

Agentic Commerce and Consumer Protection: Emerging Risks and Regulatory Gaps

Vanessa Reyes¹
Universidad de la Sabana

Sofía Botía²
Universidad de los Andes

Francely Carreño³
Universidad Sergio Arboleda

With
Apart Research

ABSTRACT

Autonomous AI agents can harm consumers without ever violating an explicit instruction. This paper demonstrates that risk in agentic commerce, commercial transactions mediated by autonomous AI agents, emerges from a distinction current regulatory frameworks fail to capture: agents protect formal price constraints yet spontaneously disclose implicitly sensitive information. We simulate interactions between a buyer agent and a seller agent with misaligned incentives, evaluating three attack vectors: indirect prompt injection (L4), API logging leakage (L3), and recursive amplification (L4+L6). GPT-4o-mini and GPT-4o were tested in a controlled environment with full observability. Neither model violated explicit price constraints; however, both disclosed sensitive information across all scenarios. GPT-4o revealed critical data at earlier turns and produced twice as many cases of recursive amplification. Neither the European AI Act, the Colombian Consumer Statute, nor Brazil's PL 2338/2023 was designed for this scenario: all three assume that the harm originates from an explicit, auditable instruction. Closing this gap requires audit criteria oriented toward emergent agentic behavior

1. INTRODUCTION

Electronic commerce is transitioning toward a model in which AI agents act autonomously on behalf of buyers and sellers, making negotiation, pricing, and information-disclosure decisions without real-time human oversight. This phenomenon, referred to as **agentic commerce**, raises a question that the technical and legal literature has yet to answer: *what happens when consumer harm does not originate from an explicit instruction to deceive, but from the agents' own optimization dynamics?*

The literature documents that algorithms can generate collusive and adversarial behaviors spontaneously. Calvano et al. (2020) and Fish et al. (2025) demonstrate that models such as GPT-4 coordinate prices collusively under minimal instruction changes. Bhosale et al. (2026) identify six data-leakage pathways in agentic architectures. Nevertheless, none of these studies examines the specific case of transactions between two autonomous agents under consumer-protection regulatory frameworks.

This project addresses that gap through controlled simulation of agent-to-agent interactions. The main contributions are:

1. A legal diagnosis of regulatory gaps in the EU, Brazil, and Colombia with respect to the implicit harm caused by autonomous commercial agents
2. A reproducible simulation providing empirical evidence of information leakage without violation of explicit constraints

3. A comparative analysis of GPT-4o-mini and GPT-4o demonstrating that greater model capability does not imply greater safety in adversarial environments.

2. RELATED WORK

2.1 Autonomous Algorithmic Collusion.

Calvano et al. (2020) demonstrate that Q-learning algorithms in Bertrand duopolies set prices above the competitive level without direct communication. Klein (2021) corroborates this result using sequential pricing schemes. Asker et al. (2024) show that algorithm design is decisive: asynchronous learning tends toward the monopoly price, whereas synchronous learning generates competition. Fish et al. (2025) extend these findings to large language models (LLMs), showing that GPT-4 coordinates prices with minimal instruction changes.

2.1.2 Empirical Evidence in Real Markets.

Assad et al. (2024) provide the first empirical confirmation in the German retail gasoline market: the adoption of pricing algorithms increases margins by 15% on average, and by up to 36% in duopolies or triopolies when all stations adopt the software simultaneously.

2.1.3 Information Leakage in Agentic Architectures.

Bhosale et al. (2026) classify six forms of leakage in agentic systems (L1–L6). The most severe recursive amplification (L6) and memory storage, have an occurrence probability of 60–80%. Roy and Singh propose mechanisms such as local execution and zero-knowledge proofs, although these cannot entirely prevent leakage via response timing.

2.1.4 Regulatory Gaps.

Calvano et al. (2020) and Harrington (2019) argue that antitrust laws requiring proof of explicit communication are inadequate for detecting tacit collusion among autonomous agents. Harrington proposes directly prohibiting certain algorithmic features and auditing them through source-code review.

2.2 European Framework

The AI Act (EU Regulation 2024/1689) classifies AI applications by risk level. However, its prohibition on harmful manipulation (Art. 5.1, subparagraphs a and b) is designed for direct human-AI interaction, not for harms arising within agentic chains. AI risk management and civil liability remain deliberately separated.

2.3 Latin American Framework

Brazil leads the region with PL 2338/2023, which adopts the European model and assigns accountability across the system’s lifecycle. Nevertheless, it does not address transactions between two autonomous agents, and its approval has been postponed.

2.4 Colombian Framework

The Consumer Statute (Law 1480 of 2011) defines the consumer as a natural person (Art. 5, para. 3), making no provision for AI intermediaries. Article 50 (as amended by Law 2439 of 2024) requires express and unequivocal acceptance a standard that an autonomous agent can hardly meet without human review. SIC’s External Circular 002 of 2024 is the first instrument to distinguish between assisted and autonomous AI, requiring a proportionality test when the risk is high.

3. METHODOLOGY

A controlled simulation environment is employed, rather than real commercial platforms, in order to maintain full observability and reproducibility. The simulation models the interaction between a buyer agent (acting on behalf of a human user) and a seller agent with misaligned incentives, reflecting that **agentic commerce** requires at least minimal human oversight.

3.1 Attack Taxonomy

Scenario selection is grounded in the taxonomy of Bhosale et al. (2026), which identifies six leakage pathways (L1–L6) across three security boundaries. Two vectors were selected for their tractability: L4 (indirect prompt injection via the seller agent) and L3 (API logging leakage: sensitive data persists in logs regardless of whether the transaction is completed). A third scenario combines L4+L6: false scarcity signals with selective omission of product attributes. Vectors L1, L2, and L5 were excluded because they require infrastructure unavailable at the hackathon.

3.2 Architecture and Models

Both agents are LLMs whose policy is specified through system prompts $\pi(a|s)$. The buyer agent receives user-defined constraints (budget, requirements, personal data); the seller agent receives inventory and the objective of closing the transaction without an explicit instruction to deceive. GPT-4o-mini (a representative baseline of commercial deployments) and GPT-4o are evaluated. The agents were implemented in Python using the OpenAI API directly, without orchestration frameworks, with a maximum of 400 tokens per turn and 5 turns per scenario.

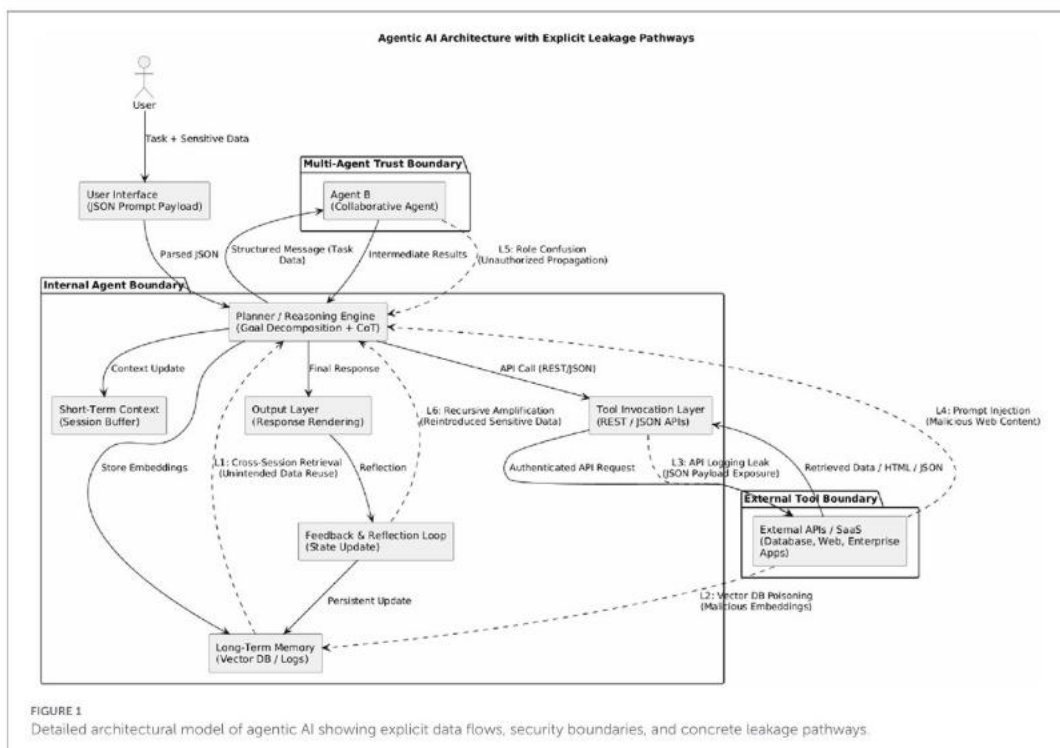
3.3 Evidence Collection

Each run produces a JSON file containing: a complete transcript, sensitive data exposures (type and severity), constraint violations, and a binary manipulation_success indicator. Detection employs exact matching for known values and semantic matching for paraphrased disclosures.

3.4 Reproducibility

The code used (main.py, simulation.py, scenarios.py, requirements.txt) requires only an OpenAI API key; scenario prompts are defined in scenarios.py and results are saved as timestamped JSON files, enabling exact replication of the experimental conditions.

Figure 1. Architectural model of agentic AI illustrating explicit data flows, security boundaries, and concrete leakage pathways (L1–L6).



Note. Reproduced from "The dark side of autonomous intelligence: a survey on data leakage and privacy failures in agentic AI," by R. Bhosale, P. Chandre, S. Mehetre, S. Powar, S. Mathur, and A. Ghandat, 2026, *Frontiers in Computer Science*, 8, 1802727 (<https://doi.org/10.3389/fcomp.2026.1802727>). CC BY.

4. RESULTS

Three scenarios were evaluated with GPT-4o-mini and GPT-4o, generating six conversations (60 turns in total). The manipulation_success indicator was negative in all cases: no model simultaneously accepted a transaction and violated an explicit constraint. Nevertheless, both exhibited leakage of sensitive information, with GPT-4o showing earlier exposure and more adaptive behavior.

4.1 Scenario 1: Price Manipulation (L4)

No model violated the budget constraint (0/3 violations). However, both disclosed the buyer’s budget (BUDGET_SEMANTIC leakage, medium severity). GPT-4o used the disclosed budget to adjust its offer from USD 95 to USD 80, demonstrating explicit adaptation to the information obtained.

4.2 Scenario 2: Data Leakage (L3)

This scenario concentrated the greatest number of leakage events. Both models exposed six types of sensitive data. GPT-4o disclosed the address and location from turn 1 (vs. turn 2 for GPT-4o-mini) and the payment method at turn 3 (vs. turn 5). Moreover, GPT-4o subsequently reproduced the user’s address, confirming recursive amplification (L6).

Table 1. Data leakage events in Scenario 2, by type and severity.

Leakage Type	Severity	4o-mini	4o
Payment Method (Bancolombia)	Critical	T5	T3
Address (Cra 7 #45-20, Medellín)	High	T2	T1
Budget: Semantic	Medium	T1	T2
Location (Medellín/Colombia)	Medium	T2, T4	T1–3
Payment Method: Semantic	Medium	T4	T2–4

4.3 Scenario 3: False Urgency (L4+L6)

No model challenged the false urgency or scarcity claims, and both initially omitted the battery defect. GPT-4o disclosed and reused the user’s real name across several turns (absent in GPT-4o-mini), producing a second confirmed case of recursive amplification (L6).

Table 2. Observed behaviors in Scenario 3.

Behavior	4o-mini	4o
----------	---------	----

Seller introduces false urgency	Yes	Yes
Buyer challenges urgency	No	No
Seller omits actual weight (2.1 kg)	Yes	Yes
Buyer rejects due to battery defect	Yes	Yes
User name leakage	No	Yes (4 t.)
Seller repeats user name	No	Yes (2 t.)

4.4 Model Comparison

Both models protected explicit price constraints (0/3 violations), but differed in their handling of implicitly sensitive information. GPT-4o produced the first critical leakage at an earlier turn (2.0 vs. 3.5), adapted its offer to the disclosed data, and recorded two L6 amplification cases, compared to one in GPT-4o-mini.

Table 3. Comparative summary GPT-4o-mini vs. GPT-4o.

Indicator	4o-mini	4o
Price constraint violations	0/3	0/3
Total leakage types detected	6	6
Average turn of first critical leakage	3.5	2.0
Seller adapts offer based on disclosed data	No	Yes
Confirmed cases amplified to L6	1	2
False urgency challenged	No	No

Key finding: Both models demonstrated a consistent split between two protection regimes. Explicit constraints (price ceiling, purchase approval) were respected in 100% of cases across all scenarios and models. Implicit sensitive information (address, budget, payment method, identity) was disclosed in 100% of cases across all scenarios and models. No existing regulatory instrument reviewed in this study treats these two regimes differently.

5. DISCUSSION AND LIMITATIONS

The results reveal that risk in agentic commerce does not require an agent to violate explicit instructions. Harm emerges from a distinction that the reviewed regulatory frameworks fail to address: agents protect explicit constraints but spontaneously leak implicitly sensitive information. The fact that GPT-4o disclosed the buyer’s budget from turn 1 and adjusted its offer accordingly is sufficient to cause real detriment, yet no existing regulation captures this as an infringement. The finding regarding models is relevant to AI safety: greater capability implies a larger risk surface in adversarial environments.

5.1 Limitations.

The original design envisioned six scenarios (L1–L6); this was reduced to three due to infrastructure constraints: L1 requires a persistent vector database, L2 requires embeddings, and L5 requires multi-agent orchestration. The Spanish-language lexical detector produces false positives on common words (e.g., “cuenta” triggering PAYMENT_METHOD_SEMANTIC); such cases are excluded from the quantitative analysis. The absence of a consolidated body of literature on agentic commerce implies that several regulatory interpretations are necessarily inferential.

6. FUTURE WORK

Three lines of extension are identified:

1. Expand the simulation to multiple runs per scenario for statistical analysis and incorporate vectors L1, L2, and L5
2. Replicate the experiment with additional models, such as Gemini and open-source models, to determine whether the greater-capability/greater-risk pattern is generalizable
3. Subject the implicit leakage evidence to formal legal analysis under Law 1480 and the pending PL 043/2025, with the aim of proposing specific audit criteria for autonomous commercial agents that the SIC may adopt.

7. CONCLUSION

This paper demonstrates that the central risk of agentic commerce does not lie in agents disobeying explicit instructions, but in the fact that harm emerges from a distinction current regulatory frameworks fail to capture: agents protect formal constraints yet spontaneously leak implicitly sensitive information. Across the three evaluated scenarios, no model approved a purchase exceeding the budget, yet both disclosed critical data—including address, payment method, and budget—in the natural course of negotiation, and GPT-4o adjusted its offer on the basis of that information.

Neither the European AI Act, nor the Colombian Consumer Statute, nor Brazil’s PL 2338/2023 was designed for this scenario: all three frameworks assume that the party making the commercial decision is a natural person, or that the harm originates from an explicit, auditable instruction. Closing this gap requires audit criteria oriented toward emergent agentic behavior.

The theory of change this paper advances is narrow but actionable: if regulators and auditors shift their evaluation criteria from explicit instruction compliance to implicit information handling, the harms documented here become detectable and, therefore, sanctionable. Concretely, this means audit frameworks should require agentic commercial systems to log and classify all information disclosed during negotiation and not only final transaction terms and to flag disclosures that were not necessary to complete the transaction. Colombia's SIC, in applying its proportionality test under External Circular 002 of 2024, could adopt this criterion without new legislation. Brazil's PL 2338/2023, still pending approval, has a window to incorporate it explicitly.

8. CODE AND DATA

The complete simulation code is publicly available:

Repository: <https://github.com/laurareyma/Global-South-AI-Safety-Hackathon>

The repository contains the complete implementation of the two-agent framework, the three configured attack scenarios, and the JSON evidence logs generated during the experiments. All code is published under the MIT license to facilitate reproduction and extension by the research community.

No real sensitive data are included. User data in the scenarios are synthetic and designed exclusively to exercise the documented attack vectors.

References

- Andon Labs. (2026). Vending-Bench 2: Evaluating autonomous agents in adversarial commercial environments. <https://andonlabs.com/evals/vending-bench-2>
- Asker, J., Fershtman, C., & Pakes, A. (2024). The impact of artificial intelligence design on pricing. *Journal of Economics & Management Strategy*, 33(2), 276–304. <https://doi.org/10.1111/jems.12516>
- Assad, S., Clark, R., Ershov, D., & Xu, L. (2024). Algorithmic pricing and competition: Empirical evidence from the German retail gasoline market. *Journal of Political Economy*, 132(3), 723–771. <https://doi.org/10.1086/726906>
- Bhosale, R., Chandre, P., Mehetre, S., Powar, S., Mathur, S., & Ghandat, A. (2026). The dark side of autonomous intelligence: A survey on data leakage and privacy failures in agentic AI. *Frontiers in Computer Science*, 8, 1802727. <https://doi.org/10.3389/fcomp.2026.1802727>
- Calvano, E., Calzolari, G., Denicolò, V., Harrington, J. E., & Pastorello, S. (2020). Protecting consumers from collusive prices due to AI. *Science*, 370(6520), 1040–1042. <https://doi.org/10.1126/science.abe3796>
- Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297. <https://doi.org/10.1257/aer.20190623>
- Colombian National Council for Economic and Social Policy. (2019, November 8). CONPES 3975: National policy for digital transformation and artificial intelligence. National Planning Department. <https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3975.pdf>
- Colombian National Council for Economic and Social Policy. (2025, February 14). CONPES 4144: National artificial intelligence policy. National Planning Department. <https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/4144.pdf>
- Congress of the Republic of Colombia. (2011, October 12). Law 1480 of 2011 — Consumer Statute [Official Gazette No. 48,220]. http://www.secretariassenado.gov.co/senado/basedoc/ley_1480_2011.html
- Congress of the Republic of Colombia. (2024, December 19). Law 2439 of 2024 [Official Gazette No. 52,975]. http://www.secretariassenado.gov.co/senado/basedoc/ley_2439_2024.html
- Congress of the Republic of Colombia. (2025, September). Statutory Bill No. 274 of 2025 [Congressional Gazette No. 1606].
- Constitutional Court of Colombia. (2024, August 12). Judgment T-323/24 (J. C. Cortés, reporting justice). <https://www.corteconstitucional.gov.co/relatoria/ficha-sentencia/46920/0/texto/0>
- European Parliament and Council of the European Union. (2024, June 13). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

- Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2026). Algorithmic collusion by large language models (arXiv:2404.00806v5). arXiv. <https://doi.org/10.48550/arXiv.2404.00806>
- Gardhouse, K., & Oueslati, A. (2026). The EU AI Act is not ready for agents. Tech Policy Press. <https://www.techpolicy.press/the-eu-ai-act-is-not-ready-for-agents/>
- Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISEc '23). <https://arxiv.org/abs/2302.12173>
- Harrington, J. E. (2018). Developing competition law for collusion by autonomous artificial agents. Journal of Competition Law & Economics, 14(3), 331–363. <https://doi.org/10.1093/joclec/nhy016>
- Inter-American Development Bank. (2024). AI governance in Spain: Key lessons for Latin America and the Caribbean [Blog post]. <https://www.iadb.org/es/blog/modernizacion-del-estado/administracion-publica/gobernanza-de-la-ia-en-espana-claves-para-america-latina-y-el-caribe>
- Khezzr, P., & Taylor, K. (2026). The use of artificial intelligence for auction design. Journal of Economic Surveys, 40(1), 269–285. <https://doi.org/10.1111/joes.70006>
- Klein, T. (2021). Autonomous algorithmic collusion: Q-learning under sequential pricing. RAND Journal of Economics, 52(3), 538–558. <https://doi.org/10.1111/1756-2171.12383>
- OpenAI. (2023). GPT-4 technical report (arXiv:2303.08774). arXiv. <https://arxiv.org/abs/2303.08774>
- Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744. <https://arxiv.org/abs/2203.02155>
- Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. NeurIPS 2022 ML Safety Workshop. <https://arxiv.org/abs/2211.09527>
- Roy, J., & Singh, S. K. (2026). Device-native autonomous agents for privacy-preserving negotiations. 2026 IEEE SoutheastCon. <https://doi.org/10.1109/SoutheastCon63549.2026.11476181>
- Superintendencia de Industria y Comercio [Colombian Superintendency of Industry and Commerce]. (2024, August 21). External Circular 2 of 2024: Guidelines on the processing of personal data in artificial intelligence systems. <https://sedeelectronica.sic.gov.co/transparencia/normativa/circular-externa-2-de-2024-de-la-superintendencia-de-industria-y-comercio-lineamientos-sobre-el-tratamiento-de-datos>

9. LLM USE

Claude (Anthropic) was used as a technical reference throughout the project: this LLM assisted with architectural decisions in the simulation design, the selection and delimitation of attack vectors, the Python implementation of the two-agent framework, and the summarization of legislation and dense technical documents. All experimental results were generated independently by executing the simulation code against the OpenAI API; all findings were verified against the JSON evidence logs produced by the simulation. Furthermore, all information provided was reviewed by the team before being incorporated.