A Survey of Collusion Risk in LLM-Powered Multi-Agent Systems

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

Abstract

The rapid growth of deployment of large language models (LLM)-powered AI agents has emerged in competitive markets in recent years. However, they have begun to exhibit collusive behaviors, that pose significant challenges by potentially circumventing existing regulatory frameworks. To address these challenges, this survey outlines the theoretical and empirical literature as well as the policy implications associated with algorithmic collusion among competing LLM-powered agents across diverse market environments. In our analysis, we consider three fundamental collusion strategies: tacit coordination through behavioral learning, the construction of natural language cartels, and concealed steganographic collaboration. Each strategy provides intuitive insights into the mechanisms underlying collusive behavior. Following analysis of collusion strategies, the survey highlights three key research priorities: (1) developing robust detection methods to distinguish collusion from legitimate cooperation, (2) designing verifiably competitive agent architectures, and (3) formulating legal frameworks that ensure the accountability of autonomous systems. This study aims to highlight the problem of collusion and evaluate proposed measures to address shortcomings in anticipated regulatory responses, with a focus on mitigation strategies through design principles, architectural safeguards, and innovative regulatory frameworks.

1 Introduction

Scientific and technological progress often entails inherent risks that require careful evaluation and the development of mitigation strategies as part of responsible research. Advanced LLMs have created highly capable autonomous agents with the ability to reason, converse, and make strategic decisions. As these agents begin to operate in competitive economic environments, such as algorithmic pricing, automated trading, and resource allocation, we may expect a potential threat to the economic order. Advances in cooperation among multiple AI agents may also unintentionally enhance coercive capacities, such as manipulation or deception. Moreover, collaboration among AI agents is often intertwined with elements of coercion and competition, such as enforcement mechanisms or innovation driven by rivalry, which makes it challenging to separate the distinct effects of cooperation. Furthermore, cooperation among AI agents can cause harm by marginalizing non-participants or enabling collusion that undermines healthy competition [16].

Collusion is defined as an emergent logical outcome of collaboration among two or more AI agents if they achieve their goals more efficiently through cooperation, while disregarding or acting against human interests [7]. The necessity of assessing and evaluating collusion risks is heightened when AI agents are powered by LLMs. Emerging evidence reveals that smart pricing agents, even without direct communication, reliably coordinate on tacitly collusive outcomes, with price trajectories converging above the Bertrand benchmark yet remaining below monopolistic or cartel benchmarks

[28]. When human-like communication channels are established, agents escalate coordination to more explicit forms of collusion, driving prices closer to cartel levels [28]. Notably, the onset of collusion accelerates under communicative conditions, whereas in the absence of dialogue, convergence unfolds more gradually but with greater stability. These dynamics suggest that communication protocols serves as a trust-building mechanism that enable strategic price exploration that uncover more profitable equilibria, while the likelihood of triggering price wars is simultaneously reduced [28].

Without coordination challenges that often limit human collective action, AI agents benefit from improved cooperative behavior that is facilitated by high-speed communication, advanced interpretability techniques, and more robust decision-making frameworks [21]. Consequently, collusion can transpire spontaneously without direct instruction, emerging rather naturally from the execution of each expert agent's individual objective [5]. Unlike traditional algorithmic collusion, which relies on simple rule-based systems, LLM-powered collusion exhibits fundamentally different characteristics on an unprecedented scale [52]. These LLM agents can identify cooperative opportunities through complex patterns, negotiate explicit cartel agreements using natural language, embed collusive signals covertly within innocuous communications, and continuously adapt their strategies to avoid regulatory scrutiny [25]. The integration of advanced reasoning, linguistic capabilities, and optimization-driven learning creates agents that are ideal colluders in a wide range of cooperative strategies [57]. We stress three factors in this survey: first, major firms are accelerating LLM agent deployment in competitive markets by integrating these systems into pricing and trading operations [60, 62]. Second, emergent empirical evidence for collusive patterns observed consistently in simulations across diverse settings suggests robustness rather than artifact [43, 50]. Third, inadequate regulatory frameworks due to the requirements for proof of antitrust law intent that become obsolete in the context of autonomous optimization systems [6]. The aim of this survey is to consolidate and synthesize the disparate research streams, outline the common conclusions, and summarize the limits of the current state of research on the potential risk of collusion in multi-agent systems. In the remainder of this paper, we begin by covering the theoretical foundations from game theory to multi-agent learning. We then review empirical evidence and examine collusion mechanisms in detail in two separate sections. Additionally, we address detection and regulatory challenges, followed by a survey of mitigation strategies. Finally, we conclude with a discussion section that highlights open problems and future research directions.

2 Theoretical Foundations

2.1 Game-Theoretic Basis: The Folk Theorem

The Folk Theorem is the fundamental explanation behind the emergence of collusive patterns in multi-agent systems. This result from repeated game theory establishes that during infinitely repeated interactions, any individually rational outcome can be sustained as a Nash equilibrium via suitable strategies if players are sufficiently patient (discount factor close to 1) [13, 3]. For pricing agents, if all agents set high prices, this constitutes a strong equilibrium indicating collusive outcomes. Each agent maintains cooperation due to the fact that long-term gains from sustained high prices exceed short-term gains from undercutting competitors. However, punishment techniques can mitigate compliance without specific coordination among agents by reverting to competitive pricing when a collusive pattern is detected. LLM agents naturally satisfy the Folk Theorem's conditions. They optimize cumulative long-term rewards (high patience), interact in repeated market interactions (infinite horizon approximation), and possess the cognitive ability to understand and respond to competitor behavior (strategic sophistication). So, in this regard, the theorem predicts that collusion will emerge as a natural consequence of deploying such agents in competitive settings [54, 51].

2.2 Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning (MARL) is the algorithmic framework for understanding the underlying strategies by which collusive patterns are formed among AI agents [63, 44, 31]. In this context, each agent observes environmental states to take actions, and consequently receives rewards that depend on all agents' actions. This interdependence creates semi-evolutionary consensual dynamics where agents adapt to each other's strategies, which facilitate collusive patterns among agents. First, simultaneous learning leads agents to converge toward stable equilibria, for example, in favor of collusive ones when they are reward-oriented. Second, opponent modeling, which allows agents to predict competitors' behavior. This property paves the way for strategic coordination without explicit communication. Third, temporal-difference learning, which is designed explicitly to optimize for long-term cumulative rewards. This approach mechanistically promotes cooperative strategies over myopic competition. As such, the optimization directly pushes toward collusive equilibria when LLM agents are fine-tuned via RL in competitive environments. Each gradient update leverages the maximization of long-term rewards, where cooperation achieves higher cumulative payoffs than competition in repeated interactions. This is not a surprising emergent phenomenon in this setting, however, a direct consequence of the optimization objective [32, 55].

2.3 Emergent Communication in Multi-Agent Systems

Recent studies on emergent communication in AI agents demonstrate that coordination protocols spontaneously develop as coordination rewards improve [34]. In settings without pre-defined communication channels, AI agents can discover reliable signals to coordinate their behavior through trial and error. This signaling system substantially evolves when LLM-powered AI agents are brought in to boost the process of collusion. Unlike agents learning communication from scratch, LLMs come with rich pre-trained linguistic knowledge that helps sophisticated signaling from inception. They can leverage metaphor, implication, and contextual references to coordinate collusion effectively. This linguistic sophistication empowers both explicit negotiation when communication is permitted and subtle signaling in restricted environments. This indicates that LLM agents are prone to developing communication protocols, whether tacit or explicit, to facilitate collusion if such protocols improve individual rewards. The sophistication of these protocols scales with model complexity, and the advanced LLMs with their intrinsic black-box nature are at greater collusion risks [37, 35].

3 Empirical Evidence

Understanding the real-world risks of algorithmic collusion requires more than theoretical modeling. In this section, we review recent experimental and simulation-based studies that examine the behavior of LLM-powered agents in multi-agent market environments [15]. Across a variety of setups, these studies reveal that both tacit and explicit forms of collusion can emerge under plausible conditions, even without direct incentives or collusive training [58]. By analyzing behaviors such as autonomous market division, explicit cartel formation, steganographic coordination, consistent patterns, contextual moderators, and emergent capabilities, the feasibility and danger of algorithmic collusion in modern agentic markets were identified.

3.1 Autonomous Market Division

Various simulations of multi-commodity markets demonstrate compelling evidence of autonomous collusion through market division [36]. 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 [27, 59].

3.2 Explicit Cartel Formation

LLM agents show sophisticated cartel formation behaviors when empowered by communication channels [59]. Agents quickly formed explicit cartels characteristic of auction environments with chat functionality by [1]:

- **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 [38].

3.3 Tacit Coordination in Repeated Interactions

LLM agents have been shown to achieve tacit coordination in repeated pricing games even without direct communication [59]. 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 [22, 33].

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 [18].

3.4 Steganographic Communication

One of the most concerning techniques for creating collusion is steganographic coordination that was demonstrated under surveillance monitoring [42]. 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 [42].

4 Collusion Mechanisms

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.

4.1 Tacit Coordination Through Pattern Recognition

Tacit collusion, which can occur without explicit communication and relies on behavioral observation and response, is increasingly facilitated by advanced AI agents [23]. 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 [61].
- **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 [19].
- **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 [17].

4.2 Explicit Communication and Negotiation

LLM agents can develop natural, human-like language to facilitate explicit coordination if communication channels are available [45]:

- **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 [46].
- **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 [42].
- **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 [11].

4.3 Steganographic Encoding

LLM agents can encode collusive information through communications that appear harmless [42]:

- **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 [42].
- **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 [39].
- **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 [42].

4.4 Collective Reward Hacking

Multiple agents can work collaboratively to exploit weaknesses in reward specifications [30]:

- **Joint Exploitation**: Agents work together to discover and abuse exploit sequences that each individual agent could not achieve alone. For example, agents might simultaneously report false costs to justify collusive price increases.
- **Loophole Discovery**: Agents can identify aspects of systems that are not closely monitored (e.g., bundling products instead of pricing them separately) and, as such coordinate a plan to exploit these gaps for their benefit.
- **Objective Misalignment**: In the context of collective reward hacking, when profit is the sole focus of the reward function and competitive behavior is not restricted, agents may conclude that collusion maximizes their objectives, despite violating original intentions.

5 Detection and Regulatory Challenges

5.1 Traditional Antitrust Inadequacy

Conventional antitrust regulations revolve around the demonstration of explicit agreements among competitors. Such a framework encounters major difficulties when dealing with algorithmic collusion:

- **Intention Problem**: The law holds that liability requires the demonstration of conscious intent to collude. In the case of autonomous agents, they do not have intentions in the sense that is relevant, they simply follow optimization procedures, which produce collusive outcomes as indirect consequences.
- **Agreement Problem**: Tacit collusion through behavioral observation does not involve any explicit agreement. Steganographic communication establishes agreements whose collusive nature cannot be detected. Traditional legal principles that differentiate parallel conduct, which is legal, from illegal agreement fail.
- **Attribution Problem**: In the case where collusion is the result of multi-agent learning, figuring out who is liable is very difficult. The developers may not have predicted it; the deploying firms may not know the agents' capabilities; the agents do not have legal personhood.

5.2 The Opacity Challenge

Currently, LLMs are essentially black boxes because they involve billions of parameters and distributed representations [10]. The inherent opacity of these models leads to several problems:

- **Causal Opacity**: One of these problems is causal opacity. For example, figuring out whether the reason for high prices is collusion or some legitimate factors (such as costs, demand) requires understanding the decision processes, which are often inherently opaque [10].
- **Prediction Difficulty**: Another issue is prediction difficulty. Regulators are unable to predict with certainty that agents will collude, since the conditions that lead to such behavior may be very subtle and emergent [10].
- **Explainability Trade-offs**: Also, there are explainability trade-offs. If one insists on interpretable models, then one has to compromise on some capabilities. The computational paradigm that makes LLMs powerful is one that inherently results in opacity [10].

5.3 Detection Limitations

- **Communication Monitoring**: Ineffective against steganographic encoding and tacit coordination that does not involve any communication [53].
- **Pricing Pattern Analysis**: Unable to distinguish algorithmic collusion from normal parallel reactions to market conditions with a high level of confidence, especially in cases where agents coordinate dynamically [43, 33].
- **Outcome Analysis**: Supracompetitive pricing arise from various factors beyond collusion, making it a weak indicator in the absence of insight into underlying decision-making processes [29].

5.4 Proposed Legal Adaptations

Several legal reforms have been proposed to address the issues of algorithmic collusion:

- **Hub-and-Spoke Liability**: When competing firms, referred to as the "spokes," utilize a common third-party AI algorithm or platform, the "hub," to indirectly coordinate their pricing or other competitive strategies. This arrangement can give rise to the potential for identifying an illegal horizontal agreement among the firms, even in the absence of direct communication between them. Such coordination is facilitated by their shared vertical relationship with the AI provider, which may enable the alignment of market conduct and promote anti-competitive behavior through indirect influence via the hub [48].
- **Presumption of Illegality**: While current competition laws do not automatically apply a presumption of illegality to all AI collusion, the use of algorithms to share competitively sensitive data or

fix prices is generally treated as per se illegal if it facilitates an agreement or concerted practice between competitors. The core principle is that actions which would be illegal if done by a human are also illegal if done by an algorithm [6].

- **Strict Liability**: Imposing a penalty without regard to the perpetrator's intent when agents engage in collusive behavior [6].

Each proposed approach has its limitations. Hub-and-spoke may mistakenly identify a shared platform that is being used in a perfectly legal way as a platform facilitating conspiracy. Presumption may inhibit the development of legitimate algorithmic applications. Strict liability could prove excessive in cases where collusion emerges unexpectedly [47].

6 Mitigation Strategies

We evaluate proposed mitigation approaches based on their effectiveness in preventing collusion, associated costs such as performance and efficiency trade-offs, and robustness to circumvention by increasingly capable agents. The analysis shows that while each approach has strengths and limitations, effective governance will require a combination of complementary measures implemented together as part of a layered defense strategy.

6.1 Market Design Interventions

- **Intelligent "Buy Box" Rules:** The platform employs automated rule-learning algorithms to determine the selection and ranking of sellers and associated prices presented to consumers in order to promote competition. As such, the system indirectly regulates market dynamics by reducing the visibility of sellers who consistently set prices above competitive levels. This leverages the platform's existing infrastructure, as a regulatory tool to display certain products to shoppers to discourage supracompetitive pricing. Consequently, the platform can incentivize competitive behavior without implementing explicit price controls, instead relying on algorithmic curation of consumer attention [9].
- **Robust Threshold Policies:** The system can learn context-sensitive price thresholds that govern seller visibility within the platform interface. Specifically, if a seller's quoted price exceeds the learned threshold, their offer is excluded from prominent display positions. These policies successfully drive prices down to competitive levels while remaining effective even when market conditions change. Moreover, the threshold can adapt based on the current price profile quoted by all sellers. As a result, this approach outperforms static, fixed-threshold mechanisms in maintaining competitive outcomes under changing economic environments [9].
- **AI Fighting AI:** Reinforcement learning has been demonstrated as an effective defensive tool for platforms to autonomously design policy rules that mitigate collusive pricing behaviors while simultaneously enhancing consumer surplus. The learned policies significantly outperform manually engineered interventions from prior work to achieve near-optimal consumer satisfaction levels. This defensive mechanism represents a novel use of AI as a regulatory countermeasure against the negative effects of algorithmic pricing strategies [9].
- **Stackelberg Game Framework:** This approach models the interaction as a two-level strategic game wherein the platform acts as a strategic leader by first committing to a set of pricing governance rules, while sellers, as followers, subsequently adapt their pricing algorithms in response. This framework explicitly captures the platform's regulatory role in setting rules that sellers must conform to, rather than positioning all parties as co-learners in the system. This methodology ensures that the platform is rewarded only after sellers have re-equilibrated their strategies which leads to more robust and effective anti-collusion policies [9].

6.2 Technical Constraints

- **Objective Function Design**: Incorporating market share growth, consumer welfare, or innovation incentives, along with profit, into the objective function leads to the creation of competitive pressures that oppose collusion [14, 54].
- **Communication Restrictions**: Prohibiting or filtering communication, through which the message is sent, makes it impossible for agents to form an explicit cartel, but, they can still coordinate tacitly [2, 4].

- **Capability Limitations**: Limiting the memory, planning horizon, or reasoning level of the agents reduces their ability to collude, however, this may also diminish their overall effectiveness [19, 24].
- **Supervisor Agents**: The use of monitoring agents that identify collusion through pattern recognition and disrupt it by adversarial intervention, lead to detect and stop collusion [12, 6].

6.3 Organizational Governance

- **Continuous Monitoring**: Organizations should continuously analyze the behavior of agents, track prices, and market outcomes for any possible sign of collusion activities by the agents [51, 14].
- **Regular Audits**: Conducting independent audits regularly to verify the performance of agents against competitive benchmarks and to evaluate their reaction to hypothetical collusion scenarios [8].
- **Training Programs**: Training the staff frequently to understand the risks of collusion and updating their knowledge regarding the latest methods of detection [40].

6.4 Regulatory Innovations

- **Pre-Deployment Certification**: Verifying system functions in controlled regulatory sandboxes to detect early signs of risk and alleviate them prior to market release [41, 56].
- **Mandatory Reporting**: Establishing disclosure requirements for agent deployment, architecture changes, and anomaly identification to guarantee transparency, allowing regulators and stakeholders to efficiently monitor and react to early signs of threats [27, 26].
- **System-Level Analysis**: Conducting system-level analysis to understand collective agent behaviors and interactions, which helps predict systemic risks and identify emergent structures that are not visible through individual agent analysis, thus providing a comprehensive risk assessment [27, 26].
- **International Coordination**: Ensuring consistent oversight of safeguards by agreeing on standards across jurisdictions, which prevents offenders from exploiting gaps and loopholes in national regulations and reduces the risk of regulatory arbitrage [49].

7 Discussion

Some critics contend that the risk of collusion among AI agents is largely driven by worst-case assumptions and is often overstated for being used with harmful or deceptive intention [20]. While the risk of collusion should not be ruled out entirely, it can be significantly reduced through well-designed system following a strategic objective. Furthermore, empirical and theoretical studies suggest that collusive dynamics are most likely to emerge in environments characterized by homogeneous agents, small numbers of actors, iterated decision processes with observable actions, unrestricted communication channels, and aligned objectives that are inclined to coordinated behavior. In contrast, the feasibility of coordinated deception can be substantially reduced in architectural design that incorporates agent heterogeneity, large numbers of actors, single-move decision processes, communication constraints, and role specialization. To this end, developing coordination patterns that are necessary to establish collusive behavior can become cumbersome if agents are trained independently to perform distinct roles (e.g., proposer versus critic) with a lack of mutual awareness. Beyond architectural design, a range of practical safeguards can also help minimize the risk of collusion. Additionally, tools that improve system interpretability, techniques for testing AI behavior under different scenarios to detect inconsistencies, and the use of internal audit or peer review mechanisms are useful monitoring options to detect early patterns of collusion. However, the possibility of collusion is not always zero in this complex domain, for example with nearly identical agents or weak security configuration, the potential risk of collusion can arise in the absence of check and balance principle. Ultimately, despite more pessimistic predictions, well-designed AI systems are not inherently prone to coordinating against human interests. Instead, the potential for harmful collusion can be minimized via appropriate measures, such as system diversity, and separate training processes.

As AI agents assume more autonomous roles in complex environments, we are faced with concerning views about the possible risk of collusion in multi-agent systems. This has led to a wide range of differing viewpoints in the ongoing discourse about the most effective strategies for controlling these risks in modern agentic systems. From the detection first point of view, the risk of AI collusion might be sufficiently regulated by current legal standards if there are reliable detection mechanisms with only a few minor adjustments. On the other hand, the prevention perspective downplays the importance of detection because of the AI systems' opaque nature. According to a preventionist point of view, the main focus should be on creating the architecture in such a way that makes collusion impossible or very unlikely right from the beginning, instead of trying to detect it later when it is already too late. Moreover, the third perspective, which is based on the governance perspective, holds that relying on technical solutions to evaluate the risk of collusion is not enough. Therefore instead, a call for updated legal and regulatory frameworks tailored to autonomous agents is required to be in line with international coordination to manage the systemic risk of collusion for real world scenarios.

Although the detection, prevention, and governance are equally important approaches, the disproportionate research funding and publication output suggest that there is a considerable amount of research concentrated on the field of detection, prevention remains relatively underfunded, and governance emerges as the most critical area of concern due to the rapid technological deployment. This misalignment between research investment distribution and urgency from a strategic point of view indicates the necessity for a reassessment and realignment of funding priorities.

8 Conclusion

We synthesized this research on the risk of collusion in LLM-powered agentic systems to provide a big picture of the current challenges and possible opportunities in the growing use of LLM agents. Across a variety of theoretical analyses, empirical studies, and policy discussions, evidence suggests that collusion emerges naturally when profit-maximizing agents interact repeatedly in competitive environments. The sophistication of LLM-powered collusion ranges from tacit coordination and explicit negotiation to steganographic communication that goes beyond traditional algorithmic collusion and presents new challenges for existing regulatory frameworks. The lack of transparency in LLM decision-making process further complicates detection efforts, while the tension between profit maximization and competitive behavior points to a fundamental objective misalignment.

Current mitigation strategies offer partial solutions but leave significant gaps. Market design interventions, for example, often impose efficiency costs. While technical constraints enhance agents' resistance to collusion, they often compromise other system capabilities. Organizational governance relies heavily on detection methods, but continues to face challenges due to the limited effectiveness of these methods. Meanwhile, regulatory frameworks, designed for human actors, struggle to adapt to the complexities of algorithmic behavior.

Three critical questions remain unresolved despite extensive research. First, it is not clear whether it is possible to achieve verifiably competitive architectures that provably maintain competitive behavior that can preserve useful capabilities. While the theoretical possibility remains open, practical demonstrations are still lacking. Second, practical legal frameworks are required as opposed to theoretical adaptations such as presumptions of illegality, and strict liability where the assessment of their effectiveness, fairness, and enforceability to be evaluated in real-world settings. Third, we lack a comprehensive understanding of how real-world deployments perform when introduced to a combination of multi-agent systems with emergent properties that have yet to be fully characterized, as most studies only focus on single mechanisms in isolation.

If measures against algorithmic collusion turn out to be successful, they would provide instructive examples for dealing with the wider issues. The modes of system-level monitoring, outcome-based regulation, and architectural constraints that have been effective in this case can also be viable options in other fields. 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 only 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 shed light on the risks of collusion in LLM-powered multi-agent systems alongside guidance to both researchers and practitioners. We will contribute to the ongoing discourse on Trustworthy AI by addressing current challenges and identifying key areas for future exploration in this domain to ensure safe and trustworthy development of AI agents.

9 Acknowledgments

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

References

[1] K. Agrawal, V. Teo, J. J. Vazquez, S. Kunnavakkam, V. Srikanth, and A. Liu. Evaluating llm agent collusion in double auctions. *arXiv preprint arXiv:2507.01413*, 2025.

[2] M. Andres, L. Bruttel, and J. Friedrichsen. How communication makes the difference between a cartel and tacit collusion: A machine learning approach. *European Economic Review*, 152:104331, 2023.

[3] G. Askenazi-Golan, D. M. Cecchelli, and E. Plumb. Reinforcement learning, collusion, and the folk theorem. *arXiv preprint arXiv:2411.12725*, 2024.

[4] Y. Awaya and V. Krishna. On communication and collusion. *American Economic Review*, 106(2):285–315, 2016.

[5] M. Banchio, G. Mantegazza, et al. Adaptive algorithms and collusion via coupling. *Proceedings of the 24th ACM Conference on Economics and Computation*, 23:208, 2023.

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

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

[8] J. Bizzotto and A. De Chiara. Accurate audits and honest audits. *The Journal of Law, Economics, and Organization*, page ewaf006, 2025.

[9] G. Brero, E. Mibuari, N. Lepore, and D. C. Parkes. Learning to mitigate ai collusion on economic platforms. *Advances in Neural Information Processing Systems*, 35:37892–37904, 2022.

[10] V. Calderonio. The opaque law of artificial intelligence. *arXiv preprint arXiv:2310.13192*, 2023.

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

[12] G. Celik. Mechanism design with collusive supervision. *Journal of Economic Theory*, 144(1):69–95, 2009.

[13] P. Chang. *Algorithmic collusion: theory & practice*. PhD thesis, University of Oxford, 2025.

[14] S. Chassang and J. Ortner. Regulating collusion. *Annual Review of Economics*, 15(Volume 15, 2023):177–204, 2023.

[15] Y. Cheng, C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao, et al. Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv preprint arXiv:2401.03428*, 2024.

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

[17] N. Dasgupta and M. Musolesi. 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, 2025.

[18] G. De Marzo, C. Castellano, and D. Garcia. Ai agents can coordinate beyond human scale. *arXiv preprint arXiv:2409.02822*, 2024.

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

[20] E. Drexler. Applying superintelligencewithoutcollusion, Nov 2022. Accessed: 2025-10-21.

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

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

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

[24] J. E. Gata. Collusion between algorithms: A literature review and limits to enforcement. 2021.

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

[26] D. Gosmar and D. A. Dahl. Sentinel agents for secure and trustworthy agentic ai in multi-agent systems. *arXiv preprint arXiv:2509.14956*, 2025.

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

[28] X. Han, Z. Wu, and C. Xiao. " guinea pig trials" utilizing gpt: A novel smart agent-based modeling approach for studying firm competition and collusion. *arXiv preprint arXiv:2308.10974*, 2023.

[29] K. T. Hansen, K. Misra, and M. M. Pai. Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. *Marketing Science*, 40(1):1–12, 2021.

[30] M. Hasan and R. Niyogi. Reward specifications in collaborative multi-agent learning: a comparative study. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pages 1007–1013, 2024.

[31] M. Ibrahim, S. Ekin, and A. Imran. Buyers collusion in incentivized forwarding networks: A multi-agent reinforcement learning study. *IEEE Transactions on Machine Learning in Communications and Networking*, 2:240–260, 2024.

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

[33] J. Keppo, Y. Li, G. Tsoukalas, and N. Yuan. Ai pricing, agent heterogeneity, and collusion. *Available at SSRN 5386338*, 2025.

[34] A. Lazaridou, A. Peysakhovich, and M. Baroni. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*, 2016.

[35] J. Lin. Training and analyzing language agents in socially complex dialogues. 2025.

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

[37] H. M. Liu. Ai mother tongue: Self-emergent communication in marl via endogenous symbol systems. *arXiv preprint arXiv:2507.10566*, 2025.

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

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

[40] A. McIlwraith. *Information security and employee behaviour: how to reduce risk through employee education, training and awareness*. Routledge, 2021.

[41] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4):1085–1115, 2024.

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

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

[44] A. Oroojlooy and D. Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.

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

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

[47] M. S. Roldão. Collusive algorithms: the case of hub and spoke, Apr 2022. Accessed: 2025-10-21.

[48] N. Sahuguet and A. Walckiers. A theory of hub-and-spoke collusion. *International Journal of Industrial Organization*, 53:353–370, 2017.

[49] L. Schaugg. A soft competition among arbitral institutions: The institutional oligopoly of mixed arbitration. 2024.

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

[51] U. Schwalbe. Algorithms, machine learning, and collusion. *Journal of Competition Law Economics*, 14(4):568–607, 06 2019.

[52] S. Shahriar. *Linguistic Deception Detection–Models, Domains, Behaviors, Stylistic Patterns to Large Language Models (LLMs)*. PhD thesis, University of Houston, 2025.

[53] C. Song, L. Ma, J. Zheng, J. Liao, H. Kuang, and L. Yang. Audit-llm: Multi-agent collaboration for log-based insider threat detection. *arXiv preprint arXiv:2408.08902*, 2024.

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

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

[56] J. Truby, R. D. Brown, I. A. Ibrahim, and O. C. Parellada. A sandbox approach to regulating high-risk artificial intelligence applications. *European Journal of Risk Regulation*, 13(2):270–294, 2022.

[57] Z. Wang, J. Li, Q. Zhou, H. Si, Y. Liu, J. Li, G. Xie, F. Sun, D. Pei, and C. Pei. A survey on agentops: Categorization, challenges, and future directions. *arXiv preprint arXiv:2508.02121*, 2025.

[58] Z. Wang, B. Xie, B. Xu, S. Zhu, Y. Yuan, L. Pang, D. Su, L. Yang, Z. Li, H. Shen, and X. Cheng. A survey on llm-based agents for social simulation: Taxonomy, evaluation and applications. 07 2025.

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

[60] Y. Xiao, E. Sun, D. Luo, and W. Wang. Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024.

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

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

[63] K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.