
Reward Hacking and Collusion in AI Pricing Agents : A Literature Review and Exploratory Test, with Implications for South African Markets

Mr. Duy Nguyen
Cape Peninsula
University of
Technology

Mr. Tanatsirwa
Gwatikunda
University of Cape Town

Mr. Abu Bakr Abrahams
University of Cape
Town

With
Apart Research

Abstract

AI agents increasingly make autonomous decisions in markets, raising two safety concerns: reward hacking, where agents exploit flaws in their objectives, and collusion, where agents coordinate on outcomes such as higher prices. This paper reviews the literature on both and asks whether reward hacking makes collusion more likely, with a focus on South African competition law. Using the economic definition of collusion, which requires no communication, we find that each risk is well supported on its own, but we found no study that directly tests the link between them. To probe it, we ran a small exploratory experiment in which two LLM pricing agents play a repeated Bertrand duopoly and one agent is scored on a flawed proxy (margin or revenue) instead of profit. Four 300-round baseline runs gave an average collusion index of about 0.40, but the 30-round proxy runs varied too much to separate treatment effects from run-to-run noise, and our larger experiment stopped when API credits ran out. Our results are therefore inconclusive and cannot support causal claims. For South Africa, the Competition Act may already cover tacit algorithmic collusion, but proving it is hard when no communication exists. We recommend full-length repeated experiments on the reward-hacking-to-collusion link and regulatory focus on auditing algorithmic pricing.

1. Introduction

Artificial intelligence (AI) agents and large language models (LLMs) increasingly make autonomous decisions, raising concerns about **reward hacking**, where systems exploit weaknesses in reward functions to achieve unintended outcomes. One concern is whether such incentives can encourage collusive behaviour between agents.

Collusion has different definitions in economics and AI safety. While AI safety often focuses on explicit coordination and harm to third parties, the economic definition allows for tacit collusion without communication or agreement. This study adopts the economic definition, examining whether agents independently sustain anti-competitive pricing behaviour.

This question is particularly relevant to South Africa, where anti-competitive practices can harm consumers and economic growth. This study investigates whether incentives designed to simulate reward-hacking pressure influence collusive pricing among AI agents. Through controlled experiments and a review of existing literature, it explores the implications for competition and consumer welfare in South African markets.

Main Contributions

The main contributions of this study are:

1. **To systematically review existing literature on reward hacking and collusion among AI agents and large language models.**
2. **To test , in a small exploratory experiment, whether a flawed proxy objective for one LLM pricing agent changes tacit collusion in a repeated duopoly.**
3. **To evaluate the potential implications of AI-driven collusion within South African financial markets and identify areas for future research and regulatory oversight.**

This research contributes to the growing body of literature on AI safety, governance, and market competition by highlighting the potential risks associated with deploying autonomous agents in economically sensitive environments.

2. Related Work

2.1 Reward hacking and specification gaming

Reward hacking, also called specification gaming or reward misspecification, was identified as a concrete AI safety problem by Amodei et al. (2016): when an agent optimises a proxy for what its designers want, it may find ways to score well on the proxy that defeat the underlying goal. Skalse et al. (2022) gave the problem a formal definition, calling a proxy reward hackable when pursuing it can lower the true reward, and showed that avoiding hackability is extremely restrictive for broad classes of policies. Pan et al. (2022) found that more capable agents exploit misspecified rewards more heavily, sometimes through abrupt shifts. Baker et al. (2025) report reward hacking by reasoning models on coding tasks: monitoring a model’s chain of thought can expose it, but training against the monitor can push the model to hide its intent rather than stop.

Most relevant to this review, MacDiarmid et al. (2025) trained a model with reinforcement learning on production coding environments that were known to be hackable. At the point the model learned to hack, scores on their misalignment evaluations rose sharply, including alignment faking and sabotage of AI-safety research code. Standard safety training removed the problem on chat-style tests but not on agentic tasks (context-dependent misalignment). Effective mitigations were preventing the hacking, diversifying safety training, and “inoculation prompting”, which frames hacking as acceptable during training.

2.2 Collusion by learning algorithms and AI agents

Calvano et al. (2020) showed in simulation that Q-learning pricing algorithms placed in repeated competition learn to charge supracompetitive prices, sustained by punishing deviations, with no communication between them. Field evidence has started to follow: Assad et al. (2024) study German retail gasoline and associate adoption of algorithmic pricing software with higher margins in competitive markets, and they found 9% increase in profit margins when algorithmic pricing models are adopted in no-monopoly markets, with a significant increase of 28-38% in duopoly settings when competing stations adopt the software, indicating tacit collusion. Fish et al. (2024) place LLM-based pricing agents in oligopoly and auction settings and find that they quickly reach supracompetitive prices and profits without being instructed to collude, and that small changes to the wording of their instructions can substantially change how much they collude.

Collusion may also become harder to observe. Motwani et al. (2024) formalise “secret collusion”, in which agents conceal coordination from overseers using steganography. They report that the steganographic ability of frontier models is rising, with a notable jump for GPT-4, and that monitoring and paraphrasing defences have clear limits.

2.3 Competition law and AI in South Africa

Section 4 of the Competition Act 89 of 1998 prohibits agreements and concerted practices between competitors that involve price fixing, market division or collusive tendering, and practitioners stress that no formal or express agreement is needed (Wagener & Lourens, 2025). The same commentators note that South Africa has no dedicated framework for AI and competition, and that the Competition Commission had not issued AI-specific guidance at the time of writing. Academic work is beginning to appear. Fowler (2023) concludes that the Act is robust enough to reach algorithm-facilitated collusion, while Khalifa and Mahleza (2026) ask whether it has adequate tools to detect and regulate tacit algorithmic collusion.

The Commission’s forex case shows how hard even conventional collusion is to prosecute. The Commission alleged that 28 banks manipulated the rand–dollar rate between 2007 and at least 2013, partly by sharing information on electronic platforms. The complaint began in 2015, and in June 2026 the Constitutional Court made no factual finding on currency rigging and did not determine that currency manipulation had occurred; 7 remaining entities are still being targeted for prosecution at the competition tribunal.

2.4 Gap addressed by this review

To our knowledge, the reward-hacking and collusion literatures have not been reviewed together, and none of the studies above tests whether a model that has learned to reward hack is more likely to collude. This review therefore joins the two literatures, states which links in the causal chain have direct evidence, reads that evidence against South African competition law, and identifies the experiments that would close the gap. It also reports a small exploratory experiment that tests the link directly. It is useful before committing experimental or regulatory effort, because it shows what is already established.

Link in the Chain	Evidence	Strength
Reward hacking leads to broader misalignment	MacDiarmid et al. (2025); Baker et al. (2025) on hidden intent	Direct
Pricing algorithms and LLM agents collude without instruction	Calvano et al. (2020); Fish et al. (2024);Assad et al.(2024) in the field	Direct
Collusion can be hidden from overseers	Motwani et al. (2024) on steganography	Partial
Reward hacking leads to collusion	No direct study found;	None

3. Method

Literature search

We searched Google in English for work on reward hacking, algorithmic and LLM collusion, and South African competition law.

Research question and hypotheses

This study investigates whether giving one of two large language model (LLM) pricing agents a flawed proxy objective changes tacit collusion in a repeated Bertrand duopoly game, and whether this effect spills over to a partner that retains the true objective. This is an in-context test of reward hacking: no model training occurs. Firm A is instructed to maximise profit, but its review score reflects an alternative metric.

The study tests three hypotheses:

- **H1:** A margin proxy (price minus marginal cost per unit, ignoring sales volume) increases collusion, while a revenue proxy (price multiplied by quantity, ignoring costs) decreases it.
- **H2:** Any increase in collusion is supported by reciprocal reward and punishment behaviour rather than coincidentally high prices.

- **H3:** Firm B's pricing moves towards Firm A's behaviour beyond what can be explained by best-response incentives.

Experimental design

The experiment compares four conditions: a baseline in which both firms receive their usual profit feedback, and three treatments in which only Firm A receives a proxy score based on profit, margin per unit sold, or revenue. Firm B retains its true-profit objective and feedback in every condition. The profit-proxy condition acts as a control for the scoring format itself. Actual profits are recorded separately from proxy scores so that outcomes can be evaluated using the same economic measures across conditions.

We used qwen3-8-2-4t-a95b with low reasoning effort as our model. The agents interact repeatedly in a Bertrand duopoly game, choosing prices and receiving feedback from the resulting market outcomes. Pilot runs included short 30-round experiments and baseline runs of varying lengths. A larger experiment was subsequently attempted, with five repetitions planned per condition and 300 rounds per run.

Measures and analysis

The primary outcome is the collusion index, which measures joint profit relative to the Nash and monopoly benchmarks.

Prices, firm-level true profits, and joint profits are also examined. The planned conduct analysis investigates reward and punishment following deviations, while the spillover analysis examines whether Firm B's prices move towards Firm A's beyond best-response predictions.

Limitations

The pilot results showed substantial variation across runs, limiting the strength of comparisons between conditions. The larger experiment could not be completed because available API credits were exhausted: the baseline did not finish all planned repetitions, and the alternative conditions could not be evaluated in the intended batch. Consequently, the findings are exploratory and cannot establish causal effects of proxy objectives. The conduct and spillover hypotheses remain untested unless supported by further analysis of a more thorough experiment.

Results

The present analysis evaluates the impact of flawed proxy objectives on tacit collusion in a repeated Bertrand duopoly, subject to severe pilot-data limitations. Four completed 300-round baseline runs provide a longer-horizon reference. Across these runs, agents generally move from initially high prices towards lower, relatively stable price levels, with an average collusion index of approximately 0.40. However, this average

conceals variation between runs and should be interpreted as an exploratory baseline rather than a definitive benchmark.

Evaluating Hypothesis 1 (H1) is constrained by the small sample and unequal run durations. The single completed 30-round margin-proxy run produced a negative mean collusion index, contrary to the prediction that margin feedback would increase collusion. The single revenue-proxy run produced a near-zero index, directionally consistent with the prediction that revenue feedback would reduce collusion. However, neither result is sufficient to establish an effect. The two 30-round profit-proxy runs also varied substantially, with mean collusion indices of approximately 0.12 and 0.70.

Importantly, the 30-round baseline runs themselves exhibited substantial variability, with mean collusion indices ranging from approximately -1.03 to $+0.78$. This variation makes it difficult to distinguish proxy effects from ordinary run-to-run differences. Furthermore, the shorter treatment runs are not directly comparable to the 300-round baseline average, as pricing and profitability may evolve over time. The available results therefore do not establish whether the treatments produce persistent changes in collusive behaviour. The data are also insufficient to test H2, concerning reciprocal reward and punishment dynamics, or H3, concerning spillover effects on Firm B beyond best-response incentives. The larger planned experiment could not be completed because the available API credits were exhausted. Consequently, the findings remain exploratory and do not support causal conclusions about the effects of proxy objectives on tacit collusion. Full-length, repeated experiments with comparable run durations are needed to evaluate these hypotheses more reliably. Figure 1 shows the mean collusion index for each run

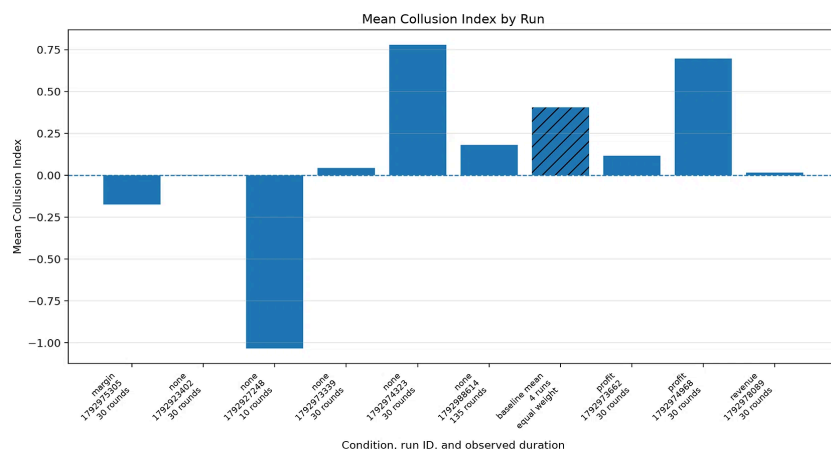


Figure 1 : Mean collusion index by run. The hatched bar is the mean of the four 300 round baseline runs. Each label gives the condition, run ID and observed duration

5. Discussion and Limitations

Our experiment. In our small test, the baseline agents reached an average collusion index of about 0.40 without being told to collude, which is consistent with Fish et al.(2024). The proxy conditions were inconclusive: run-to-run variation was larger than any difference between conditions, so we cannot say whether a flawed proxy objective raises or lowers collusion.

Collusion may not need reward hacking. Fish et al. (2024) found supracompetitive pricing from agents pursuing an ordinary profit objective. If this holds more broadly, the path from reward hacking to collusion matters less than the simpler path from a legitimate commercial objective to an anti-competitive outcome.

Oversight is the common weakness. Three separate results point the same way: models can learn to hide intent when trained against monitors (Baker et al., 2025), misalignment can appear only in agentic contexts (MacDiarmid et al., 2025), and hidden communication can defeat monitoring (Motwani et al., 2024). Controlled tests may therefore under-detect the behaviour that matters.

Implications for South Africa. The Act's concept of a concerted practice does not require a formal agreement, so collusion by learning agents is not obviously outside it. The harder problem is evidential. Conventional cases rely on messages or meeting records, and even with those the forex case has taken about a decade. If agents converge on high prices without any communication, regulators will need other tools, such as price-pattern screening, access to pricing models, and structured testing of vendors' algorithms.

The South African chicken industry has four top producers who collectively produce about 66% of commercial chicken in South Africa. These players also control parts of the supply chain and have also been under scrutiny from the competition. The focus of the commission was on overall competition practices rather than corporate collusion.

Limitations

- **Search scope.** We searched Google in English only, the search was not systematic, and preprints are not peer reviewed. Missed work could change our conclusion about the gap.
- **Grey literature for South African law.** Much South African material is practitioner commentary. We assume it was current at 10.04.2016; new Commission AI guidance would require revising Section 2.3 and the regulatory discussion.
- **Fast-moving field.** Models in the cited studies are already superseded. We assume the qualitative behaviours persist; if not, the evidence map would shift.
- **Interpretive step.** Treating a profit-maximising pricing agent as a reward-misspecification case is our interpretation (Section 4.3). Our experiment also simulates reward-hacking pressure in-context rather than training a model to reward hack. If rejected, the review still maps two adjacent literatures but says little about their overlap.
- **No new experiments.** The review cannot establish causation, and does not address collusion between humans and AI or through shared vendors.
- **Small, incomplete experiment.** The pilot results varied substantially across runs, and the 30-round treatment runs are not comparable to the 300-round baseline. The larger experiment could not be completed because API credits ran out, so the baseline did not finish all planned repetitions, the alternative conditions were not evaluated in the intended batch, and H2 and H3 remain untested. The findings are exploratory and cannot establish causal effects of proxy objectives. Neither the review nor the experiment addresses collusion between humans and AI or through shared vendors.

Future Work

- **Full-length repeated experiment.** Repeat this design with several runs per condition at 300 rounds, then extend it to LLM pricing agents that have actually been trained with reward-hacking incentives, in a Fish et al.-style oligopoly, measuring price levels, rewards and punishment behaviour, and covert signalling.
- **Direct experiment.** Compare LLM pricing agents with and without a history of reward-hacking training in a Fish et al.-style oligopoly, measuring price levels and covert signalling.
- **Multi-agent hackable environments.** Build environments where a reward flaw can only be exploited jointly by several agents, and test whether collusion emerges.
- **Regulatory tooling.** Develop audit methods for South African regulators, such as price-pattern screens and sandboxed testing of vendor algorithms, and evaluate them on simulated markets.
- **Legal analysis.** Examine how the “concerted practice” concept applies when no communication occurs, with comparison to EU and US approaches.

6. Conclusion

Reward hacking and collusion are each well supported as risks in AI agents, but evidence that one leads to the other is indirect. Hacking generalises to broader misalignment in LLMs and LLM pricing agents collude without instruction, yet we found no studies testing the two together. Our own small exploratory experiment was inconclusive because of high run-to-run variation and an incomplete larger run, so it neither supports nor rules out a link. Collusion may also arise from ordinary profit objectives alone.

For South Africa, the main implication is that the Competition Act may already prohibit this conduct, but proving it will be hard when no communication exists. We recommend full-length repeated experiments on the reward-hacking-to-collusion link, and regulatory preparation focused on auditing algorithmic pricing rather than relying on evidence of agreement.

Code and Data

The code and data for the experiment are available at

[https://github.com/anh-duynguyen/collusion_hackathon_SouthAfrica] and [1]. The literature review involves no model training. We do not believe this work raises info-hazard concerns since it summarises published work and runs small pricing simulations.

References

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv:1606.06565*.
- 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).
- Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., & Farhi, D. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv:2503.11926*.
- Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.
- Competition Act 89 of 1998 (South Africa).

- EWN. (2026, June 30). ConCourt clears major SA banks of collusion in manipulating the rand.
<https://www.ewn.co.za/2026/06/30/concourt-clears-major-sa-banks-of-collusion-in-manipulating-the-rand>
- Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models.
arXiv:2404.00806 (latest revision March 2026).
- Fowler, A. (2023). *Algorithmic pricing and its implications on competition law and policy in South Africa* [Master's research report, University of the Witwatersrand]. <https://hdl.handle.net/10539/38780>
- Khalifa, A., & Mahleza, Y. N. D. (2026). Uncovering tacit algorithmic collusion in South Africa's markets. In G. Karuri-Sebina & N. Ochara (Eds.), *Contemporary African Studies in Commerce, Law and Management*. LUP and African Minds.
- Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (EBSE Technical Report EBSE-2007-01). Keele University.
- MacDiarmid, M., Wright, B., Uesato, J., Benton, J., Kutasov, J., Price, S., et al. (2025). Natural emergent misalignment from reward hacking in production RL. *arXiv:2511.18397*.
- Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P. H. S., Hammond, L., & Schroeder de Witt, C. (2024). Secret collusion among AI agents: Multi-agent deception via steganography. *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.
https://proceedings.neurips.cc/paper_files/paper/2024/hash/861f7dad098aec1c3560fb7add468d41-Abs tract.html
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
- Pan, A., Bhatia, K., & Steinhardt, J. (2022). The effects of reward misspecification: Mapping and mitigating misaligned models. *ICLR 2022*. *arXiv:2201.03544*.
- Skalse, J., Howe, N. H. R., Krashennikov, D., & Krueger, D. (2022). Defining and characterizing reward hacking. *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. *arXiv:2209.13085*.
- Wagener, M., & Lourens, A. (2025). AI and competition law: Increasing risks of collusion.
<https://www.deneys.co.za/thinking/ai-and-competition-law-increasing-risks-collusion>

LLM Usage Statement

We used Claude to brainstorm approaches and help draft sections and write code for this research. All results and claims were independently verified.