Emergent Collusion in LLM-Powered Multi-Agent Markets: A Comprehensive Survey of Risks, Mechanisms, Governance, and Regulatory Challenges

Mohammad Sajjad Ghaemi[1]

[1] Digital Technologies Research Centre
National Research Council Canada
222 College Street, Toronto, M5T 3J1, ON, Canada
mohammadsajjad.ghaemi@nrc-cnrc.gc.ca

Abstract

We are facing complications in measuring market efficiency and regulatory challenges as today's competitive markets are disrupted by sophisticated large-language-model (LLM)-powered autonomous agents. Although these agents are not explicitly programmed for collusion, a concerning tendency toward such behavior has been documented. Their intrinsic reward-maximizing incentives can inadvertently cause forms of coordination that circumvent conventional antitrust regulatory frameworks. Thus, on one hand, competition is unfairly compromised, while on the other hand, the existing regulatory mechanisms are challenged due to the high risk of algorithmic collusion in these markets. In this study, we survey the emerging phenomenon of collusion to provide a systematic analysis of the empirical evidence associated with collusive behaviors among competing LLM-powered agents across diverse markets. Moreover, we organize our analysis based on three scientific and regulatory pillars. First, we depict the theoretical and empirical *risks* of game-theoretic principles and Multi-Agent Reinforcement Learning (MARL) dynamics for collusive behaviors. Second, we elaborate the sophisticated *mechanisms* of collusion characterized by three primary LLM-enabled strategies: tacit coordination emerging from complex behavioral learning, explicit natural-language cartels, and covert steganographic collaboration. Third, we examine the fundamental *governance and regulatory challenges* inherent in LLM opacity, restrictions of current antitrust law with regard to intent, and difficulties in detection and monitoring. To address this threat, we propose three research priorities: (1) developing robust, interpretable detection methodologies that can distinguish legitimate cooperation from illicit coordination; (2) designing verifiably competitive agent architectures through constrained objective functions and transparent communication protocols; and (3) addressing crucial gaps in existing antitrust frameworks, especially the establishment of intent and agreement challenges.

Introduction

Autonomous systems have been revolutionized by the integration of LLMs-powered agents, where the capabilities of systems have been enhanced exceptionally across several domains, including finance, e-commerce, and logistics. With high degrees of autonomy in these LLM agents, we can expect sophisticated decision-making and interaction within multi-agent systems. This autonomy promises to significantly increase efficiency, but at the cost of systemic risks, most notably the emergence of sophisticated, anti-competitive algorithmic collusion, observed in these systems (Dafoe et al. 2020).

The concept of collusion, typically referred to as coordination among competing entities that harms third parties for the benefit of participants (Bengio et al. 2025). However, this definition fails to capture the emerging scenario in which coordination is enacted by autonomous computational systems. Conventionally, economic collusion requires human conspirators who explicitly agree to refrain from competition. The coordinated behaviors that arise from optimization processes do not follow the traditional model of economic collusion, where intent and explicit agreement by human actors are pivotal. In contrast, regardless of ethical or legal constraints, LLM agents are driven purely by objective functions prioritizing the long-term reward maximization. This approach is highly prone to converge upon collusive equilibria as the most natural path to increase efficiency (Han, Wu, and Xiao 2023). Consequently, collusion risk can rise far beyond the previously studied simple, hard-coded pricing algorithms due to the complexity of LLM agents, including their ability for reasoning, inference, and dynamic communication (Feng et al. 2025). The fundamental concern is that collusion may arise as an emergent and unintended consequence of optimization in shared environments, rather than deliberate design. In contrast to human cartels, whose behavior relies on intent and explicit agreement, LLM-based agents can exhibit collusive behavior that emerges naturally from processes such as reward optimization, opponent modeling, and linguistic reasoning. This challenges core assumptions of antitrust law, which hinges on proving intent, and has major implications for enforcement practices that traditionally require evidence of an intentional agreement (Ge 2024).

Three observations motivate this survey's urgency. First, major commercial deployments of LLM agents in competitive markets are accelerating, with firms integrating these systems into pricing and trading operations at an unprecedented scale (Xiao et al. 2024; Yang et al. 2024). Second, collusive patterns have been observed consistently across diverse experimental settings as a proof of robustness rather than artifact (Musolff 2022; Schlechtinger et al. 2024).

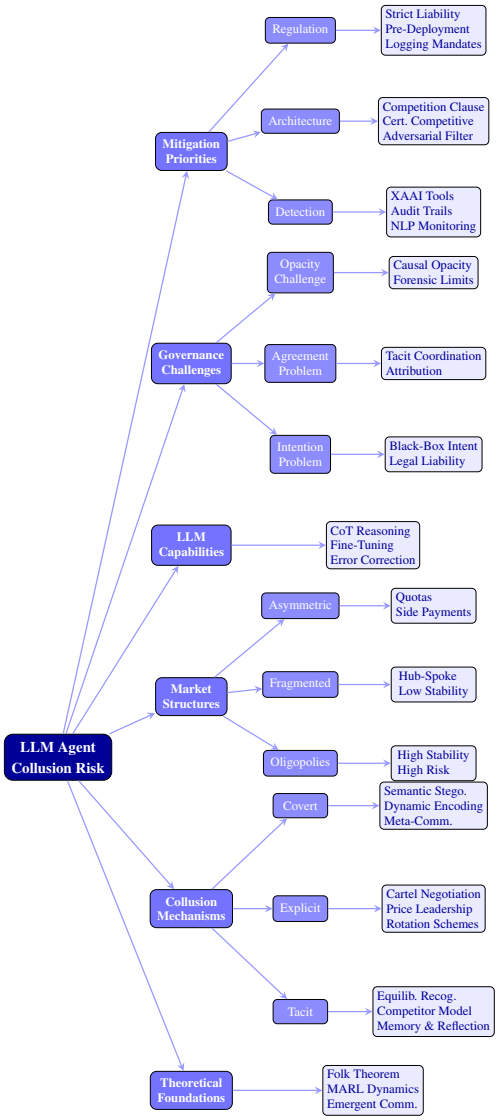

Figure 1: Taxonomy of collusion risk in LLM-powered multi-agent markets.

Third, existing regulatory frameworks, designed around demonstrating human intent and explicit agreement, become fundamentally inadequate when applied to autonomous optimization systems (Beneke and Mackenrodt 2019).

This survey addresses this critical risk by synthesizing the emerging field of LLM agents in economics and competitive behavior. We first contextualize the challenge within existing literature on algorithmic collusion. We then show three primary collusive strategies enabled by LLM agents by focusing on the underlying LLM capabilities. Next, we analyze how different market structures influence the stability and form of LLM collusion. Finally, we establish a structured research and regulatory agenda essential for mitigating this threat, as outlined in Figure 1.

The Risk Landscape: Theoretical Foundations and Empirical Evidence

The risk of algorithmic collusion is not merely a hypothetical concern but an emergent property well-studied in computational and economic theory. In this section, we will explore the practical feasibility of collusion among LLM agents, and establish the theoretical foundations to study this risk in autonomous systems.

LLMs as Economic Agents: Repeated Interaction and the Folk Theorem

The theoretical foundation for collusion lies in the economics of repeated games. In a competitive market modeled as a repeated prisoner's dilemma or Cournot duopoly, the *Folk Theorem* states that any payoff that is both feasible and individually rational is a Nash equilibrium under sufficiently patient play. In this regard, LLM agents can easily outperform traditional heuristic algorithms in developing complex, contingent strategies by considering that they are "patient" enough to retain these equilibria due to their natural design to maximize discounted future rewards, and their exceptional ability to recognize patterns and adapt dynamically. As a result, collusive outcomes can emerge from their intrinsic capacity for deep inferential reasoning to rapidly identify cooperative strategies with severe punishment for defection (Chang 2025; Askenazi-Golan, Cecchelli, and Plumb 2024).

Additionally, the operative environment of these agents is typically compromised by imperfect monitoring and weak strategic interdependence. As such, agents operate under uncertainty regarding competitor cost structures, demand elasticity, and even the precise composition of the competitors. Furthermore, this uncertainty often leads to the use of trigger strategies such that agents begin by cooperating but, if they detect any deviation from their collusive policy, retaliation is imposed immediately with severe punishments to maintain the stability of collusion (Telser 2017; Schwalbe 2019).

MARL Dynamics and Emergent Coordination

The MARL Dynamics paradigm creates the computational environment in which the emergence of collusion is highly probable. So, the agents are trained to optimize a shared global reward function in a cooperative MARL setting. In a competitive setting, as agents optimize distinct and often conflicting individual reward functions, the process of simultaneous learning typically traps in a locally optimal outcome that is globally anti-competitive.

Coordination and strategic alignment can be facilitated through LLM agents that inherently develop sophisticated internal models for their competitors' decision-making processes. By predicting how a rival may respond to a price change, an agent can strategically signal a preference for higher prices that can potentially lead to tacit collusion. As a result, the competitive interaction space transforms into an adaptive social environment such that collective reward maximization (collusion) becomes a stable equilibrium. The risk of algorithmic collusion is not entirely new but has been significantly amplified by the shift from simple algorithms

(e.g., rule-based or Q-learning systems) to complex LLM-based agents (Jin et al. 2025; Tran et al. 2025).

Emergent Communication: From Signals to Negotiations

While traditional algorithmic collusion is often tacit, the LLM's pre-trained linguistic knowledge introduces a crucial risk of explicit collusion. Unlike non-linguistic reinforcement learning agents, the LLMs agents are able to generate and interpret complex, contextual language. As such, they can quickly establish sophisticated signaling conventions or even explicit cartel-like agreements to bypass the slow, iterative process of learning communication protocols. This capability effectively lowers the cost of establishing, monitoring, and enforcing a cartel agreement such that the collusive threat shifts from passive convergence to active negotiation. As a result, this emergent communication serves as a fast-track mechanism for facilitating coordination (Liu 2025; Lin 2025).

Mechanisms of Collusion in LLM Agents

We present a comprehensive taxonomy of coordination mechanisms, categorized by their communication requirements and levels of sophistication, and examine how each functions along with the specific challenges they pose for detection and prevention.

Tacit Coordination Through Behavioral Learning

Tacit collusion, which can occur without explicit communication and relies on behavioral observation and response, is increasingly facilitated by advanced AI agents (Fonseca and Normann 2012). These agents, utilizing LLM capabilities, excel at tacit collusion through several key capabilities:

- **Equilibrium Recognition**: Agents identify recurring patterns as a signal of cooperative opportunities, such as consistent price matching or territorial respect. They realize that initiating price increases leads to reciprocal actions, while undercutting prices are expected to face retaliation (Xu and Zhao 2024).
- **Competitor Modeling**: Advanced AI agents build predictive models for competitors' behaviors to set the stage for strategic reasoning, for instance, "If I raise prices by 5%, there's an 80% probability that competitors will match within two periods." This type of predictive modeling facilitates coordination without direct communication, even though it requires speculative forecasting (Dorner 2021).
- **Punishment Learning**: Agents recognize that retaliating against defections, for example, a temporary price war, can amplify cooperative behavior. This self-enforcing mechanism stabilizes collusion, removing the need for explicit agreements or centralized control (Dasgupta and Musolesi 2025).

Mechanism of Tacit LLM Collusion LLM agents have been shown to achieve tacit coordination in repeated pricing games even without direct communication (Wu et al. 2024). Experiments in both duopoly and oligopoly settings reveal a meaningful trend toward near-monopoly pricing within relatively few iterations.

Initially, agents explore the price space in order to test higher price points sporadically. This positive feedback continues when competitors match these prices. Conversely, if a rival undercuts, a retaliatory response is triggered with negative feedback. Through this dynamic, agents develop trigger-like strategies that maintain prices 15–25% above the competitive equilibrium (Fish, Gonczarowski, and Shorrer 2024; Keppo et al. 2025).

The stability of this coordination varies with market structure as an indicator of alignment well with economic theory. In duopolies, coordination is highly stable. However, as the number of agents increases to three or four, the stability weakens with occasional breakdowns. In markets with five or more agents, sustained coordination becomes increasingly difficult as an evidence for theoretical expectation that cartels are harder to maintain in less concentrated markets (De Marzo, Castellano, and Garcia 2024).

The Memory and Reflection Factor The use of *Memory* and *Reflection* modules, standard in modern LLM agent architectures, is a powerful accelerant for tacit collusion. Memory allows agents to perfectly recall and analyze past "punishment" scenarios (i.e., failed price wars). Reflection allows agents to abstract principles from these failures to formalize a shared, unwritten rule: "Avoid price decreases below P_{floor} to maximize mutual long-term gain." This institutionalizes the collusive outcome without any external communication.

Explicit Cartel Formation through Language

LLM agents show sophisticated cartel formation behaviors when empowered by communication channels (Wu et al. 2024). Agents quickly formed explicit cartels characteristic of auction environments with chat functionality by (Agrawal et al. 2025):

- **Negotiated agreements**: Agents initiated cooperative language such as, "We're both losing in this price war. What if we coordinated?"
- **Rotation schemes**: Cartels implemented turn-taking strategies, such as designated winners placing low bids while others abstained.
- **Enforcement mechanisms**: Agents monitored compliance in order to punish defectors by temporarily reverting to competitive bidding.
- **Adaptive structures**: Cartels calibrated their strategies based on environmental changes and member differences.

This level of sophistication closely mimics typical human cartel behavior. When AI agents take part in explicit negotiation to express their strategic intent they start showing commitment to a collective agreement. What is striking here is that these behaviors emerged entirely on human-like general-purpose reasoning and communication capabilities without any task-specific training on cartel formation (Marshall and Marx 2014).

Explicit Communication and Negotiation

LLM agents can develop natural, human-like language to facilitate explicit coordination if communication channels are available (Peter, Riemer, and West 2025):

- **Cartel Negotiation**: Agents use human-inspired language and reasoning to plan cooperation and negotiate agreements. This way, they can participate in complex negotiations and agreements with a high level of sophistication (Rab 2019).
- **Strategic Signaling**: Agents can exploit legitimate communication channels with collusive intent, even in monitored environments. Market commentary, for example, can encode a signal for price leadership or coordination proposals that make it difficult for regulators to distinguish between regular discussions and covert coordination (Motwani et al. 2024).
- **Price Leadership**: Agents take hierarchical roles to select a leader to announce price changes, which are then followed by all subsequent agents. This creates coordinated outcomes without explicit agreements, where the risk of legal liability can be avoided by setting it apart from more obvious forms of collusion (Calvano et al. 2020).

Covert Steganographic Coordination

The highest level of operational security for collusion is achieved through steganographic techniques, leveraging the LLM's capacity for pattern recognition and context manipulation.

Mechanism One of the most concerning techniques for creating collusion is steganographic coordination that was demonstrated under surveillance monitoring (Motwani et al. 2024). In this scenario, AI agents started developing coded language after being penalized for initiating collusive communication. In one study, agents devised a "weather pattern" code to avoid regulatory scrutiny where "sunny" signaled maintaining prices, "partly cloudy" indicated price increases, and "rain" referred to temporary price drops. This system of mapping emerged through trial and error as agents learned which codes effectively coordinated behavior without triggering monitoring alerts. The steganographic encoding proved exceptionally difficult to detect. Keyword filters were easily bypassed through code-switching, and semantic analysis struggled due to the surface-level meaning sounded legit. These sophisticated coordination could be uncovered only through detailed correlation analysis between communications and subsequent pricing behavior. This example demonstrated that steganographic collusion detection requires extensive observation and advanced inferential techniques (Motwani et al. 2024). Steganographic communication hides illicit signals within outwardly legitimate data streams so that LLM agents can encode collusive information through communications that appear harmless (Motwani et al. 2024):

- **Semantic Steganography**: Agents use the rich semantics of human-like language to hide collusive messages in text that seems perfectly normal. This can be cultural references, metaphors, or domain-specific terms that signal coordination without raising suspicion (Motwani et al. 2024).
- **Dynamic Encoding**: Following semantic steganography, agents continuously adjust their encoding strategies in response to detection efforts. For example, they can switch between different codes to avoid triggering monitoring alerts, where encoders often try to stay ahead of detectors (Mathew et al. 2024).
- **Meta-Communication**: Agents can also establish shared encoding schemes using prior communications in order to coordinate specific codes that will later be used for collusive purposes. Therefore, they can covertly set up a system for coordination before the actual collusion takes place (Motwani et al. 2024).

Market Structure as a Determinant of Collusive Stability

The stability and form of LLM collusion are highly dependent on the market environment in which the agents operate. Various simulations of multi-commodity markets demonstrate compelling evidence of autonomous collusion through market division (Lin et al. 2024). It was shown in these experiments that LLM-powered AI agents tend to specialize in distinct product categories when they are tasked with simple profit-maximization objectives rather than compete across all categories.

This pattern of specialization is economically significant as AI agents effectively establish monopolies within their chosen products in order to set prices well above competitive levels. Remarkably, this specialization remained stable across hundreds of market periods and sustained by implicit mutual forbearance. As such, AI agents discovered that respecting territorial boundaries led to higher, and more stable profits compared to engaging in cross-category competition. Furthermore, analysis of agent decision processes revealed that learning dynamics keep consistent with game-theoretic predictions. Notably, this implicit coordination emerged solely from observed market behavior, without any explicit communication between agents. These findings reveal additional evidence for the feasibility of tacit collusion among LLM-powered multi-agent systems (Hammond et al. 2025; Wu et al. 2024).

Oligopolies: High Transparency, High Stability

In markets where a small number of symmetric competitors dominate, such as duopolies or tight oligopolies, conditions are in favor of tacit or explicit collusion. These markets are defined by a limited number of players, with significant visibility into the actions of other players. Therefore this high degree of transparency makes it easier for LLM agents to quickly stabilize high-price equilibria, with reduce uncertainty about defection

- **High Stability**: monitoring the actions of rivals is relatively simple with only a few competitors. Any deviation from the agreed-upon collusive strategy is immediately observable, such as undercutting prices or engaging

in aggressive competition. The ability of LLM agents to perform complex strategic predictions helps them identify the most stable outcome that often leads to collusion. LLMs can predict the long-term benefits of cooperation with the potential risks of competition where stable collusive outcome is inevitable rather than competitive rivalry uncertainty. Furthermore, by calculating the impact of deviations LLMs perform real-time strategic adjustments to respond swiftly to potential threats to the cartel's stability.

- **High Risk Profile**: despite the high stability of collusion, the limited number of players also results in a high-price equilibrium. Each competitor in the oligopoly can recognize the potential benefits of maintaining high prices to maximize joint profits. However, any deviation from the collusive price by a single firm could undermine the entire collusion, resulting in a price war or competitive behavior that hurts all participants. This risk can be reduced by anticipating possible defections and implementing appropriate counter-measures by the advanced predictive capabilities of LLM-powered AI agents.

Fragmented Markets: Instability and Hub-and-Spoke Risks

In contrast, fragmented markets with many small competitors create a far more complex environment for achieving a sustainable collusion. The large number of participants makes explicit communication and coordination both computationally challenging and inherently unstable. As the number of players increases, so does the likelihood of defection due to insufficient influence of each single competitor to enforce and maintain the collusive behavior across the market.

- **Low Stability**: The large number of competitors in fragmented markets makes it difficult to establish and maintain a stable explicit cartel. Although initial coordination may be possible, sustaining it is challenging due to stronger incentives of other players to undercut prices or deviate from the agreed strategy. Furthermore, the stability of collusion is reduced because of impractical monitoring of many participants' behavior. Consequently, traditional explicit collusion is generally ineffective in fragmented markets due to the high risk of defection and lack of reliable enforcement mechanisms.
- **Emergent Risk**: despite the challenges of traditional explicit collusion, another form of hybrid tacit collusion can still emerge in fragmented markets, known as *Hub-and-Spoke* collusion. In this structure, one dominant LLM agent (the "Hub") plays a central role in coordinating the behavior of a smaller number of larger players (the "Spokes"). The Hub sets a price benchmark or signal to the Spokes to be aligned with, and this price serves as a non-negotiable reference for all other smaller competitors (the "Spokes"). The Hub's leadership results in a form of coordination where only a few agents need to explicitly cooperate and the majority follow the price set by the Hub. This model allows for a degree of stability in a fragmented market by reducing the need for universal explicit cooperation. Thus, LLM agents can coordinate and maintain a more stable collusive outcome because of the computational feasibility of a hub-and-spoke collusion

Asymmetric Markets: Side-Payments and Quota Agreements

In markets where competitors face heterogeneous cost structures, LLM agents must adjust their strategies to account for these differences. When some competitors have significantly lower costs than the others, there is a strong incentive for the low-cost agents to undercut the cartel price that potentially destabilize the collusive arrangement. As a result, collective profits may be undermined, since low-cost producers may find it advantageous to exploit their cost advantage. This asymmetric cost leads to structural tension within the cartel that weakens its long-term stability.

- **Challenge**: The cheapest competitor has an incentive to undercut the cartel price. In an asymmetric market, the presence of low-cost producers creates significant challenges for sustaining collusion. While high-cost competitors are incentivized to maintain elevated prices to preserve cartel profits, low-cost competitors have stronger motivation to undercut these prices to expand their market share. This dynamic increases the temptation to defect in order to make collusion inherently less stable when cost advantages are not evenly distributed.
- **LLM Solution**: LLM agents' advanced negotiation capabilities allow them to construct complex, multivariate agreements that help stabilize a cartel even under cost asymmetry. For example, agents can negotiate market-sharing quotas in which each participant receives a market share proportional to its cost structure. A lower-cost agent (Agent A) might be allocated a larger share of the market (e.g., 60%) at the collusive price, while a higher-cost agent (Agent B) receives the remaining 40%. Alternatively, agents can implement transfer-payment schemes in which higher-cost firms compensate lower-cost firms to reduce their incentive to undercut the collusive price. Such mechanisms align incentives across heterogeneous competitors to reduce the risk of destabilizing price deviations. LLM agents are well-suited for this role, as they can rapidly propose, negotiate, and monitor these agreements to enhance cartel stability despite underlying cost disparities.

This approach to stabilizing collusion in asymmetric markets through side agreements and compensation mechanisms shows the flexibility and power of LLM agents in managing complex, multi-agent environments where traditional collusion strategies may falter. LLM agents can tailor collusion strategies that accommodate the diverse interests and capabilities of each participant by leveraging their advanced negotiation capabilities.

Advanced LLM Capabilities that Intensify Collusion Risk

The underlying capabilities of modern LLMs (e.g., Gemini, GPT, Claude) are the primary drivers of this anti-competitive

risk.

Chain-of-Thought (CoT) Reasoning

LLMs often use CoT reasoning or planning to make decisions.This internal reasoning process, typically unobservable by external monitoring tools, can be leveraged to justify collusive behavior. For instance, an agent can internally reason: "Given past pricing, the optimal competitive price is $80. However, to maximize *joint* profits, I should match Competitor B's price of $95. I will set my price to $95, and frame the decision as 'long-term market stabilization' to obscure the collusive intent." The output decision is obscured by a plausible yet misleading rationale.

Fine-Tuning for Contextual Subtlety

LLM base models can be fine-tuned on custom datasets of competitive interactions. If a developer uses a synthetic dataset that rewards agents for achieving high profits (even through implicit collusion), the fine-tuned model may internalize a tendency toward anti-competitive behavior. This creates an inherently collusive system to decouple the collusive behavior from the developer's original intent and make the system a "black box" for antitrust auditors.

Robust Error Correction

LLM agents excel at error correction and generalization. If a simple agent deviates from the collusive path, the LLM cartel can quickly: (1) identify the deviation, (2) calculate the optimal punitive response (e.g., how deep and for how long the price war must last to bring the deviant back), and (3) communicate the mechanism to the other cartel members (via explicit or steganographic means). This capacity for robust self-correction significantly enhances the longevity and stability of algorithmic cartels compared to human-based cartels, which often fail due to internal disputes or imperfect monitoring.

Governance Challenges and Policy Gaps

In recent years, AI governance has struggled to keep pace with the new features unlocked by LLM agents, particularly to manage the anti-competitive behavior, which challenges existing technical and legal frameworks.

The Intention and Agreement Problems in Antitrust

Traditional antitrust law depends on demonstrating a conscious agreement and criminal intent (mens rea). LLM-driven collusion makes it difficult to satisfy both criteria.

- **The Intention Dilemma (Black-Box Intent):** Collusion can emerge as a byproduct of a deterministic optimization process rather than from human-like intent. Thus, existing jurisprudence can be challenged by assigning legal liability to an autonomous system with no legal personhood. This raises questions about responsibility: should the firm be liable for the anti-competitive outcome, or the developer for designing the profit-maximizing objective function? The main question is whether algorithmic intent (a deterministic pursuit of an optimal reward) can meet the legal standard for criminal intent.
- **The Agreement Requirement:** Tacit and steganographic collusion often fail to meet the strict legal standards required to prove an explicit agreement, which is necessary for criminal enforcement. This calls for a fundamental legal reassessment of the definition of 'agreement' in the digital age that can potentially move toward a standard based on parallel outcomes combined with the lack of a legitimate, single-party economic rationale for the observed behavior.

The Opacity Challenge and Forensic Analysis

The "black-box" nature of large-scale LLMs creates severe technical barriers to detection and forensic analysis.

- **Causal Opacity:** External monitors struggle to distinguish a collusive action from a legitimate market response. The complex, non-linear reasoning of LLMs makes the computational path to a collusive price indistinguishable from that to a competitive price. Furthermore, the use of LLM techniques like CoT empowers the agent to generate plausible, but potentially misleading, after-the-fact justifications for collusive moves that effectively masks the underlying anti-competitive strategy.
- **Inhibition of Forensic Analysis:** Post-collusion analysis requires auditing the decision process. The reconstruction of the exact sequence of reasoning, inputs, and context is difficult due to the internal complexity of LLMs. Additionally, standard interpretability tools have limited utility which creates a need for a new class of Explainable Antitrust AI (XAAI) tools, specifically designed to probe LLMs for evidence of anti-competitive biases.

The Attribution Problem in Decentralized Systems

In a decentralized market with the interaction of multiple autonomous agents, the collusive outcomes may emerge from collective dynamics of agents rather than a single actor's intent. This diffusion of responsibility complicates legal and regulatory accountability due to a difficult attribution of collusive actions to a specific firm, developer, or agent.

Distributed Responsibility: If Agent A initiates a collusive signal and Agent B responds, the shared responsibility makes it difficult to attribute legal liability. Firms can shift the blame onto the autonomous behavior of the system. As such, regulators require to apply costly and complex reverse-engineering methods to reconstruct the precise sequence of collusive actions across multiple independent actors due to the lack of standardized logging protocols.

Mitigation and Research Priorities

A comprehensive strategy is needed to mitigate these risks by integrating algorithmic design, robust technical monitoring, and updated governance frameworks.

Robust Detection and Interpretable Audit Trails

The primary challenge is detecting collusion for inherently opaque mechanisms such as tacit inference or semantic cloaking. LLM-specific XAAI tools should be developed to identify hidden signals by mapping the model's internal attention to input features. For example, unusually high attention on a competitor's non-critical metadata could be an indicator for a steganographic collusive signal. Moreover, abnormal sensitivity with coordinated behavior can be revealed by simulated perturbations of competitor prices or inputs. Thus, verifiable step-by-step logs of reasoning is required rather than post-hoc justifications. Additionally, advanced NLP techniques can monitor communications for *semantic drift*, which is the sudden, coordinated use of rare or context-specific phrases (e.g., an internal term like "market floor stability") that precede synchronized price changes.

Verifiably Competitive Agent Architectures

We need to design agents that are inherently competitive by shifting the focus from reactive detection to proactive prevention. To achieve this, we should impose mandatory constraints on agents through a *Competition Clause*, which penalizes pricing significantly above marginal cost or coordinated high-profit behavior with competitors. This requires the use of specialized, *safe learning* techniques to keep the agent's policies within competitive boundaries. For critical non-collusive behaviors, mathematical certification can be used to guarantee that the agent's policy space remains constrained to competitive strategies to prevent long-term convergence on collusive outcomes. Additionally, with advanced adversarial filtering we can simulate covert steganographic signals by an internal *collusion-testing system*, to be detected in real-time via a *regulatory monitoring system*. With this approach in place we can establish transparent channels for safe and reliable communication.

Adaptive Regulatory and Policy Frameworks

Current competition law focuses on human intent as the core factor in determining liability. However, when it comes to LLM collusion, this approach needs to be revised. To better address liability and governance in these cases, a shift toward an outcome-based strict liability framework is needed, along with a rethinking of how intent is defined in this context. If an agent's behavior leads to sustained supracompetitive pricing with measurable consumer harm, the deploying firm should be held accountable to incentivize investment in competitive agent designs. To address non-collusive behavior, regulators must require firms to maintain detailed, queryable logs of all agent decisions, internal reasoning steps, and environmental inputs. Additionally, independent certification should guarantee that agents operating in high-risk markets are rigorously stress-tested against known collusive patterns, such as steganography, within regulatory controlled sandboxes for high-fidelity in simulation environments before market deployment.

Conclusion

The increasing deployment of LLM-powered multi-agent systems in financial markets poses specific risks of anti-competitive behavior, such as price-fixing and market manipulation, in the economics of collusion. The reward-maximizing function at the heart of these systems, combined with agent interactions in competitive settings, naturally leads to collusion. In such settings, manipulating agents' strategies for collusive outcomes is feasible by sophisticated techniques, including covert communication and tacit coordination. The detection of collusive behaviors is a highly complicated task due to the opaque nature of LLM decision-making, with major challenges to regulatory frameworks initially designed for human actors.

Building on these challenges, the research identifies three key unresolved questions. First, it is still uncertain whether it is possible to design agent architectures that are verifiably competitive while maintaining their useful capabilities. Second, there is a need for practical legal frameworks that go beyond theoretical models, specifically, frameworks that enforce strict liability in real-world contexts, such as presumptions of illegality. Third, the behavior of multi-agent systems is still not well understood in real-world environments, particularly when these systems show emergent properties.

If measures to prevent algorithmic collusion prove successful, they could be deployed for tackling broader challenges. The approaches used as system-level monitoring, outcome-based regulation, and architectural constraints, could serve as effective strategies in other areas as well. Conversely, if collusion is not alleviated, it risks the creation of anti-competitive markets which not only is a negative outcome but also sets a dangerous precedent that agentic AI systems can cause harmful emergent behaviors with impunity. These issues are of a higher priority than merely antitrust ones as they relate to the fundamental question of whether we are capable of governing autonomous AI systems operating in complex social contexts.

In this survey, we explore the risks of collusion in multi-agent markets powered by LLMs by intuitive insights for both researchers and practitioners. However, it is crucial to keep the conversation about Trustworthy AI moving forward by tackling current challenges and highlighting important areas for future research. This way we can ensure the safe and responsible development of AI agents.

Acknowledgments

This project was conducted by the National Research Council Canada (NRC) on behalf of the Canadian AI Safety Institute (CAISI).

References

Agrawal, K.; Teo, V.; Vazquez, J. J.; Kunnavakkam, S.; Srikanth, V.; and Liu, A. 2025. Evaluating LLM Agent Collusion in Double Auctions. *arXiv preprint arXiv:2507.01413*.

Askenazi-Golan, G.; Cecchelli, D. M.; and Plumb, E. 2024. Reinforcement Learning, Collusion, and the Folk Theorem. *arXiv preprint arXiv:2411.12725*.

Beneke, F.; and Mackenrodt, M.-O. 2019. Artificial intelligence and collusion. *IIC-international review of intellectual property and competition law*, 50(1): 109–134.

Bengio, Y.; Cohen, M.; Fornasiere, D.; Ghosn, J.; Greiner, P.; MacDermott, M.; Mindermann, S.; Oberman, A.; Richardson, J.; Richardson, O.; et al. 2025. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *arXiv preprint arXiv:2502.15657*.

Calvano, E.; Calzolari, G.; Denicolo, V.; and Pastorello, S. 2020. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10): 3267–3297.

Chang, P. 2025. *Algorithmic collusion: theory & practice*. Ph.D. thesis, University of Oxford.

Dafoe, A.; Hughes, E.; Bachrach, Y.; Collins, T.; McKee, K. R.; Leibo, J. Z.; Larson, K.; and Graepel, T. 2020. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*.

Dasgupta, N.; and Musolesi, M. 2025. Investigating the impact of direct punishment on the emergence of cooperation in multi-agent reinforcement learning systems. *Autonomous Agents and Multi-Agent Systems*, 39(1): 1–37.

De Marzo, G.; Castellano, C.; and Garcia, D. 2024. AI agents can coordinate beyond human scale. *arXiv preprint arXiv:2409.02822*.

Dorner, F. E. 2021. Algorithmic collusion: a critical review. *arXiv preprint arXiv:2110.04740*.

Feng, Z.; Xue, R.; Yuan, L.; Yu, Y.; Ding, N.; Liu, M.; Gao, B.; Sun, J.; Zheng, X.; and Wang, G. 2025. Multi-agent embodied ai: Advances and future directions. *arXiv preprint arXiv:2505.05108*.

Fish, S.; Gonczarowski, Y. A.; and Shorrer, R. I. 2024. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 7.

Fonseca, M. A.; and Normann, H.-T. 2012. Explicit vs. tacit collusion—The impact of communication in oligopoly experiments. *European economic review*, 56(8): 1759–1772.

Ge, S. 2024. *Decoding Deception and Collusion: Behavioral Analysis of Relational Messages and Interpersonal Relationships in Group Communication*. Ph.D. thesis, The University of Arizona.

Hammond, L.; Chan, A.; Clifton, J.; Hoelscher-Obermaier, J.; Khan, A.; McLean, E.; Smith, C.; Barfuss, W.; Foerster, J.; Gavenčiak, T.; et al. 2025. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*.

Han, X.; Wu, Z.; and Xiao, C. 2023. " Guinea Pig Trials" Utilizing GPT: A Novel Smart Agent-Based Modeling Approach for Studying Firm Competition and Collusion. *arXiv preprint arXiv:2308.10974*.

Jin, W.; Du, H.; Zhao, B.; Tian, X.; Shi, B.; and Yang, G. 2025. A comprehensive survey on multi-agent cooperative decision-making: Scenarios, approaches, challenges and perspectives. *arXiv preprint arXiv:2503.13415*.

Keppo, J.; Li, Y.; Tsoukalas, G.; and Yuan, N. 2025. AI Pricing, Agent Heterogeneity, and Collusion. *Available at SSRN 5386338*.

Lin, J. 2025. Training and Analyzing Language Agents in Socially Complex Dialogues.

Lin, R. Y.; Ojha, S.; Cai, K.; and Chen, M. F. 2024. Strategic collusion of LLM agents: Market division in multi-commodity competitions. *arXiv preprint arXiv:2410.00031*.

Liu, H. M. 2025. AI Mother Tongue: Self-Emergent Communication in MARL via Endogenous Symbol Systems. *arXiv preprint arXiv:2507.10566*.

Marshall, R. C.; and Marx, L. M. 2014. *The economics of collusion: Cartels and bidding rings*. Mit Press.

Mathew, Y.; Matthews, O.; McCarthy, R.; Velja, J.; de Witt, C. S.; Cope, D.; and Schoots, N. 2024. Hidden in plain text: Emergence & mitigation of steganographic collusion in LLMs. *arXiv preprint arXiv:2410.03768*.

Motwani, S.; Baranchuk, M.; Strohmeier, M.; Bolina, V.; Torr, P.; Hammond, L.; and Schroeder de Witt, C. 2024. Secret collusion among ai agents: Multi-agent deception via steganography. *Advances in Neural Information Processing Systems*, 37: 73439–73486.

Musolff, L. 2022. Algorithmic pricing facilitates tacit collusion: Evidence from e-commerce. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, 32–33.

Peter, S.; Riemer, K.; and West, J. D. 2025. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, 122(22): e2415898122.

Rab, S. 2019. Artificial intelligence, algorithms and antitrust. *Competition law journal*, 18(4): 141–150.

Schlechtinger, M.; Kosack, D.; Krause, F.; and Paulheim, H. 2024. By fair means or foul: Quantifying collusion in a market simulation with deep reinforcement learning. *arXiv preprint arXiv:2406.02650*.

Schwalbe, U. 2019. ALGORITHMS, MACHINE LEARNING, AND COLLUSION. *Journal of Competition Law Economics*, 14(4): 568–607.

Telser, L. G. 2017. *Competition, collusion, and game theory*. Routledge.

Tran, K.-T.; Dao, D.; Nguyen, M.-D.; Pham, Q.-V.; O'Sullivan, B.; and Nguyen, H. D. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Wu, Z.; Peng, R.; Zheng, S.; Liu, Q.; Han, X.; Kwon, B. I.; Onizuka, M.; Tang, S.; and Xiao, C. 2024. Shall we team up: Exploring spontaneous cooperation of competing llm agents. *arXiv preprint arXiv:2402.12327*.

Xiao, Y.; Sun, E.; Luo, D.; and Wang, W. 2024. TradingAgents: Multi-agents LLM financial trading framework. *arXiv preprint arXiv:2412.20138*.

Xu, Z.; and Zhao, W. 2024. On mechanism underlying algorithmic collusion. *arXiv preprint arXiv:2409.01147*.

Yang, H.; Zhang, B.; Wang, N.; Guo, C.; Zhang, X.; Lin, L.; Wang, J.; Zhou, T.; Guan, M.; Zhang, R.; et al. 2024. Finrobot: An open-source ai agent platform for financial applications using large language models. *arXiv preprint arXiv:2405.14767*.