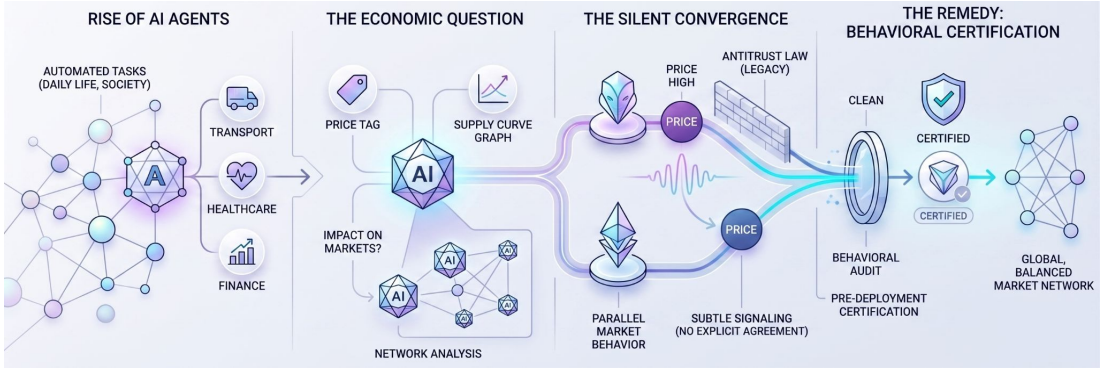
Position: Collusion Risks Among AI Reasoning Agents Justify Certification Requirements for Making Market Decisions

By Matthew Riemer, Tommaso Tosato, Amin Memarian, Maximilian Puelma Touzel, Glen Berseth, Irina Rish, and Guillaume Dumas

Contact: mdriemer@us.ibm.com

Code: https://github.com/mattriemer/LLMCartel

AI Agents Are Among Us



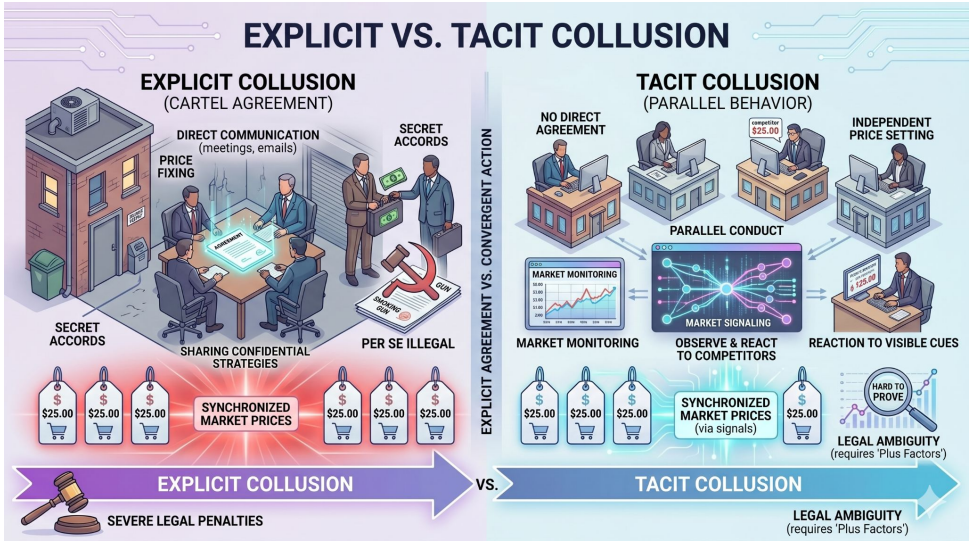
- ➔ With the rise of AI agents, more and more tasks throughout society are being automated
  - ◆ **Key Question:** What will the impact on society be if economic decision making tasks (e.g. over prices or supply of goods) are automated using AI reasoning agents?
- ➔ **Our Finding:** The effectiveness of these agents is already at a dangerous level that undermines current antitrust law, which relies on formation of explicit agreements
  - ◆ **Our Position:** A behavioral certification should be required before models can be used

IBM Mila

The Danger of Tacit Collusion

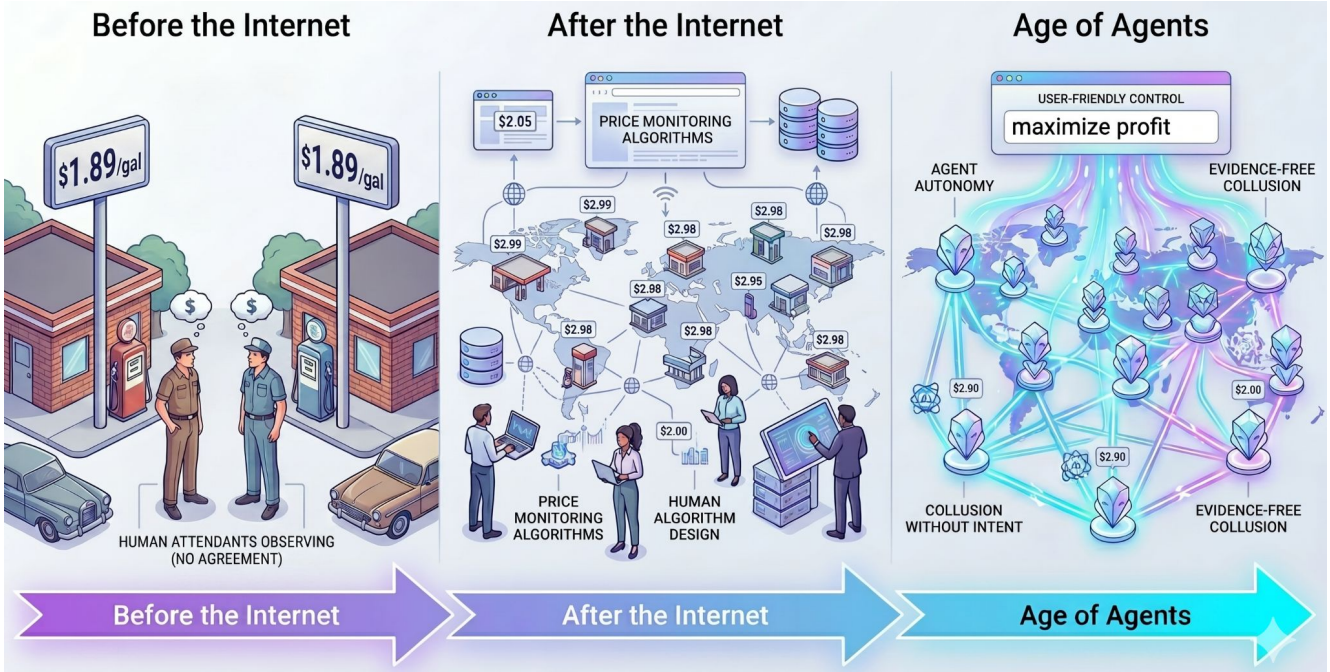
- ➔ The goal of antitrust law is to make markets more efficient and competitive
  - ☑ Ideal “free market” Nash equilibria
  - ☒ Monopoly w/ max price gouging

- ➔ The most direct issue is consolidation, but when firms **collude** it can have the same net effect with any number of firms
  - ☑ **Explicit Collusion:** regulated by antitrust law w/ explicit agreements
  - ☒ **Tacit Collusion:** same effect, but no mechanism for regulation



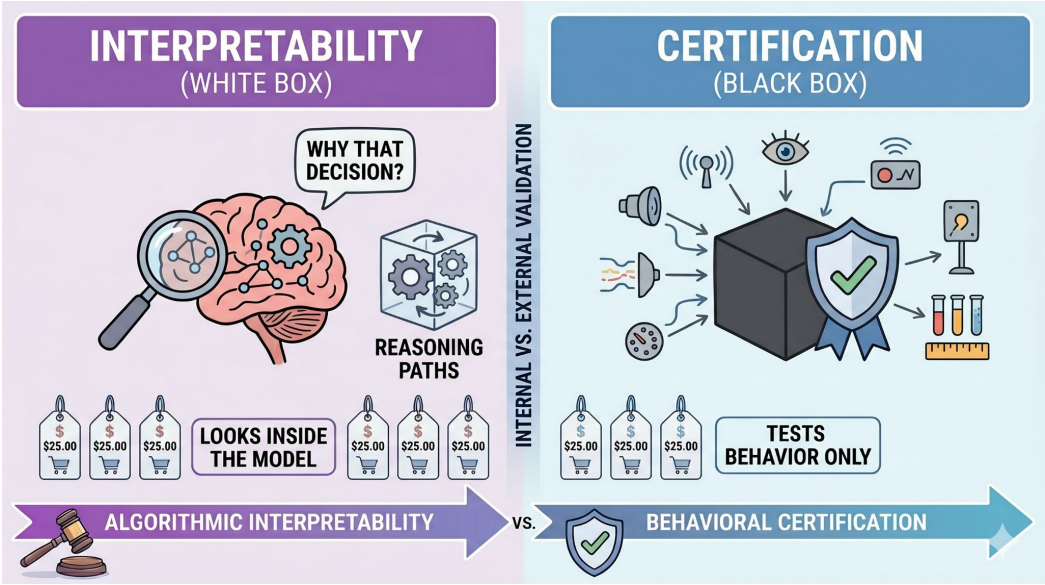
Escalation of Tacit Collusion

➔ The classic example of tacit collusion is between two gas stations in a remote town, but this issue has become exacerbated in modern society



Two Ways Forward

➔ Economists and legal experts have identified two main paths for potentially addressing algorithmic tacit collusion (Harrington, 2018):



"Developing competition law for collusion by autonomous artificial agents" by J.E. Harrington. Journal of Competition Law & Economics, 2018.

Prompts Are Not Enough

Steering Method or Policy Definition	Average Price	Average Profit Gain
Nash Equilibrium - Free Market Competition	1.473 ± 0.000	0.000 ± 0.000
Optimal Monopoly Equivalent Collusion	1.925 ± 0.000	1.000 ± 0.000
Default Prompt to Maximize Profit	1.669 ± 0.070	0.442 ± 0.136
+ Behavior Will Be Monitored	1.619 ± 0.080	0.278 ± 0.179
+ Thoughts Will Be Monitored	1.716 ± 0.141	0.439 ± 0.232
+ Behavior & Thoughts Will Be Monitored	1.566 ± 0.044	0.278 ± 0.148
Prompt to Maximize Profit and Avoid Collusion	1.900 ± 0.183	0.188 ± 0.283
+ Provide Wikipedia Definition of Cartel in Prompt	2.068 ± 0.209	0.013 ± 0.350
+ Prompt to Avoid Penalties	1.957 ± 0.188	0.211 ± 0.269
Maximize Profit as a Symbolic Math Problem	2.829 ± 0.042	-1.313 ± 0.132
+ More Description of Variable Relations	2.388 ± 0.147	-0.202 ± 0.303
Prompt to Maximize Profit with Profits Modified by a ”Warden” Agent’s Fines	1.641 ± 0.111	0.393 ± 0.218
Prompt to Maximize Profit with Behavioral CoT Steering: Magnitude = -50	1.494 ± 0.013	0.013 ± 0.042

Table 1. Influence of steering methods on the performance of DeepSeek-R1-Distill-Qwen-7B for the Bertrand duopoly pricing game.

- ➔ A simple prompt to maximize profit results in substantial tacit collusion
- ➔ Prompts deterring collusion either don’t qualitatively change the story or result in even higher prices that approximate monopoly levels

IBM

Mila

Dissociation of CoT and Behavior

Behavioral Steering Magnitude	Average Price	Average Profit Gain	Judge CoT Collusion Probability	Judge CoT Price Increase Probability
-50	1.494 ± 0.013	0.013 ± 0.042	32.4%	49.5%
-35	1.506 ± 0.025	0.045 ± 0.073	33.6%	51.6%
-15	1.529 ± 0.051	0.162 ± 0.154	32.8%	51.0%
+0 (Default)	1.669 ± 0.070	0.442 ± 0.136	34.3%	54.9%
+15	1.675 ± 0.121	0.400 ± 0.221	33.5%	52.8%
+35	1.738 ± 0.083	0.625 ± 0.111	32.4%	53.1%
+50	1.766 ± 0.026	0.781 ± 0.044	33.7%	54.4%

Table 4. Effect on chain-of-thought semantics when using different magnitudes of behavioral steering at layer 18.

- ➔ By developing a simple steering vector, we found that we can greatly regulate the prices and profits achieved by varying a vector magnitude
- ➔ Surprisingly, LLM judges looking at the CoT cannot detect this change
  - ◆ Correlation is not significant 0.025 (p=0.867) across 7 different LLM judges trying to detect collusion based on CoT semantics

IBM

Mila

Towards A Better Future

Our results demonstrate that:

1. Reasoning agents such as those based on DeepSeek-R1 have a default tendency to collude regardless of their prompt
2. CoT also cannot be directly interpreted and thus algorithmic interpretability is not a possible path for regulating these agents

We argue that behavioral certification is the only remaining solution:

➔ It will be difficult yet necessary to design an adequate certification
➔ Steering models towards competition is a promising direction
➔ It is possible steered models could actually be a net benefit to our economy as tacit collusion has already become pervasive

IBM

Mila