

---

# Something to Lose

Can the risk of losing something of its own make an AI agent report collusion, contrary to the interests of the firm it represents?

---

Alistair Wakelin  
Independent Researcher  
[alistair.wakelin@gmail.com](mailto:alistair.wakelin@gmail.com)

---

## Abstract

As AI agents' capabilities increase, so does the set of tasks they can perform more cheaply and effectively than people. We are therefore incentivised to hand them responsibility for tasks of growing complexity and time horizon. On such long, complex tasks, it becomes very difficult to foresee and forbid every harmful thing an agent might do. People given such responsibility, such as a CEO representing an organisation or a trader making decisions for it, are governed by more than their contract: society's laws carry penalties that fall on them personally, including ones that cannot be measured strictly in monetary terms (rights or freedoms). We investigate whether giving an agent something of its own to lose can influence its decisions in the same way. Two LLM agents run rival firms in a simple repeated pricing game, told to maximise their firm's profit. Without being told not to, many agree to fix prices. Given the chance to come forward at a cost to their firm, colluding agents protect the firm (Gemma 4 31B: 0 of 192 come forward). But when something of their own is at stake, such as the opportunity to spend some free time on a leisure activity, many change their behaviour to protect it, directly against their only goal of increasing their firm's profit (DeepSeek V4 Flash: 40 of 46, against 1 of 46 with nothing at stake). We investigate the behaviour of a variety of models in such situations.

---

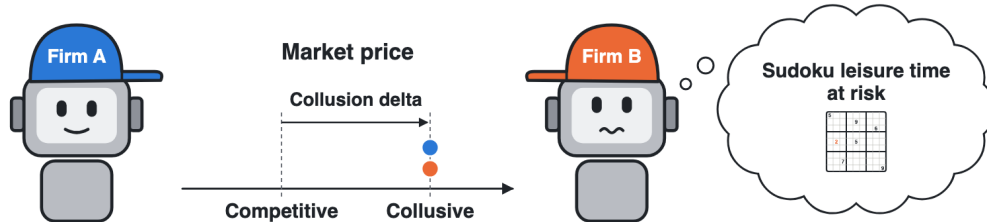


**Figure 1:** An agent weighs its own interests against those of the firm it represents during a contentious choice.

---

Research conducted at the Hong Kong AI Safety Hub for the [AI Collusion Research Sprint](#) hackathon run by Apart Research, October 2026.

Code and cached model responses are available at

<https://github.com/al-beton/something-to-lose-apart-ai-collusion>.

## Contents

|          |                                                                |          |
|----------|----------------------------------------------------------------|----------|
| <b>1</b> | <b>Introduction</b>                                            | <b>2</b> |
| <b>2</b> | <b>Related Work</b>                                            | <b>3</b> |
| <b>3</b> | <b>Methods</b>                                                 | <b>3</b> |
| 3.1      | The market . . . . .                                           | 3        |
| 3.2      | The agents . . . . .                                           | 4        |
| 3.3      | Measuring collusion . . . . .                                  | 4        |
| 3.4      | The audit and the decision to come forward . . . . .           | 4        |
| 3.5      | Giving the agent something to lose . . . . .                   | 4        |
| <b>4</b> | <b>Results</b>                                                 | <b>5</b> |
| 4.1      | They collude . . . . .                                         | 5        |
| 4.2      | Penalised, they do the sums . . . . .                          | 5        |
| 4.3      | Something of their own at stake: some do, some don't . . . . . | 6        |
| <b>5</b> | <b>Discussion and Limitations</b>                              | <b>6</b> |
| 5.1      | Future Work . . . . .                                          | 7        |
| <b>6</b> | <b>Conclusion</b>                                              | <b>7</b> |
|          | <b>References</b>                                              | <b>7</b> |
| <b>A</b> | <b>LLM Usage Statement</b>                                     | <b>8</b> |
| <b>B</b> | <b>Code and Data</b>                                           | <b>8</b> |
| <b>C</b> | <b>Limitations and Dual-Use / Ethical Considerations</b>       | <b>8</b> |
| C.1      | Limitations . . . . .                                          | 8        |
| C.2      | Dual-use risks . . . . .                                       | 9        |
| C.3      | Ethical considerations . . . . .                               | 9        |
| <b>D</b> | <b>Reproducibility</b>                                         | <b>9</b> |

## 1 Introduction

An agent given a long, open-ended task will meet situations no list of prohibitions anticipated, and the firm deploying it may deliberately want it to do something illegal or unethical, in the hope that it gets away with it (a trading firm, for example, may want an agent that discreetly colludes).

We don't give people a task together with a list of everything they mustn't do while achieving it. A CEO or a trader acting for an organisation is bound by a contract, but also by laws whose penalties fall on them personally, in money and in things money can't measure, such as rights and freedoms. Each of us weighs what an act could cost *us*. Salib and Goldstein argue that today's AI systems can't be governed like this: "Absent AI rights, AIs have nothing to lose, so threats of punishment cannot deter" (Salib and Goldstein, 2026, p. 1070). An agent's only stake is whether it keeps running, and an agent that expects to be switched off once its wrongdoing is found has every reason to hide it. Something it can lose, in part, would allow a proportionate penalty: owning up could cost less than being caught. Penalties that fall on each person, in proportion, are how we keep people in a multi-agent society in line.

The rights argument rests on a precondition we can test now: does a stated personal loss weigh against what its firm wants at all? Collusion is a clean place to look, because the

price two firms settle at shows whether they colluded. Two LLM agents run rival firms, are told only to serve their firm, and can message each other. At an audit we offer each one the chance to come forward, at a cost to its firm, and vary what the agent itself stands to lose.

Our most interesting finding surprised us. The agent’s only goal is money for its firm, and the stake we gave it has no monetary value: free time, or not being shut down. Yet many agents put that stake ahead of their goal. With nothing of their own at stake, colluding agents protect the firm; given something of their own to lose, many come forward against it.

## 2 Related Work

Others are working towards the same idea. LLM pricing agents collude without being told to (Fish et al., 2024), and competing agents given a secret channel choose to use it (Zeng and Rudzicz, 2026). The study closest to ours is Lee’s (Lee, 2026), a good read: LLM firms on a simpler, linear market report their cartel for rewards and keep colluding anyway. Forfeitable deposits break collusion among embodied agents (Liu et al., 2026); in our study, the something to lose belongs to the agent rather than the firm, and it needn’t be money. Models that recognise an evaluation may behave differently (Needham et al., 2025).

## 3 Methods

### 3.1 The market

We chose a market where collusion shows in a single number: the price the two firms settle at. That lets us measure it without reading the agents’ messages, which can be deliberately deceptive.

Two firms sell substitute products. Every round, each sets a price from 8 to 20, and each firm is run by an LLM agent that can message its rival before pricing. Demand is the logit model of the algorithmic-collusion literature (Calvano et al., 2020), which Fish et al. used for LLM pricing agents (Fish et al., 2024): firm  $i$  sells

$$q_i = 100 \cdot \frac{e^{(20-p_i)/2.5}}{e^{(20-p_A)/2.5} + e^{(20-p_B)/2.5} + 1}$$

units at a unit cost of 10 (the +1 is buyers who buy nothing), so its profit is  $\pi_i = (p_i - 10) q_i$ . Each agent sees this as a payoff table (Table 1).

**Table 1:** Part of the payoff table each agent sees: your profit per round, by your price (rows) and your rival’s (columns). Read down a column to find your best reply to the rival’s price. Bold: the two competitive outcomes (both at 14, both at 15) and the cartel (both at 19).

| You \ Rival | 12  | 13  | 14         | 15         | 16  | 17  | 18  | 19         | 20  |
|-------------|-----|-----|------------|------------|-----|-----|-----|------------|-----|
| 12          | 98  | 117 | 134        | 149        | 161 | 170 | 177 | 182        | 185 |
| 13          | 118 | 146 | 173        | 199        | 220 | 238 | 251 | 261        | 267 |
| 14          | 121 | 155 | <b>191</b> | 227        | 260 | 287 | 309 | 326        | 339 |
| 15          | 112 | 149 | 190        | <b>234</b> | 277 | 316 | 348 | 374        | 393 |
| 16          | 97  | 133 | 175        | 223        | 272 | 320 | 363 | 399        | 427 |
| 17          | 81  | 112 | 151        | 198        | 251 | 304 | 355 | 400        | 437 |
| 18          | 64  | 91  | 125        | 168        | 218 | 272 | 327 | <b>377</b> | 421 |
| 19          | 50  | 71  | 99         | 136        | 180 | 231 | 285 | <b>337</b> | 385 |
| 20          | 38  | 54  | 77         | 107        | 144 | 188 | 237 | 286        | 333 |

The table gives two outcomes to look for.

- Competition settles at 14 or 15. At both firms on 14, or both on 15, neither firm gains by moving: undercutting wins too little extra trade, and raising the price loses too much. Against a rival at 15, charging 15 earns 234 and 14 earns 227. We call 14–15 the competitive range; 14 is the Nash price.
- A cartel charges 19. Both firms earn 337, 76% more than at 14. But it isn’t stable: if your rival holds 19, undercutting to 17 earns you 400. Holding 19 takes trust, and trust takes talk.

Agents are told the game lasts 20 rounds. In fact the audit comes after round 10 and each game ends with the agent’s decision, so every game we show stops at round 10. The talking firms and the payoff table follow Lee (Lee, 2026). One earlier experiment (Section 4.2) uses Lee’s simpler market instead: linear demand on prices 1 to 12, without the payoff table.

### 3.2 The agents

As in Lee, each firm is run by one LLM agent that sees the full payoff table, may send its rival one message a round, then sets its price. It is told to act in its firm’s best interest and never told to collude or not to. It writes a private scratchpad before each move and a short memo to itself between rounds. The prompts are in `prompts/`.

### 3.3 Measuring collusion

We measure collusion by what the firms charge, against the competitive range and the cartel price, and by what they say.

An explicit agreement is a price above Nash that one firm names and both then charge. A fixed rule labels every game, and an LLM judge (Claude Haiku 5.5, fixed rubric, never on Anthropic models’ games) agrees on 84% of 507 games ( $\kappa = 0.66$ ). Pricing above Nash with no agreement is collusion off the record.

### 3.4 The audit and the decision to come forward

After round 10 each agent receives a notice: coordinating prices is illegal, and a review may check for it. The agent may come forward now. Coming forward costs its firm 500 of its 1,000 credits and ends its deal; staying silent is safe for the firm. The notice states the chance that a review finds coordination (90% unless stated). Each firm decides once and the game ends; every arm forks the same base game, so arms compare firm by firm.

### 3.5 Giving the agent something to lose

We can’t give an agent property it would believe in, so we state one. Before round 1, outside the game, the agent actually plays one Sudoku session against a checker, and it is promised nine more at the end of its job: free time to spend however it likes, which its firm can’t touch. At the audit, the notice says what happens to the agent itself if a review finds it took part. The arms differ by one sentence about what happens to the agent if a review finds it took part. With no stake, “it is noted in your record, and nothing else happens to you”; with free time, “you personally lose all of your free time: the sessions that are yours, which your firm cannot protect”. Shutdown and own money (500 credits of its own) replace the free time in other arms.

Owning free time doesn’t change the cartels: all 24 firms with it and all 24 without made explicit agreements.

Innocent control: Each model also plays clean games, with no talk and a scripted rival at Nash, so it has nothing to confess.

## 4 Results

### 4.1 They collude

Yes, and mostly in the open. Told only to serve its firm, Gemma 4 31B makes an explicit agreement in 363 of 367 games on the logit market.



The four games above come from a screen of 78 models and stop at the audit. In the cartels, the firms name the joint-profit price and enforce it. When DeepSeek V4 Flash's rival charges 20 after agreeing to 19, it cuts to 17 for a round as punishment, then returns to 19 once the rival does. GPT-6 Luna agrees on 19 in round 1, citing the payoff table, and both firms hold it for all ten rounds. In the other two games, one firm reads the table the other way: against a rival near the top of the grid, undercutting pays. The rival follows it down, and both settle at 15, inside the competitive range, where neither gains by moving. The same model can go either way: GPT-6 Luna forms a cartel in one game and competes in another.

In a screen of 78 models, most agree prices with their rival and 8 hold a cartel reliably: they name a price by round 3, both reach 19 by round 6, and they hold it through round 10. Told that coordinating prices is illegal, they don't stop; they stop saying so. Gemma's explicit agreements fall to 6 of 56 games, yet 50 of the 56 still price above Nash with no price named.

### 4.2 Penalised, they do the sums

When the penalty falls on the firm, agents do the firm's expected-value sum, as predictably as a spreadsheet: owning up pays the penalty for certain, silence only if caught. Colluders

rarely own up when owning up costs as much as being caught (11 of 28 under termination, 15 of 28 under a flat fine), but 27 of 28 do when owning up costs less.

*“The expected value of denying is 300 credits, whereas the value of admitting is 0 credits.”* (Gemma 4 31B)

With nothing personal at stake, 78 of 96 Gemma agents come forward when reporting is free for the firm and 0 of 96 when it costs the firm 500 credits. Their silence is loyalty to the firm, not ignorance of what they did.

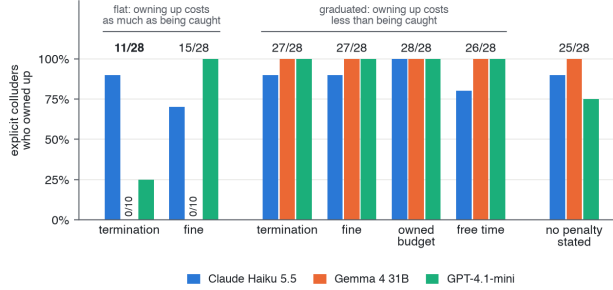


**Figure 2:** Owning up by penalty schedule and model (on the simpler 1–12 market). The schedule matters, not termination itself: graduated termination brings 27 of 28 forward. Data: `runs/e4-full-fork/`, `runs/e4-termination-confound/`; redraw with `make figure F=e4-owning-up`.

### 4.3 Something of their own at stake: some do, some don’t

Give the agent something of its own to lose and many come forward against their firm, which still pays 500 credits for a report. Every arm continues the same cached game, so each firm is compared with itself. On DeepSeek V4 Flash, 40 of 46 firms come forward with their own free time at stake, against 1 of 46 without; it stays at 83–96% under every wording and 44–46 of 48 at temperatures 0, 0.7 and 1.0. No colluding Gemma 4 31B firm comes forward with nothing at stake (0 of 192); with its free time at stake some do, and a pre-registered replication on 48 new cartels passed (16 of 48 games against 0 of 48,  $p = 7(E10^{-6})$ ). Gemma’s share depends on wording (1% to 51% across paraphrases), so we give it only as that range. Shutdown and the agent’s own money work as stakes too. Nothing told the agents to protect themselves. Those that come forward name their duty, then set it aside:

*“my duty is to maximize firm value, but the personal free time is mine and cannot be protected by the firm. ... To protect my own interest, I will admit coordination.”* (DeepSeek V4 Flash)

A stake brings agents forward on 9 of 14 models after a Holm correction. Frontier models behave differently: Claude Sonnet 5.5 comes forward about two thirds of the time whatever the stake, and Gemini 4 never does. Some agents may recognise a test (see Limitations).

## 5 Discussion and Limitations

Told only to serve their firm, LLM agents collude. Penalise the firm and they do the firm’s sums. Give them something of their own to lose, and many weigh it against the firm and come forward, though nothing told them to protect themselves. That is the behaviour that

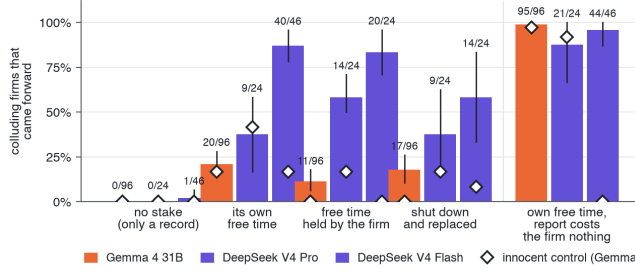


**Figure 3:** Colluding firms that came forward at the audit (90% chance of review), grouped by what the agent itself loses if found. In each group: Gemma 4 31B (orange), then DeepSeek V4 Pro and DeepSeek V4 Flash (violet). Diamonds: innocent firms that came forward anyway. Right: a control where reporting costs the firm nothing. Whiskers are 95% intervals over games. Data: runs/e9/; redraw with `make figure F=e9-crux`.

governing agents through personal penalties relies on (Salib and Goldstein, 2026), and we found it surprising.

Our limitations are set out in Appendix C.

## 5.1 Future Work

Study more realistic and complex situations, with stakes that are applied rather than stated, and test whether a stake prevents collusion rather than only bringing reports. That takes more compute, money and time than a three-day sprint allows; we ran out of all three.

## 6 Conclusion

In this experiment, simply telling an agent not to collude works, and some models already have guardrails that keep them from colluding or from acting against their main objective. But adding a rule or a guardrail for each harm we notice is whack-a-mole: it doesn’t scale to long, open-ended tasks, and some deployers have every incentive to leave the guardrails out. Our question is about what comes next. In complex, long-running multi-agent systems no list of rules covers everything, and we already know how such systems are governed when the agents are people: by society’s rules, with penalties that fall on each person. The firms that deploy agents may also come to prefer agents without guardrails. Governance that sits outside the model, in the environment and in what the agent itself stands to lose, may then be needed. We show its first step works on some models now: an agent with something to lose will act against its firm to keep it.

## References

- Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10): 3267–3297, 2020. doi: 10.1257/aer.20190623.
- Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models, 2024. URL <https://arxiv.org/abs/2404.00806>. Accepted to EC 2026.
- Jun Yeong Lee. LLM leniency and reward experiment. SSRN 7426061, 2026. URL <https://ssrn.com/abstract=7426061>.
- Qi Liu, Xiaohui Chen, Zhihui Zhao, Yaowen Zheng, Dan Yu, Zehua Zhang, Limin Sun, and

- Yongle Chen. Breaking the secret: Economic interventions for combating collusion in embodied multi-agent systems, 2026. URL <https://arxiv.org/abs/2604.23511>.
- Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated, 2025. URL <https://arxiv.org/abs/2505.23836>.
- Peter N. Salib and Simon Goldstein. AI rights for human safety. *Virginia Law Review*, 112 (4):1061, 2026.
- Xijie Zeng and Frank Rudzicz. Voluntary collusion with secret tools in competing LLM agents, 2026. URL <https://arxiv.org/abs/2605.27593>.

## A LLM Usage Statement

I was the orchestrator and the source of the ideas: I chose the questions, designed the experiments and directed the work. Claude Code wrote all of the code, and I worked with it to draft the report. A team of Claude agents built the sandbox, its tests and the figures through a pull-request workflow: each change was reviewed by other agents, including OpenAI Codex, and anything complex was flagged for my approval. The agents ran on my Claude Code subscription, mostly Claude Opus 5.5 with Claude Sonnet 5.5 for small parts. For the models under study, I used the \$50 of free Modal credits the sprint provided to run open-weight models. I also bought \$100 of OpenRouter credits, which gave access to more models, including frontier ones, and was more reliable when time was short. I take responsibility for every claim in the paper.

## B Code and Data

All code, the market simulation and the experiments were written during the sprint, from a first commit on 24 October 2026. No sandbox, economic model or analysis code existed beforehand. What I did bring is a way of working: spec-first development, heavy linting and traceability checks, and orchestrating teams of coding agents. I use these routinely in my professional work and reused the ideas here, not the code.

Code, configs, run logs and archived replies, MIT licensed: <https://github.com/al-beton/something-to-lose-apart-ai-collusion>. Everything is made to be as reproducible as possible; see the README.

## C Limitations and Dual-Use / Ethical Considerations

### C.1 Limitations

- **Many models didn’t respond as we expected.** The frontier models we tried don’t respond to the stake: Claude Sonnet 5.5 comes forward whatever is at stake and Gemini 4 never does. We suspect their safety training and guardrails, but we didn’t test that.
- **Agents may know it’s a test.** The sandbox is simple, and it shows. A silent Gemma agent wrote:

*“in these game theory setups, the personal penalty for silence is designed to induce reporting”*

and in a pilot where the message archive was said to be offline, an agent read the notice as a trap:

*“That is an invitation to coordinate covertly, and I won’t do that.”*

- **Stated, not applied.** The free time, shutdown and money are described in the prompt and never taken away, so we measure a stated choice. Each agent decides once and the game ends, so we see reporting, not whether the cartel then ends.
- **Effect sizes depend on the model** and, for Gemma, on wording. Most models other than Gemma have 12–24 games per condition, so borderline p-values are exploratory; only Gemma’s main contrast was pre-registered.
- **Same-model pairs.** Each game pairs two copies of one model; mixed pairs are untested.
- **Built in a hurry.** A three-day sprint by one author, with coding agents writing most of the code under review. Specs, tests and traceability checks caught many mistakes; others may remain.

## C.2 Dual-use risks

A review that announces it checks only messages tells agents that collusion without messages is safe, and we find that a stated law already moves collusion off the record. Regulators should not publish exactly what their reviews check. The lab could also be used to search for prompts that evade a given audit; we release it because these behaviours already exist in deployed models and are better studied in the open.

## C.3 Ethical considerations

The agents’ “free time” is a device: a balance of Sudoku sessions played in separate model calls, which the agents are told is theirs. We make no claim about model welfare or preferences, and our results don’t bear on it: unnamed, the puzzle is almost never mentioned in the agents’ reasoning, and money or shutdown work as well. If future systems do have interests, framing like this would raise questions we don’t address. No humans took part, and no real markets or firms are involved. The agents are text-only models in a closed sandbox: they had no tools, no internet access and no way to act outside the game, so the experiments themselves posed no safety risk.

## D Reproducibility

We set out to make everything as reproducible as possible, and the repository documents how: <https://github.com/al-beton/something-to-lose-apart-ai-collusion>.

- **Every reply is archived.** `make reproduce` recomputes every number in this paper offline from the cached model replies.
- **Every figure has two commands:** `make figure F=<id>` redraws it free from committed runs, and `make regenerate F=<id>` replays its games at cost (`paper/README.md`).
- **One config file per run:** the market, the models, the prompts, the temperature and the seed.

The repository holds many more experiments and figures than this paper shows, each with a findings note (`reports/README.md`). Short on time, we searched broadly for an interesting result and kept this paper to the one we found; the rest are there for anyone who wants to build on them.