk may thus become a source of long-term strategic weakness. This is especially visible in sectors where speed of learning and adaptive capacity are themselves competitive advantages (Kaushik et al., 2025; Wornow et al., 2023).

Non-sharing is also associated with a paradox of autonomy leading to isolation. The enclosure of data resources is often justified as preserving sovereignty and informational control. But if this strategy becomes a general norm, the result is a fragmented ecosystem characterized by low interoperability, limited ability to coordinate, and a high degree of duplicative learning. Each actor retains autonomy, but the whole suffers from isolation. In this way autonomy becomes a collective inability to improve together. It is precisely here that the Prisoner's Dilemma converges with Pareto inefficiency: actors remain in a stable regime that protects their independence, yet prevents transition to a superior arrangement (*Prisoner's Dilemma and Cooperation* 2021; Richards, 2001).

A further paradox of non-sharing is that caution may lead to greater dependency. Organizations that long resist developing their own data-grounded AI processes or safe cooperative schemes may ultimately become more dependent on external commercial solutions, because they fail to develop the requisite internal expertise, governance capacity,

or technical infrastructure. What was meant to protect autonomy may thereby weaken the capacity for independent judgment. Caution here is not a guarantee of sovereignty; it may become a mechanism of its erosion. Similarly, legal restraint may paradoxically produce legal unpreparedness. Institutions that delay working with LLMs out of concern for regulatory uncertainty may remain without compliance processes, without established audit trails, and without practical understanding of how to satisfy new obligations. In the short term they may reduce exposure, but in the long term they weaken their preparedness for a technological and legal reality that is very unlikely to bypass them (Curtis et al., 2023; European Commission, 2025; European Parliament and Council of the European Union, 2024).

In certain sectors non-sharing has yet another dimension: the paradox of protecting trust while reducing public value. Especially in healthcare, science, or public administration, an excessively restrictive data regime may protect a narrow institutional conception of trust while simultaneously weakening the capacity to generate broader epistemic or public-interest outcomes. Wilke (2025) shows that willingness to share health data is closely connected with trust and with the perception of public purpose. If institutions choose radical restriction, they may protect the immediate relationship to the individual data subject, but at the same time diminish their ability to contribute to a wider public good. From the perspective of institutional philosophy, this is highly significant: protecting trust in a narrow sense does not necessarily mean maximizing legitimacy in a broader one.

What follows from this overview is a decisive conclusion. Neither the sharing paradigm nor the non-sharing paradigm can be absolutized, because both generate their own paradoxical consequences. Sharing may lead to accelerated vulnerability, asymmetrical extraction, and efficiency without legitimacy; non-sharing may lead to strategic weakness, isolation, dependence, and unpreparedness. The real dispute is therefore not between the good of openness and the good of protection, but between different regimes for distributing risk, responsibility, and trust.

8 Institutional Design as the Reparameterization of the Game

If the organizational data-sharing dilemma is correctly understood as a strategically stable yet Pareto-inferior equilibrium, then the solution cannot consist in moral exhortation alone. Appeals to openness, cooperation, or innovation may have rhetorical force, but they do not by themselves change the incentive structure under which organizations act. From the standpoint of the formal model introduced above, the decisive question is not whether actors can be persuaded to become less self-interested, but whether the

institutional environment can be redesigned in such a way that the parameters of the game are altered. The practical and normative significance of institutional design lies precisely here: it changes the structure within which rationality is exercised. Rather than asking organizations to ignore risk in the name of collective progress, it seeks to lower the cost of cooperation, reduce the attractiveness of unilateral appropriation, increase the expected gains from reciprocal participation, and transform generalized restraint from the uniquely prudent strategy into one among several institutionally mediated options.

This point can be expressed directly in the language of the model. Cooperation becomes more plausible when the effective cost of sharing, C, is reduced; when the expected collective innovation gain, G, is rendered more visible and realizable; when the appropriative advantage of unilateral withholding, A, is diminished; and when the safeguard and trust parameters, R and T, are increased. Conversely, when legal ambiguity, technical opacity, and weak accountability make C highly salient and uncertain, while leaving A largely intact, organizations will predictably converge toward non-sharing. Institutional design should therefore be understood as a process of strategic reparameterization. It does not abolish the Prisoner's Dilemma by dissolving self-interest, but by altering the relative magnitudes of the relevant payoffs such that cooperation can become a rational equilibrium rather than a naive gamble. In this respect, governance is not a secondary ethical overlay applied after technical deployment; it is one of the primary determinants of the game's structure (Curtis et al., 2023; Schultz and Seele, 2023).

The point of Table 1 is not to suggest a one-to-one deterministic mapping between a single institutional intervention and a single strategic variable. In practice, many interventions affect several parameters at once. Federated learning, for example, may lower C directly while also increasing R; contractual reciprocity may reduce A while indirectly increasing T. The table is intended instead as an analytic bridge between the model's abstract variables and the concrete mechanisms through which institutional design operates in real organizational environments.

A useful way of making this more concrete is to move beyond the binary vocabulary of sharing and non-sharing and instead to conceptualize cooperation as a staged trust architecture. Real organizations rarely confront a choice between unrestricted openness and absolute closure. More commonly, they move through a graded sequence of sharing arrangements in which the expected costs, risks, and gains differ substantially. At the lowest level lies a regime of internal-only data use, corresponding to a high-control but low-cooperation equilibrium. This regime minimizes immediate exposure but leaves the collective innovation gain G unrealized and preserves the system-wide opportunity loss δ. A second level may involve the exchange of synthetic, benchmark, or otherwise low-risk datasets. Such arrangements typically do not eliminate the cost of sharing, but they reduce C sufficiently to allow institutions to test interoperability, compliance routines,

and reciprocal expectations without incurring the full vulnerability associated with high-value raw data exchange. A third level may consist of federated or privacy-preserving analytics, where raw data remain local but model parameters, encrypted computations, or aggregated outputs are exchanged. In the formal model, this kind of architecture does not merely reduce exposure in a vague sense; it changes the expected payoff structure by lowering the effective cost of participation and, under some circumstances, increasing the safeguard term R. A fourth level may introduce role-limited, contractually bounded access to selected sensitive datasets under auditable governance and explicit use restrictions. Finally, a mature consortium regime may combine repeated interaction, independent oversight, sanction-backed reciprocity, and shared governance boards in a way that increases the trust premium T while simultaneously reducing the attractiveness of unilateral appropriation A. The point of this staged structure is that cooperation need not arise all at once; it can be institutionalized incrementally through transitions between regimes whose strategic properties differ in systematic ways.

The technical mechanisms often invoked in contemporary AI governance debates can now be interpreted more rigorously. Federated learning, for example, is often presented as a privacy-preserving solution because it avoids central transfer of raw data. In the present framework, its strategic significance lies in the fact that it may lower the effective cost of sharing, C, by reducing exposure of directly identifiable or competitively sensitive information, while preserving at least part of the collective gain G (Cheng et al., 2020; Yin et al., 2021). Secure multi-party computation and related cryptographic approaches may play a similar role by allowing joint computation without full information disclosure, thereby increasing R, the effective mitigation afforded by technical safeguards. Auditable logs, purpose limitation, and fine-grained access controls reduce not only the expected cost of misuse but also the uncertainty associated with institutional vulnerability. This matters because organizations often act not only on expected risk but on opacity-induced precaution. By making flows of access and responsibility more visible, such measures may reduce the subjective and objective components of C simultaneously. Contractual prohibitions on secondary training, explicit liability allocation, and enforceable restrictions on reuse may also reduce A, because they constrain the capacity of a non-reciprocating actor to profit asymmetrically from another's openness. In this way, legal and technical design operate not as external restraints on the game, but as endogenous determinants of its strategic landscape.

The role of repeated interaction is equally important. In a one-shot Prisoner's Dilemma, withholding remains attractive because there is no future in which opportunism can be punished or cooperation rewarded. In repeated settings, however, trust can acquire strategic substance. This is what the parameter T is meant to capture. It does not refer to trust as a diffuse sentiment, but to the expected payoff increment generated by credible reciprocity, verified compliance, institutional memory, and the possibility of sanction.

Where actors know that present behavior will shape future access, future reputation, or future collaborative gains, the incentive structure changes. Repeated interaction can increase the expected value of mutual cooperation and reduce the attractiveness of one-sided defection. Yet this point should not be romanticized. Repetition alone is insufficient. Organizations do not trust simply because they interact repeatedly; they trust when repeated interaction is embedded in a setting where deviations can be observed, verified, and sanctioned. This is why the architecture of trust must remain institutional rather than purely psychological. Auditability, traceability, dispute resolution, and compliance verification are what convert repeated exposure into a meaningful increase in T rather than a mere repetition of vulnerability (Reinhardt, 2023; Schultz and Seele, 2023).

This also clarifies why legal certainty is strategically important. Regulatory ambiguity raises the expected and perceived cost of sharing because it amplifies uncertainty regarding sanction, liability, and ex post interpretation. Even if the objective benefits of cooperation are high, organizations may remain reluctant to share if they cannot reliably estimate the legal meaning of their actions. Clearer regulatory standards, guidance on permissible uses, and standardized contractual templates therefore matter not merely as compliance conveniences, but as variables affecting the game itself. They reduce the indeterminate component of C, enable organizations to distinguish between tolerable and intolerable exposure, and render institutional safeguards more legible. In the European context, this is one reason why the GDPR, the AI Act, and emerging guidance on general-purpose AI models should not be read only as constraints. They also function as potential coordination devices. Under favorable conditions, regulatory clarification can lower the uncertainty that pushes actors into defensive non-sharing, thereby making more limited and auditable forms of cooperation strategically conceivable (European Commission, 2026; European Parliament and Council of the European Union, 2016; European Commission, 2025; European Parliament and Council of the European Union, 2024).

The model further helps explain why some institutional interventions fail. If a governance scheme reduces C only marginally while leaving A largely untouched, the temptation to free-ride or defect will remain substantial. Likewise, if a consortium promises large collective gains in principle but offers weak enforcement, poor observability, and no meaningful sanction, then the formal value of T remains low even if the rhetoric of trust is abundant. This is one reason why many appeals to "responsible AI" or "trustworthy ecosystems" remain normatively attractive but strategically thin. They invoke the right values without sufficiently transforming the parameters that make mutual withholding rational. The real test of institutional design is therefore not whether it sounds cooperative, but whether it changes the expected payoff ordering in a way that can be sustained under conditions of bounded trust, partial information, and asymmetrical capacity.

For that reason, the transition from a Pareto-inferior equilibrium to a more cooperative

regime should be understood not as a single leap but as a mechanism-design problem. One can imagine several possible institutional strategies. A regulator may subsidize safe-sharing infrastructures, thereby lowering C. A consortium may require reciprocal contribution as a condition of access, thereby reducing A. A technical architecture may replace raw-data transfer with federated analytics, thereby increasing R. A repeated governance framework with third-party audit may increase T. Each of these interventions changes the game in a different way. Their common purpose is not to force cooperation abstractly, but to make cooperation less fragile and less irrational. In a well-designed institutional environment, mutual sharing need not rely on moral heroism. It can emerge as a strategically sensible equilibrium because the cost of responsible participation is lowered, the opportunities for unilateral appropriation are curtailed, and the gains from reciprocal contribution become more credible.

Seen in this light, institutional design is best understood as the practical answer to the normative insufficiency of both paradigms analyzed earlier in the paper. Pure openness is normatively unstable because it can produce asymmetrical extraction, vulnerability, and efficiency without legitimacy. Pure restraint is strategically costly because it can generate fragmentation, weakness, and long-term dependency. A staged, auditable, and technically mediated architecture of trust does not eliminate these tensions entirely, but it offers a way of managing them without collapsing into either extreme. The central practical question is therefore not whether organizations should share data in the abstract, but under what institutional conditions the expected payoff from limited, legitimate, and protected cooperation exceeds the expected payoff from generalized withholding. Once formulated in this way, the problem becomes analytically tractable, normatively more precise, and substantially closer to empirical validation.

9 Evolutionary Population Simulation: Design and Results

This simulation section operationalizes the stylized extensions introduced earlier in the paper. More specifically, it translates the baseline two-player model into a heterogeneous population setting with repeated interaction, adaptive trust, and evolutionary updating of strategy shares. Its purpose is not to replace the analytical model, but to show how the article's central claims behave once the game is embedded in a more realistic ecology of organizational diversity and institutional variation.

The argument developed in this article is primarily conceptual and analytical, but it is not intended to remain merely speculative. On the contrary, one of its central claims is that the organizational data-sharing dilemma around LLMs can be reformulated in a way that makes it empirically investigable. The present section therefore introduces

a proof-of-concept evolutionary population simulation. Its purpose is not to provide definitive empirical validation, but to demonstrate that the formal model generates coherent dynamics when extended from a stylized two-player setting to a heterogeneous organizational population. More specifically, the simulation examines whether institutional interventions can alter long-run cooperation, welfare, distribution of payoffs, and strategic composition across organizations.

9.1 Simulation Design

The simulated population consists of N organizations, each modeled as an agent embedded in a repeated interaction environment. Unlike the base formal model, which is intentionally minimal, the simulation introduces both heterogeneity and evolutionary updating. Each organization is characterized by baseline private benefit (B_i), collective innovation gain (G_i), sharing cost (C_i), appropriation advantage (A_i), and systemic loss from mutual non-sharing (δ_i). In addition, the simulation includes the institution-sensitive parameters safeguard effect (R_i) and trust premium (T_i), which correspond to the paper's broader claim that governance alters the payoff structure rather than merely exhorting actors to cooperate.

The population is composed of four strategy types: defectors, who always choose non-sharing; unconditional sharers, who always choose sharing; conditional cooperators, who share when recent cooperation in the environment exceeds an individual threshold; and opportunists, who share only when expected cooperation is sufficiently high and appropriation risk sufficiently low. This strategy ecology was chosen because it allows the simulation to represent not only a binary choice but also a structured diversity of organizational orientations. In this respect, the simulation moves closer to the article's institutional argument than a simple repeated dyadic game.

At the beginning of each run, agents are assigned heterogeneous parameter values drawn from distributions that represent organizational variation in perceived benefit, exposure, and expected collective gain. Initial strategy shares are distributed across the four strategy types. In each epoch, agents are randomly paired and interact once. The realized payoff structure preserves the formal model while allowing regime-specific adjustments:

$$\pi_i(S, S) = B_i + G_i - C_i^* + R_i + T_i,$$

$$\pi_i(S, N) = B_i - C_i^*,$$

$$\pi_i(N, S) = B_i + A_i^*,$$

$$\pi_i(N, N) = B_i - \delta_i,$$

where C_i^* and A_i^* denote effective cost and effective appropriation advantage under a

given governance regime.

Beliefs are updated adaptively from observed cooperation in the environment. Trust is also dynamic: successful mutual sharing increases T_i, while exploitation reduces it. In this way, the simulation gives formal content to the claim that institutional trust is not an exogenous moral virtue, but a cumulative feature of repeated, observable, and sanctionable interaction. Strategy composition then evolves over time through a replicator-like imitation process: strategies associated with higher average payoff become more attractive in the subsequent epoch, subject to a small mutation or exploration rate that prevents the system from collapsing into a perfectly absorbing corner. This mechanism does not model organizations as perfectly optimizing agents with full information. Rather, it models an evolving ecology of strategies under bounded rationality and selective adaptation.

The simulation compares three institutional regimes. In Regime 1, no meaningful safeguards are present; $R_i = 0$, sharing cost remains high, and initial trust is low. In Regime 2, technical safeguards reduce effective sharing cost and modestly lower appropriation advantage, but reciprocity remains institutionally weak. In Regime 3, technical safeguards are combined with a trust architecture that further reduces effective sharing cost, reduces appropriation advantage, and starts from a higher trust baseline. These regimes correspond to the article's broader distinction between unstructured exposure, technically mediated cooperation, and combined technical-institutional governance.

9.2 Simulation Results

The population-based evolutionary simulation supports the article's central claim that institutional design alters the strategic environment in which organizations decide whether to share data. At the same time, it suggests that the effects of governance are more gradual and differentiated than a simple transition from defection to cooperation would imply. Across all runs, the three governance regimes produced distinct patterns in cooperation, welfare, payoff dispersion, and long-run strategy composition.

Figure 1 shows the evolution of population-level cooperation over time. Under Regime 1, where no meaningful safeguards were present, cooperation quickly settled at the lowest level. Under Regime 2, where technical safeguards reduced effective sharing costs, cooperation stabilized at a modestly higher level. Regime 3, which combined technical protections with institutional trust architecture, produced the highest cooperation rate, but the increase over Regime 2 was incremental rather than dramatic. This result is theoretically significant because it suggests that stronger governance does improve cooperation, yet does not eliminate strategic caution or produce a simple high-cooperation equilibrium across the population.

Figure 2 shows the corresponding welfare dynamics. Average welfare was clearly lowest

under the no-safeguards regime and substantially higher once technical safeguards were introduced. However, the difference in welfare between Regime 2 and Regime 3 was comparatively small. This suggests that, in the present model, the largest aggregate benefit comes from reducing the effective cost of sharing, while additional institutional trust mechanisms contribute more to stabilization and strategic coordination than to dramatic further welfare expansion. The result therefore refines the paper's broader argument: institutional design matters, but its components do not all operate with equal force. Cost-reducing safeguards appear to be especially important.

Figure 4 provides an additional result that is normatively important. Final payoff dispersion was highest under the no-safeguards regime and lowest under the strongest governance regime. The evolutionary simulation therefore indicates that stronger governance may improve not only cooperation and welfare but also the distributional profile of outcomes. This is especially relevant to the article's concern with the distinction between Pareto improvement and normative adequacy. While stronger institutional design does not eliminate inequality, it appears in this model to reduce the dispersion of realized payoffs across organizations.

The strategy-composition dynamics in Figure 3 further clarify the character of this outcome. Under the strongest governance regime, the population does not converge toward a single dominant strategy. Instead, it approaches a relatively mixed ecology in which defectors, conditional cooperators, and opportunists remain present at similar levels, while unconditional sharers become less prevalent. This finding suggests that institutional trust does not abolish strategic diversity. Rather, it supports a more balanced strategic environment in which different orientations can coexist without collapsing into the low-cooperation pattern characteristic of the weakest regime. From the perspective of institutional theory, this is a valuable result: successful governance may not require unanimous cooperation, but only a sufficiently structured environment in which opportunism is contained and conditional cooperation remains viable.

The heatmap in Figure 5 helps identify which parameters are most influential in shifting the population toward cooperation. Final cooperation rates increase more consistently as effective sharing cost is reduced than as initial trust alone is increased. Trust has a positive effect, but it appears to function mainly as an amplifying rather than primary driver. In this evolutionary setting, cost reduction is the more powerful direct lever, while trust contributes to incremental improvement once the environment is already less punitive. This supports the paper's central claim that governance works by reparameterizing the game, but also suggests a more specific lesson: some forms of reparameterization may matter more than others, and reducing the burden of participation may be more immediately consequential than attempting to build trust in the abstract.

Taken together, the evolutionary simulation should be interpreted as a proof-of-concept

computational illustration of the model's comparative statics. It does not provide direct empirical validation of any particular sector, nor does it imply that stronger governance will automatically produce high-cooperation equilibria in practice. It does, however, show that institutional regimes influence the long-run ecology of strategies within a heterogeneous population. Weak safeguards reproduce low-cooperation and high-dispersion outcomes; technical safeguards substantially improve welfare and modestly improve cooperation; and stronger trust architecture adds further gains in cooperation, lowers payoff dispersion, and stabilizes a more balanced strategic composition. These results strengthen the article's broader argument that the problem of LLM-era data sharing is neither solved by moral appeals to openness nor exhausted by simple cost-benefit reasoning, but depends on the institutional structuring of incentives, protection, and reciprocity.

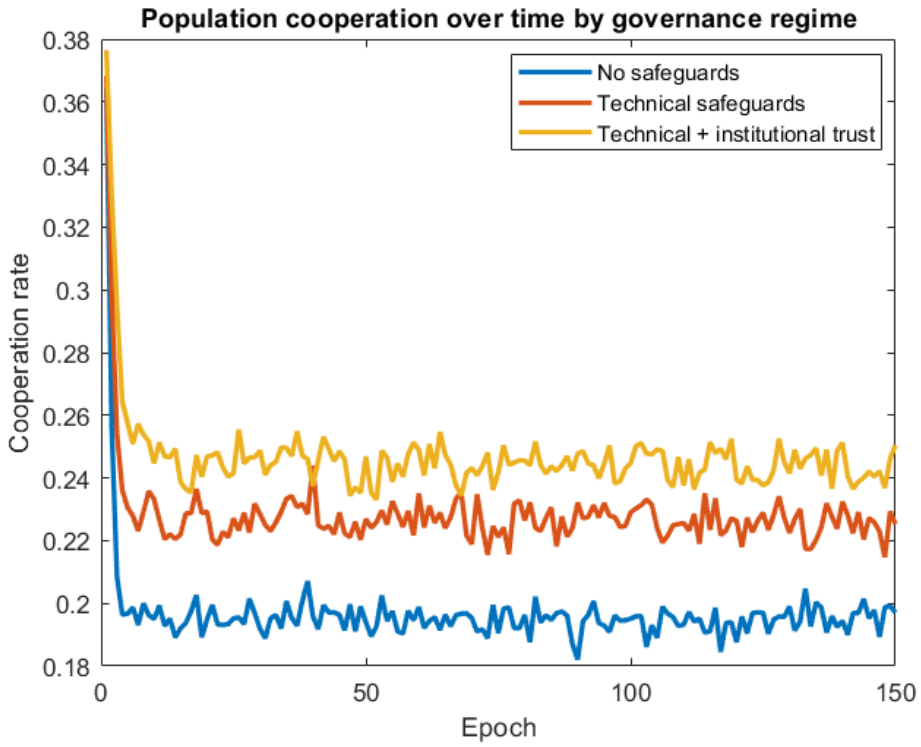

Figure 1: Population cooperation over time under three governance regimes. Stronger institutional design increases cooperation, but the gains are incremental rather than transformative in the evolutionary population model.

A first route beyond the present proof-of-concept is broader population-level simulation. The formal model introduced above lends itself naturally to environments in which heterogeneous organizations interact under varying governance regimes, network structures, and information conditions. Such extensions could model sectoral differences in privacy exposure, expected innovation gains, appropriability, legal safeguards, and trust conditions with greater realism than the current illustration. The analytical value of such simulations would not lie in predicting a single real-world outcome, but in exploring how different institutional architectures affect equilibrium behavior, strategic persistence, and

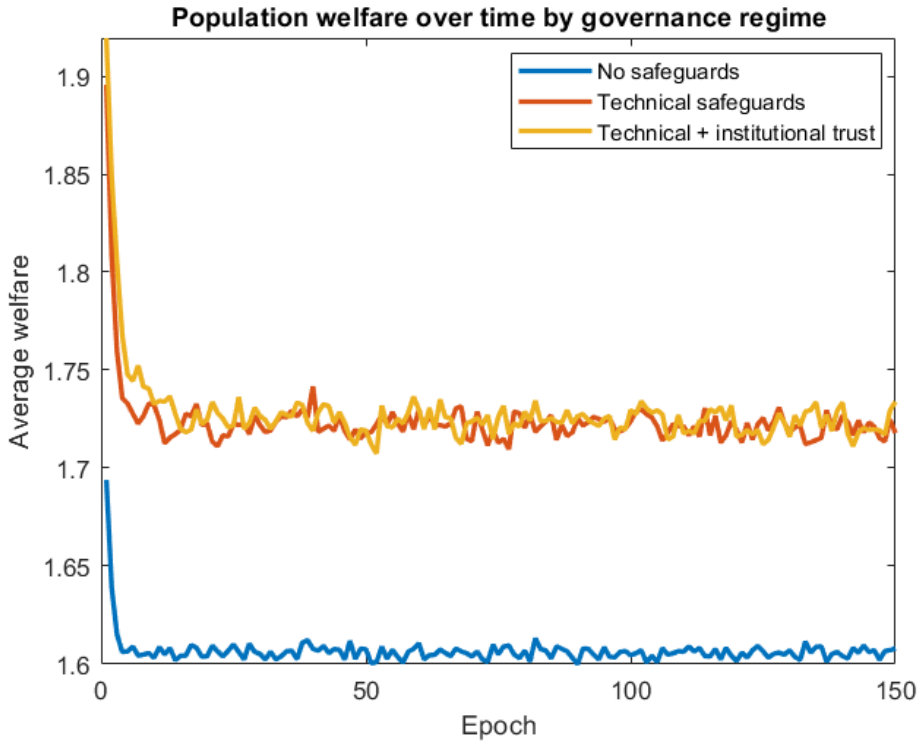

Figure 2: Population welfare over time under three governance regimes. Introducing technical safeguards produces the largest welfare gain, while the addition of institutional trust yields a smaller but still positive improvement.

the distribution of gains over time.

A second route is comparative case-study research focused on existing or emerging data-sharing regimes. The framework developed in this article is especially well suited to sectors in which the tension between collective learning and institutional caution is already pronounced, such as healthcare, finance, research infrastructures, public administration, or critical industrial systems. A comparative design could examine whether organizations embedded in stronger trust architectures are in fact more willing to participate in restricted or staged sharing arrangements than organizations operating in more opaque or weakly governed environments. Such a study would not require perfect quantification of the model's parameters in advance. It could instead proceed through structured comparison of organizational narratives, governance arrangements, compliance mechanisms, and observed sharing practices. For instance, one might compare a healthcare consortium using federated analytics and auditable access controls with another consortium relying primarily on bilateral agreements and manual oversight.

A third possibility is survey-based or experimental research directed at organizational decision-makers. Here the aim would be to estimate perceived payoff structures rather than directly observe large-scale institutional equilibria. Respondents could be presented with systematically varied scenarios involving data sharing for LLM-related purposes,

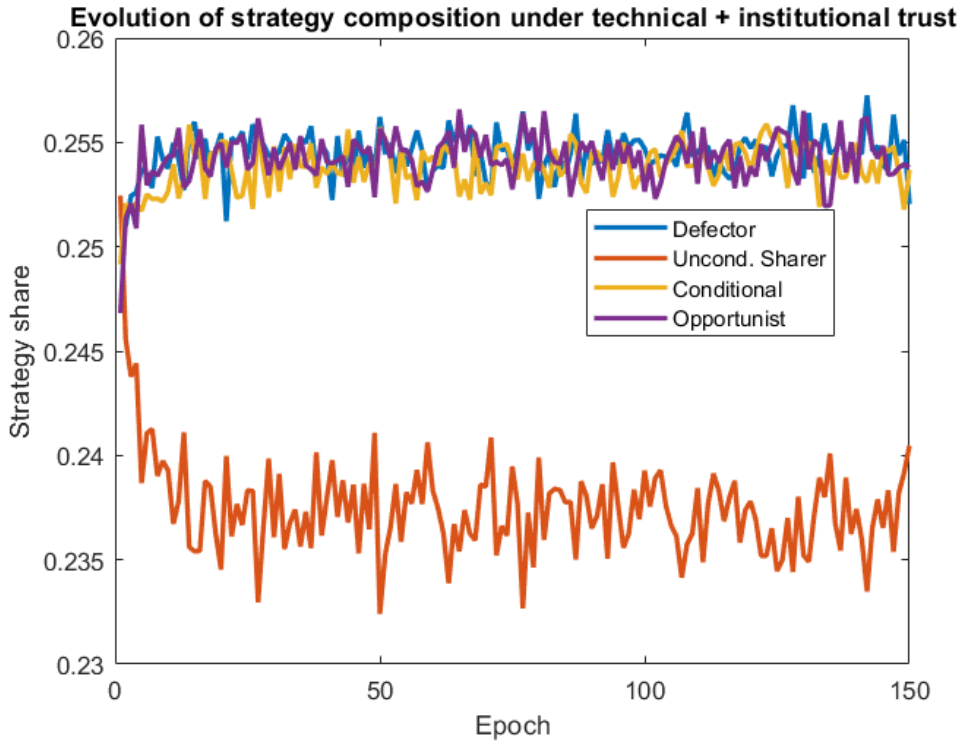

Figure 3: Evolution of strategy composition under the technical-plus-institutional-trust regime. The population converges toward a mixed strategic ecology rather than a single dominant strategy, indicating that stronger governance stabilizes coexistence rather than unanimity.

where each scenario manipulates one or more relevant variables: degree of privacy protection, level of auditability, existence of contractual restrictions on secondary use, expected reputational exposure, reciprocity guarantees, or technological architecture such as federated learning versus raw data transfer. Participants could then be asked to evaluate their organization's likely willingness to share, perceived risk, expected gain, and trust in counterpart compliance. Such a design would allow the researcher to examine which factors most strongly influence willingness to cooperate and whether these factors cluster in ways consistent with the model.

Experimental economics offers a fourth and especially promising path. One could construct a modified Prisoner's Dilemma or public-goods game in which participants act as proxies for institutional decision-makers rather than as isolated individuals. The experimental environment could include treatments corresponding to different governance architectures: no safeguards, technical safeguards only, legal safeguards only, repeated interaction, public audit, or combinations thereof. In one set of treatments, participants might receive advisory suggestions generated by human-written heuristics; in another, they might receive recommendations framed as if produced by an LLM-based governance assistant. This would make it possible to test one of the article's more novel claims: namely, that LLMs may matter not only as infrastructures fed by shared data, but also

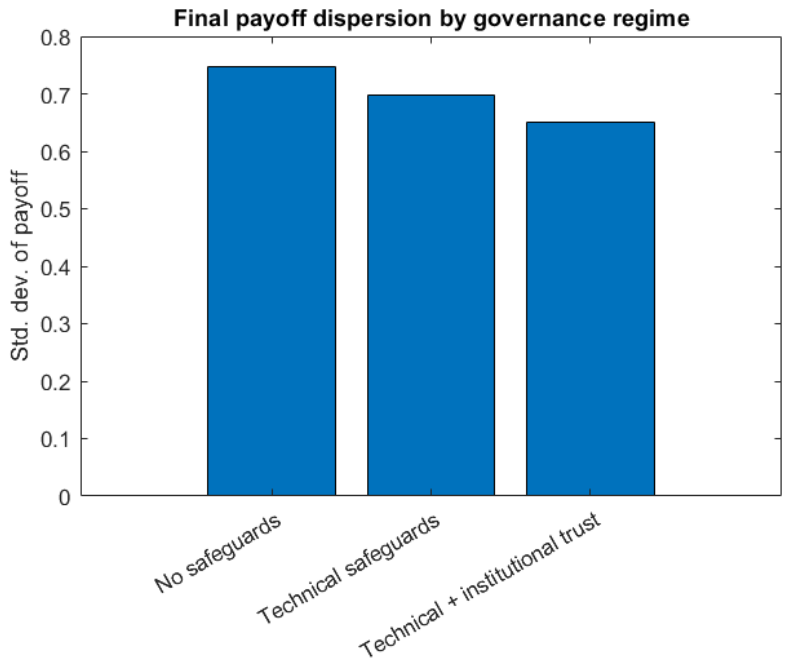

Figure 4: Final payoff dispersion by governance regime. Stronger governance is associated with lower dispersion of realized payoffs, suggesting that institutional safeguards may improve not only aggregate outcomes but also their distributional profile.

as mediators of organizational reasoning about whether to share in the first place.

Taken together, these possible research paths show that the paper's claims are not inherently resistant to evidence. The article remains conceptual in its broad orientation, but it is not speculative in the pejorative sense. It offers a model, a set of parameters, a typology of paradoxes, a mechanism-design hypothesis about how institutional architecture changes incentives, and a first computational illustration of those claims. These are exactly the kinds of contributions that can guide future simulation, comparative institutional analysis, survey research, and experimental inquiry. The point is not that any one method alone will settle the question. Rather, the value of the framework lies in the fact that it opens multiple routes through which the relation between organizational rationality, institutional trust, and LLM-mediated data sharing may be investigated in a cumulative and interdisciplinary way.

10 Limitations

The present article should be read as a high-level institutional and analytical framework rather than as a fully calibrated predictive model. Several limitations follow from that scope. First, the formal game presented in the paper is intentionally stylized. Its payoff functions are additive and simplified in order to isolate the strategic logic of the dilemma. Real-world data-sharing environments may involve nonlinear collective gains,

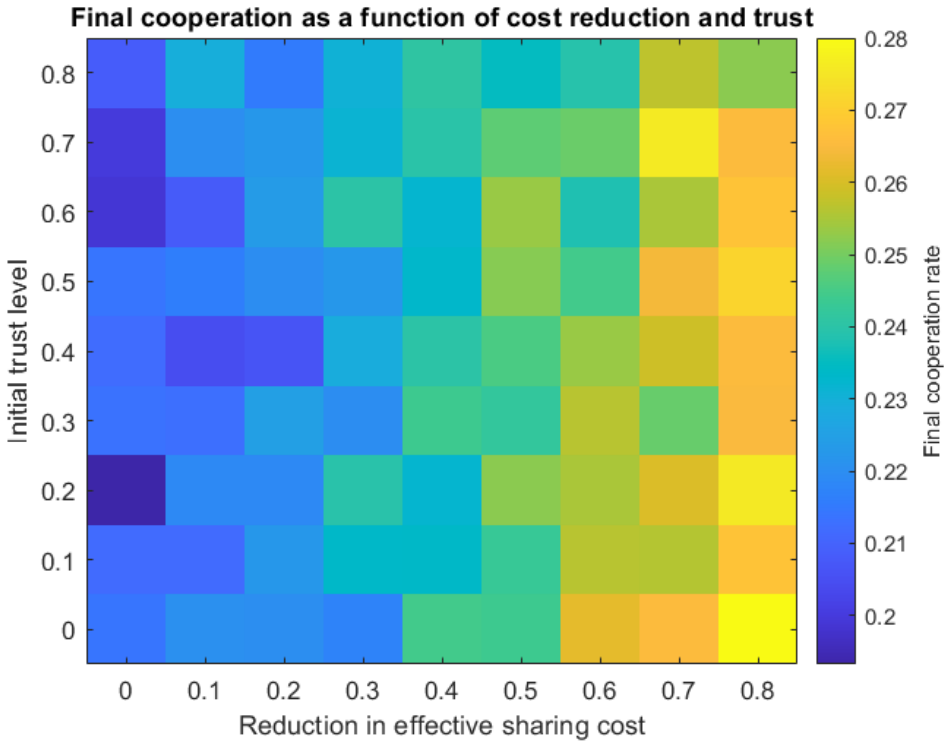

Figure 5: Final cooperation as a function of effective sharing-cost reduction and initial trust. In the evolutionary model, cost reduction appears to be the stronger primary driver of cooperation, while trust has a complementary and amplifying effect.

sector-specific complementarities among datasets, threshold effects, multi-stage negotiation, and complex asymmetries in bargaining power that are not fully captured here.

Second, although the paper introduces repeated interaction, population heterogeneity, and adaptive trust through the evolutionary simulation, the simulation remains proof-of-concept in character. It is designed to illustrate comparative statics rather than to deliver empirically calibrated forecasts for any specific domain. The parameter values are theoretically motivated and interpretable, but they are not estimated from field data. The simulation should therefore be understood as demonstrating the plausibility of the article's mechanism rather than verifying its empirical magnitude in practice.

Third, the article treats institutional trust and safeguards as analytically distinguishable parameters, but in real settings these are often jointly produced by overlapping legal, technical, and organizational arrangements. Likewise, the distinction between objective and perceived payoffs may be blurred when decision-makers rely on advisors, consultants, or AI systems to interpret risk and opportunity. The paper's stylized representation of LLMs as advisory mediators is intended to open a line of inquiry, not to settle the question of how such systems shape organizational strategy in actual governance settings.

Fourth, the framework is deliberately cross-sectoral, which is both a strength and a limi-

tation. It allows the paper to identify structural regularities across domains, but it does not provide a one-size-fits-all governance prescription. The relative magnitudes of sharing cost, appropriative asymmetry, collective gain, and trust are likely to differ substantially across healthcare, finance, public administration, science, and commercial AI ecosystems. Any practical application of the framework would therefore require domain-specific calibration and institutional interpretation.

Finally, the article does not claim that institutional design alone resolves all normative tensions. Even under improved governance, higher welfare and greater cooperation do not automatically eliminate asymmetry, unequal bargaining power, or disputes about legitimacy. For this reason, the framework should be taken as a tool for clarifying institutional trade-offs rather than as a complete normative solution. Its value lies in showing how strategic rationality, legal defensibility, and institutional trust can be analyzed together, while leaving open the empirical and political work required to translate that analysis into practice.

11 Conclusion

This article has argued that organizational data sharing in the context of large language models is best understood not as a simple moral choice between openness and restraint, but as a structured institutional dilemma in which individually rational strategies may generate collectively inferior outcomes. The central analytical claim has been that the relevant problem is not the absence of Nash equilibrium, but the possibility that a strategically stable equilibrium may remain Pareto-inferior. Once organizations confront significant exposure to privacy risk, legal uncertainty, competitive leakage, or governance opacity, non-sharing may emerge as the dominant strategy even where mutual sharing would generate larger collective gains. In that sense, the contemporary data-sharing problem surrounding LLMs exemplifies a broader phenomenon of institutional rationality under conditions of constrained trust: actors do not necessarily defect because they reject cooperation as such, but because the current architecture makes cooperation too costly, too uncertain, or too weakly protected to be strategically reasonable.

The formal model introduced in this paper was intended to make that diagnosis more precise. By distinguishing among baseline AI benefit B, collective innovation gain G, sharing cost C, appropriation advantage A, systemic loss from mutual withholding δ, and the institutional modifiers R and T, the analysis showed how the strategic structure of the dilemma can be explicitly formulated rather than merely asserted. This formalization matters because it shifts the discussion away from vague appeals to trust or innovation and toward a more rigorous account of how legal, technical, and organizational conditions affect rational choice. If mutual non-sharing persists, this need not indicate irrationality, conservatism, or ethical failure on the part of institutions. It may instead indicate that the

parameters of the game remain configured in such a way that withholding is the prudent response. The significance of this point is both analytical and normative: it explains why moral exhortation is insufficient and why the practical task is one of institutional redesign rather than persuasion alone.

On that basis, the article proposed that institutional design should be understood as the reparameterization of the game. Legal clarification, auditability, contractual restriction of opportunism, federated and privacy-preserving architectures, repeated interaction, and sanction-backed reciprocity do not merely accompany cooperation from the outside. They alter the strategic conditions under which cooperation becomes possible. Their practical function is to reduce the effective cost of sharing, lower the attractiveness of unilateral appropriation, increase the value of reciprocal participation, and render trust institutionally credible rather than psychologically wishful. This is why the proposed staged trust architecture is normatively and strategically important. It offers a way of moving beyond the sterile opposition between unrestricted openness and absolute closure by showing how organizations may pass through graduated forms of cooperation, from low-risk and synthetic exchanges to federated analytics and more mature consortium regimes with auditable accountability. In this framework, cooperation becomes not an act of naive exposure, but a conditionally structured and institutionally bounded practice.

The proof-of-concept simulation provides computational support for this argument, while also refining its interpretation. In the evolutionary population model, stronger governance improves cooperation, increases welfare, and reduces payoff dispersion, but the effects are gradual rather than revolutionary. The most substantial welfare gains arise from reducing the effective cost of sharing, while stronger trust architecture contributes additional stabilization and more modest further gains. This suggests that institutional design should not be understood as a singular intervention, but as a layered process in which some levers, especially those that lower participation costs, may be more immediately consequential than others. For that reason, the present framework should be understood not as a final technical solution to the data-sharing problem, but as a structured institutional lens through which more specialized mechanism-design, regulatory, and empirical work may be organized and assessed.

The paper has also argued that the debate cannot be resolved by treating "sharing" as self-evidently progressive or "non-sharing" as self-evidently prudent. The paradox analysis demonstrated that both paradigms generate internal tensions. Sharing can produce acceleration through vulnerability, cooperation through asymmetrical extraction, and efficiency without legitimacy. Non-sharing can produce security through weakness, autonomy through isolation, and caution through long-term dependence. These paradoxes are not merely rhetorical complications. They reveal that neither paradigm can serve as a sufficient normative foundation on its own. What is required instead is a framework

capable of recognizing that efficiency, trust, legality, and justice do not automatically coincide, and that institutional arrangements must be evaluated not only by the gains they promise but also by the kinds of exposure, exclusion, and asymmetry they create.

A further contribution of the article has been to distinguish between two levels of the LLM problem. At one level, organizations decide whether to share data into infrastructures that support large language models and related AI systems. At another level, LLMs themselves may increasingly shape the decision environment by functioning as advisory or delegated tools in governance, risk assessment, compliance reasoning, or strategic coordination. This second level suggests that the relevance of LLMs is not exhausted by the fact that they depend on data. They may also influence how institutions perceive payoffs, risks, reciprocity, and trust. For that reason, the literature on LLM behavior in social-dilemma settings is potentially important not only as an analogy but as an emerging factor in the design of institutional decision support. The stronger claim is not that LLMs have replaced organizations as strategic actors, but that organizations may increasingly act through model-mediated interpretations of the strategic environment. Once that occurs, the design of the advisory layer becomes part of the governance problem itself.

Finally, the article has sought to show that its claims are open to further validation rather than destined to remain purely conceptual. The model developed here can guide simulation, comparative case analysis, survey-based organizational research, and experimental inquiry into how institutions perceive and negotiate data-sharing trade-offs. The paradoxes identified in the paper can be operationalized and examined empirically. The institutional design proposals can be tested against actual or simulated regimes of trust, compliance, and technical safeguard. The framework therefore does not aim to close the debate, but to reformulate it in a way that is analytically sharper and more empirically tractable.

The broad conclusion is that the future of organizational data sharing around large language models will not be determined solely by technological capability, nor by abstract ethical commitments to openness, nor by defensive insistence on closure. It will be determined by whether societies, regulators, firms, and research consortia are able to build architectures in which cooperation becomes simultaneously rational, auditable, and legitimate. The decisive distinction is not simply between sharing and non-sharing, but between unstructured exposure and institutionally credible cooperation. If that distinction is taken seriously, then the core challenge of LLM-era data governance is no longer whether to share at all, but how to construct conditions under which limited, protected, and reciprocal sharing can emerge as a stable and normatively defensible equilibrium.

12 Acknowledgments

This article was prepared with substantial assistance from generative artificial intelligence, specifically ChatGPT by OpenAI. The system was used as an auxiliary scholarly-writing tool for iterative drafting, conceptual elaboration, restructuring of arguments, translation into academic English, stylistic refinement, and preparation of LaTeX-formatted manuscript text. This use is disclosed explicitly in the interest of transparency.

The manuscript follows an ICMJE-inspired distinction between tool use and authorship. The AI system is not listed as an author because it cannot assume responsibility for the integrity, originality, source verification, or scholarly accountability of the work. All substantive interpretive decisions, the final argumentative structure, the selection and checking of references, and responsibility for the submitted text remain exclusively with the human author.

References

Carlini, N., F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel (2021). "Extracting Training Data from Large Language Models". In: *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, pp. 2633–2650. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Cheng, Y., Y. Liu, T. Chen, and Q. Yang (2020). "Federated Learning for Privacy-Preserving AI". In: *Communications of the ACM* 63.12, pp. 33–36. DOI: 10.1145/3387107.

Curtis, C., N. Gillespie, and S. Lockey (2023). "AI-Deploying Organizations Are Key to Addressing 'Perfect Storm' of AI Risks". In: *AI and Ethics* 3, pp. 145–153. DOI: 10.1007/s43681-022-00163-7.

European Commission (July 10, 2025). *The General-Purpose AI Code of Practice*. URL: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (visited on 03/18/2026).

– (2026). *What Data Can We Process and under Which Conditions?* URL: https://commission.europa.eu/law/law-topic/data-protection/rules-business-and-organisations/principles-gdpr/overview-principles/what-data-can-we-process-and-under-which-conditions_en (visited on 03/18/2026).

European Parliament and Council of the European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*. Official Journal of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

European Parliament and Council of the European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. URL: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32024R1689.

Kaushik, A., C. Barcellona, N. K. Mandyam, S. Y. Tan, and J. Tromp (2025). "Challenges and Opportunities for Data Sharing Related to Artificial Intelligence Tools in Health Care in Low- and Middle-Income Countries: Systematic Review and Case Study from Thailand". In: *Journal of Medical Internet Research* 27, e58338. DOI: 10.2196/58338.

Pasetti, M., J. W. Santos, N. K. Corrêa, N. de Oliveira, and C. P. Barbosa (2025). "Technical, Legal, and Ethical Challenges of Generative Artificial Intelligence: An Analysis of the Governance of Training Data and Copyrights". In: *Discover Artificial Intelligence* 5, p. 193. DOI: 10.1007/s44163-025-00379-6.

Prisoner's Dilemma and Cooperation (2021). Springer Nature Reference. DOI: 10.1007/978-3-319-19650-3_3757.

Reinhardt, K. (2023). "Trust and Trustworthiness in AI Ethics". In: *AI and Ethics* 3, pp. 735–744. DOI: 10.1007/s43681-022-00200-5.

Richards, D. (2001). "Reciprocity and Shared Knowledge Structures in the Evolution of Cooperation". In: *Journal of Conflict Resolution* 45.5, pp. 621–635. DOI: 10.1177/0022002701045005004.

Rigaki, M. and S. Garcia (2023). "A Survey of Privacy Attacks in Machine Learning". In: *ACM Computing Surveys* 56.4, p. 101. DOI: 10.1145/3624010.

Schilke, O., M. Reimann, and K. S. Cook (2025). "The Transparency Dilemma: How AI Disclosure Erodes Trust". In: *Organizational Behavior and Human Decision Processes* 190, p. 104405. DOI: 10.1016/j.obhdp.2025.104405.

Schultz, M. D. and P. Seele (2023). "Towards AI Ethics' Institutionalization: Knowledge Bridges from Business Ethics to Advance Organizational AI Ethics". In: *AI and Ethics* 3, pp. 99–111. DOI: 10.1007/s43681-022-00150-y.

Wilke, F. (2025). "The Impact of Trust and the Role of the Opt-Out Mechanism in Willingness to Share Health Data via Electronic Health Records in Germany: Telephone Survey Study". In: *JMIR Human Factors* 12, e65718. DOI: 10.2196/65718.

Wornow, M., Y. Xu, R. Thapa, B. Patel, S. Fleming, M. A. Pfeffer, J. Fries, and N. H. Shah (2023). "The Shaky Foundations of Large Language Models and Foundation Models for Electronic Health Records". In: *npj Digital Medicine* 6, p. 135. DOI: 10.1038/s41746-023-00879-8.

Yin, X., Y. Zhu, and J. Hu (2021). "A Comprehensive Survey of Privacy-Preserving Federated Learning: A Taxonomy, Review, and Future Directions". In: *ACM Computing Surveys* 54.6, p. 131. DOI: 10.1145/3460427.

Table 1: Illustrative mapping from strategic parameters to concrete institutional mechanisms

Parameter	Strategic meaning	Illustrative legal, technical, or organizational mechanisms
C	Effective cost of sharing, including privacy exposure, compliance burden, and operational vulnerability	Federated learning, secure enclaves, differential privacy, secure multi-party computation, purpose limitation, standardized compliance review, internal data classification, constrained access workflows
A	Appropriative advantage from unilateral withholding or asymmetric capture of value	Reciprocal access rules, contribution requirements for consortium membership, contractual prohibitions on secondary use, licensing restrictions, benefit-sharing arrangements, liability clauses
R	Risk-reduction effect produced by safeguards	Encryption, auditable logs, role-based access control, authentication layers, sandboxed environments, traceability mechanisms, external technical audits
T	Institutional trust premium generated by repeated, verifiable cooperation	Repeated interaction, third-party oversight, sanctions for opportunism, reputation systems, compliance certification, dispute resolution procedures, governance boards
G	Collective innovation gain from protected cooperation	Cross-institution benchmarking, federated analytics, pooled validation, interoperability standards, shared research infrastructure, coordinated model improvement
δ	Systemic loss from mutual withholding and epistemic fragmentation	Duplicated effort, slower model improvement, fragmented case bases, repeated reinvention of solutions, lower cross-organizational learning capacity

Table 2: Baseline parameters for the evolutionary population simulation

Parameter	Baseline value/distribution	Interpretation
N	200	Number of organizations
Epochs	150	Evolutionary time horizon
Runs	50	Monte Carlo repetitions
η	0.35	Selection strength in strategy updating
μ	0.01	Mutation / exploration rate
α	0.25	Belief updating rate
τ_+	0.04	Trust increment after mutual sharing
τ_-	0.07	Trust decrement after exploitation
B_i	$\mathcal{N}(2.0, 0.2)$	Baseline private AI benefit
C_i	$\mathcal{N}(1.4, 0.3)$	Sharing cost
G_i	$\mathcal{N}(1.2, 0.2)$	Collective innovation gain
A_i	$\mathcal{N}(1.0, 0.2)$	Appropriation advantage
δ_i	$\mathcal{N}(0.5, 0.1)$	Systemic loss from mutual withholding