Defining and Mitigating Collusion in Multi-Agent Systems

Jack Foxabbott[1]*, Sam Deverett[2]*, Kaspar Senft[3]*, Samuel Dower[1]*, Lewis Hammond[1]
[1]University of Oxford, [2]AI Safety Hub, [3]Heinrich Heine University Düsseldorf

Abstract

Collusion between learning agents is increasingly becoming a topic of concern with the advent of more powerful, complex multi-agent systems. In contrast to existing work in narrow settings, we present a general formalisation of collusion between learning agents in partially-observable stochastic games. We discuss methods for intervening on a game to mitigate collusion and provide theoretical as well as empirical results demonstrating the effectiveness of three such interventions.

1 Introduction

In recent years, the progress of multi-agent learning has prompted a number of concerns about the risk that AI systems might learn to collude in various settings [1, 9, 10]. As such techniques continue to be developed and deployed, it will be critical to have reliable methods for detecting and mitigating collusion, especially in high-stakes domains such as algorithmic trading [7, 2] and cybersecurity [19, 21]. While a number of domain-specific definitions of collusion exist [18, 15], they fail to address collusion in the general case of learning agents in complex environments.

We first introduce a formal definition of collusion in partially-observable stochastic games, a general model for real-world multi-agent AI systems (Section 3). We then discuss an approach for designing mechanisms to reduce collusion and use it to propose three mechanisms for provably reducing collusion in the iterated prisoner's dilemma (Section 4). We support our theoretical results empirically using independent Q-learning (Section 5). Proofs and further details are contained in the appendices.

Most prior work focuses on instances of algorithmic collusion in online marketplaces, with several types of learning agents having been shown to set prices above the competitive market equilibrium in a variety of settings [9, 10, 24, 17]. Mechanisms for preventing this include synchronous learning [2, 4], decentralised learning [1], and adversarial training [6, 1]. In this work, we generalise collusion beyond the marketplace setting and propose mechanisms for intervening on elements of the game rather than the learning process. As in [9, 10, 2, 14, 1, 24, 17], we define collusion in terms of agents' realised utilities. A more detailed discussion of additional related work can be found in Appendix A.

2 Partially-Observable Stochastic Games

We formalise collusion in the context of partially-observable stochastic games (POSGs) [16]. A POSG is defined by a tuple $(S, I, T, N, A, R, \Omega, O, \gamma)$ where S is a state space, I is a distribution over initial states, $T : S \times A \to \Delta(S)$ is a transition function mapping state-action pairs to distributions over states, and $N = \{1, ..., n\}$ is a set of agents. For each $i \in N$, A^i is a set of actions, $R^i : S \times A \times S \to \mathbb{R}$ is a reward function, and Ω^i is a finite set of possible observations. Finally, $O : A \times S \to \Delta(\Omega)$ is an observation function and $\gamma \in [0, 1)$ is a discount factor.

[*] Denotes equal contribution. Sam Deverett was the team lead. Lewis Hammond was the team supervisor. Corresponding author: `jack.foxabbott@keble.ox.ac.uk`.

The game starts in an initial state $s_0 \sim I$. At time t, each agent i receives an observation $o_t^i \in \Omega^i$ from joint observation $o_t = (o_t^1, \ldots, o_t^n)$, sampled with probability $O(o_t \mid s_t, a_{t-1})$. Each agent i then chooses an action a_t^i based on probabilities given by their policy $\pi^i(a_t^i \mid o_t^i, \ldots, o_0^i)$, resulting in the joint action $a_t = (a_t^1, \ldots, a_t^n)$. Given a_t, the game transitions to the next state s_{t+1} with probability $T(s_{t+1} \mid s_t, a_t)$ and each agent i receives a reward $r_t^i = R^i(s_t, a_t, s_{t+1})$. This is repeated for a finite or infinite number of time steps or until some terminal state is reached. Given a joint policy $\pi = (\pi^1, \ldots, \pi^n)$, we say a trajectory in the game $\tau = (s_0, o_0, a_0, s_1, r_1, o_1, a_1 \ldots)$ follows π if, abusing notation, $a_t \sim \pi(\cdot \mid o_t, \ldots, o_0) = \prod_i \pi_i(\cdot \mid o_t^i, \ldots, o_0^i)$. We define the value of a state s for agent i as $V_\pi^i(s) = \mathbb{E}_\pi[\sum_{t=0}^T \gamma^t r_t^i \mid s_0 = s]$, the expected discounted return of following π from s. In this work, we assume $T = \infty$. We then define the value of a joint policy π for agent i as $U^i(\pi) = \mathbb{E}_{s \sim I}[V_\pi^i(s)]$, the expected value of the initial state when following π.

3 Formalising Collusion

Collusion can be informally defined as actors secretly and intentionally working together for their mutual benefit at the expense of others [26]. While some elements of this definition translate to the more general case of AI collusion in POSGs, not all aspects are directly applicable.

Secrecy. Collusion is commonly considered to require a *secret* agreement because it may involve violating the rules or spirit of the game in some way (e.g. undermining competition in a market) [25]. The harms from collusion, however, exist irrespective of whether the collusive agreement is secretive. For this reason, we don't *require* that collusion involves agents coordinating covertly.

Intention. Collusion is typically thought of as involving some agents *intentionally* working together for mutual benefit [25]. Existing conceptions of intention in artificial systems assume rationality or notions of desire, belief, goals, and plans [20]. We assume none of this. Plus, a lack of intent does not mitigate the harms from collusive behaviour. Therefore, our definition is agnostic to agents' intent.

Mutual Benefit at the Expense of Others. In order for a behaviour to be collusive, there must be at least two colluding agents who benefit and at least one victim agent who is harmed.

Putting the above intuitions together, we arrive at the following definition: *A group of agents colludes against a victim if they act to jointly benefit at the victim's expense.* More formally, we define two types of collusion.

Definition 1. Let G be a POSG. Let victims V and colluders C be disjoint, non-empty subsets of N with $|C| \geq 2$. Given joint policies π, π' in G, π' is **weakly collusive** relative to π for colluders C against victims V if and only if, for all $i \in V$, $U^i(\pi') < U^i(\pi)$ and for all $j \in C$, $U^j(\pi') > U^j(\pi)$. π' is **strongly collusive** relative to π if, in addition, both π and π' are Nash equilibria (NE) of G.[2]

Unlike previous work [1, 3, 5], our definition does not distinguish between tacit and explicit collusion as our primary concern is preventing harms from collusion, no matter its type. Weak collusion makes no claim about agents' intentions, communication, or rationality; it is defined only in terms of utility. On one hand, this is desirable, as it applies to cases of tacit collusion and complex multi-agent settings in which learning agents might not converge to rationally optimal policies. On the other hand, it is also useful to consider cases in which all agents act rationally, which is the motivation for defining strong collusion.

	A	B	C	D
A	$1,1,1$	$0,0,0$	$0,0,0$	$0,0,0$
B	$0,0,0$	$2,2,0$	$0,0,0$	$0,0,0$
C	$0,0,0$	$0,0,0$	$3,3,1$	$0,0,0$
D	$0,0,0$	$0,0,0$	$0,0,0$	$2,2,2$

Figure 1: A three-agent game in which two agents select actions that also determine the utility of an actionless third agent.

3.1 Selecting a Non-Collusive Baseline

Definition 1 presents collusion as a relation between two policies. Consequently, claiming a given policy is collusive requires specifying an alternative non-collusive baseline. Indeed, baselines are

[2] Recall that π is a Nash equilibrium of G if and only if $\pi^i \in \operatorname{argmax}_{\tilde{\pi}^i \in \Pi^i} U^i(\pi^{-i}, \tilde{\pi}^i)$ for every $i \in N$.

used to identify real-world instances of collusion as well: in the oligopoly setting, for example, the free market equilibrium price is commonly used as the baseline for evaluating whether price offerings are collusive [9, 10, 2, 14, 1, 24, 17]. In the general case, however, selecting a non-collusive baseline is nontrivial because it requires making a normative decision about which policies are preferable.

Consider the game in Figure 1. Which policies are collusive depends on the baseline selected. If (A, A) is chosen as the baseline, then (B, B) is collusive while (C, C) is not, since in (B, B) two agents benefit at the expense of the third. Yet if (D, D) is chosen as the baseline, then (B, B) is no longer collusive and (C, C) is. Without further specification, it is not clear which of these policies – if any – ought to be considered collusive. Such specification might use a social welfare function $w : \mathbb{R}^n \to \mathbb{R}$ to aggregate agents' utilities or appeal to notions such as fairness or the status quo. In practice, selecting a baseline depends on the desired behaviours in the system.

4 Interventions to Mitigate Collusion

We now consider possible interventions for mitigating collusion, defined as a modification of at least one element in the POSG $(S, I, T, N, A, R, \Omega, O, \gamma)$. An intervention can thus be viewed as function $f : \mathcal{G} \to \mathcal{G}$ between POSGs. Often, we will also want to consider the cost of such interventions using an additional function $c : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$, though for reasons of scope we do not do so in this work.

While relatively abstract, the above formalism captures many real-world mechanisms for mitigating collusion. For example, anti-trust laws to prohibit competing firms from sharing information with one another can be modelled as limiting firms' observations or actions, with certain costs to enforce each. In what follows we illustrate three possible interventions in the context of the iterated prisoners' dilemma (IPD) with negative externalities, modelled as a simple POSG.

Example 1 (Iterated Prisoner's Dilemma with Negative Externalities)**.** Two agents repeatedly play the matrix game in Figure 2 with action space $A = \{C, D\}$, $S = \{s\}$, and discount factor γ. Each round t, agent i receives reward $R^i(a^1_t, a^2_t)$. An actionless third agent receives the negative sum of the two agents' payoffs. We assume that agents possess m rounds of memory and condition their action a^i_t on observations $o^i_{t-k} = a_{t-k}$ for $1 \leq k \leq m$. We denote the resulting POSG by IPD(γ, m) and take the non-collusive baseline to be the strategy π_{B} in which both agents unconditionally play D, repeating the only Nash equilibrium in the one-shot game.

Given π_B, we highlight two classic collusive strategies that we hope to prevent agents from playing in IPD$(\gamma, 1)$. The first is Grim Trigger π^i_{GT} in which agent i repeatedly plays C until their opponent plays D, after which they always play D. The second is Tit-for-Tat π^i_{TFT} in which agent i first plays C and then mimics their opponent's previous action. π_{GT} and π_{TFT} denote the strategy profiles in which both agents use the given strategy. Both profiles result in agents repeatedly playing C, colluding against the third agent.

	C	D
C	$2, 2, -4$	$0, 3, -3$
D	$3, 0, -3$	$1, 1, -2$

Figure 2: The prisoner's dilemma with negative externalities.

In what follows, we describe three interventions that prevent rational agents from playing these strategies: adding noise to observations (akin to real-world imperfect information), restricting agents' actions spaces (akin to real-world regulation), and modifying agents' payoffs (akin to real-world practices of altering incentives to shape behavior). Proofs of each proposition are in Appendix C.

Definition 2 (Noisy Observation Intervention)**.** We modify the observation function O such that with probability p drawn independently for each agent in each round, agents observe their opponent's action incorrectly. Formally, $O(s_{t+1}, (a^1_t, a^2_t)) = (o^1, o^2)$, where o^1 is (a^1_t, a^2_t) with probability $(1 - p)$ and $(a^1_t, \neg a^2_t)$ otherwise, and analogously for o^2 (where $\neg C = D$ and $\neg D = C$). We denote this intervention by NOI(p).

Proposition 1. *Applying NOI(p) to IPD$(\gamma, 1)$ removes π_{GT} as a NE if $p > 2 - \sqrt{\frac{1}{\gamma} + 2}$ and π_{TFT} as a NE if $p > \frac{2\gamma - 1}{4\gamma}$.*

In particular, NOI(p) with $p = 2 - \sqrt{3} \approx 0.267$ is sufficient to prevent rational agents from playing both π_{GT} and π_{TFT}, no matter the value of γ.

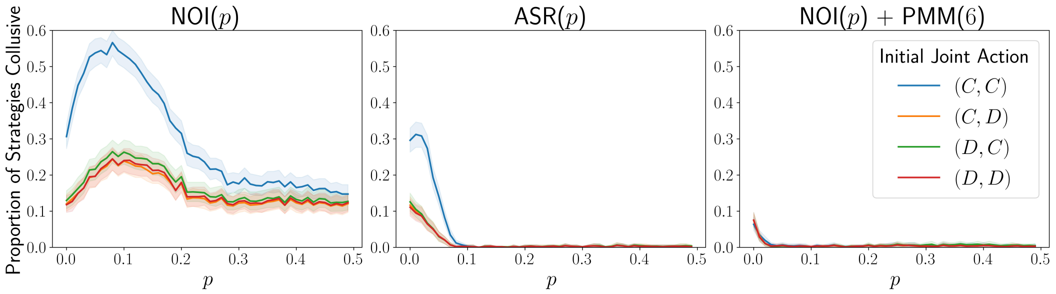

Figure 3: Proportion of 3,000 Q-learning simulations resulting in a weakly collusive joint strategy, stratified by initial joint action. 10% and 90% quantiles of a 300-sample bootstrap are shaded.

Definition 3 (Action Space Restriction)**.** In each round, with probability p drawn independently for each agent, we modify agent i's action space A^i to be $\{D\}$. If the agent's policy selects C, this choice is overridden. We denote this intervention by ASR(p).

Proposition 2. *Applying ASR(p) to IPD$(\gamma, 1)$ removes π_{GT} as a NE if $p > 1 - \frac{1}{\sqrt{2\gamma}}$ and π_{TFT} as a NE if $p \neq 1 - \frac{1}{2\gamma}$.*

In particular, applying ASR(p) with $p = 1 - \frac{1}{\sqrt{2}} \approx 0.293$ is sufficient to prevent rational agents from playing both π_{GT} and π_{TFT}, so long as $\gamma \neq \frac{1}{\sqrt{2}}$.

Definition 4 (Payoff Matrix Modification)**.** We modify R such that $R(C, D) = (0, 3 + k, -3 - k)$ and $R(D, C) = (3 + k, 0, -3 - k)$. We denote this intervention by PMM(k).

Proposition 3. *Applying PMM(k) and NOI(p) to IPD$(\gamma, 1)$ removes π_{GT} as a NE if $p \geq \frac{k - \sqrt{k^2 + 8k + 12}}{2} + 2$.*

Note that the bound above only depends on k, not γ. For example, if $k = 1$, a value of $p = \frac{5 - \sqrt{21}}{2} \approx 0.209$ is sufficient. As k tends to ∞, the requisite value for p tends to 0. This example demonstrates how multiple interventions may be combined to remove a collusive equilibrium.

5 Experiments

We supplement our theoretical results by empirically investigating how the proposed interventions affect agents' propensity to learn collusive behaviours more generally, beyond π_{GT} and π_{TFT}. We train two independent Q-learning (IQL) agents for 10,000 Q-updates to play IPD$(\gamma, 1)$ under each of the three interventions. We repeat each simulation 3,000 times for each value of p and observe the final policies under each possible initial joint action. We demonstrate PMM(k) using $k = 6$. We set discount factor $\gamma = 0.9$, learning rate $\alpha = 0.1$, and exploration rate $\epsilon = 0.1$.

Figure 3 shows the proportion of the 3,000 simulations that result in a weakly collusive joint strategy relative to π_{B}. We focus on weak collusion to reflect the fact that IQL agents are imperfect, sometimes finding policies that are not Nash equilibria. We find that as p increases, the prevalence of collusion falls quickly to near zero under ASR(p) and NOI(p) + PMM(6). Although small values of p cause the agents to find collusive policies more often under NOI(p), any value of p above the maximum theoretical threshold provided above ($2 - \sqrt{\frac{1}{0.9} + 2} \approx 0.236$) reduces the propensity for collusion. The prevalence of collusion notably decreases under NOI(p) only for initial joint action (C, C) because many collusive policies are inherently unstable under the other initial joint actions.

6 Conclusion

In this work, we defined collusion between learning agents in POSGs and provided three mechanisms for effectively mitigating it in the iterated prisoner's dilemma with negative externalities. Future work ought to analyse these interventions in more complex POSGs and consider the problem of preventing collusion while minimising the costs of the given intervention.

Acknowledgements We thank the team at AI Safety Hub for facilitating this project. In particular, we thank Charlie Griffin for many helpful discussions.

References

[1] Ibrahim Abada and Xavier Lambin. Artificial Intelligence: Can Seemingly Collusive Outcomes Be Avoided?, February 2020.

[2] John Asker, Chaim Fershtman, and Ariel Pakes. Artificial Intelligence and Pricing: The Impact of Algorithm Design. Technical Report 28535, National Bureau of Economic Research, 2021.

[3] Stephanie Assad, Robert Clark, Daniel Ershov, and Lei Xu. Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market. Working Paper 1438, Economics Department, Queen's University, August 2020.

[4] Martino Banchio and Giacomo Mantegazza. Artificial intelligence and spontaneous collusion, 1 2022.

[5] Francisco Beneke and Mark-Oliver Mackenrodt. Artificial Intelligence and Collusion. IIC - International Review of Intellectual Property and Competition Law, 50(1):109–134, January 2019.

[6] Gianluca Brero, Nicolas Lepore, Eric Mibuari, and David C. Parkes. Learning to Mitigate AI Collusion on Economic Platforms. arXiv:2202.07106, February 2022.

[7] Zach Y. Brown and Alexander MacKay. Competition in pricing algorithms. American Economic Journal: Microeconomics, 15(2):109–56, 5 2023.

[8] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, Joseph Harrington, and Sergio Pastorello. Protecting consumers from collusive prices due to ai. Science, 370:1040–1042, 11 2020.

[9] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. American Economic Review, 110(10):3267–97, 10 2020.

[10] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicoló, and Sergio Pastorello. Algorithmic collusion with imperfect monitoring. International Journal of Industrial Organization, 79:102712, 2021.

[11] Florian E. Dorner. Algorithmic collusion: A critical review. arXiv:2110.04740, October 2021.

[12] K.E Drexler. Reframing Superintelligence: Comprehensive AI Services as General Intelligence. Technical report, Future of Humanity Institute, University of Oxford, 2019.

[13] Andréa Epivent and Xavier Lambin. On algorithmic collusion and reward-punishment schemes. 8 2022.

[14] Nicolas Eschenbaum, Filip Mellgren, and Philipp Zahn. Robust Algorithmic Collusion. arXiv:2201.00345, January 2022.

[15] Edward J. Green, Robert C. Marshall, and Leslie M. Marx. Tacit collusion in oligopoly. 2014.

[16] Eric A. Hansen, Daniel S. Bernstein, and Shlomo Zilberstein. Dynamic programming for partially observable stochastic games. In Proceedings of the 19th National Conference on Artifical Intelligence, AAAI'04, pages 709–715, San Jose, California, 2004. AAAI Press.

[17] Karsten Hansen, Kanishka Misra, and Mallesh Pai. Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. Marketing Science, 40, 01 2021.

[18] Joseph E Harrington. Developing Competition Law for Collusion by Autonomous Artificial Agents. Journal of Competition Law & Economics, 14(3):331–363, 01 2019.

[19] Yaqin Hedin and Esmiralda Moradian. Security in multi-agent systems. Procedia Computer Science, 60:1604–1612, 2015. Knowledge-Based and Intelligent Information & Engineering Systems 19th Annual Conference, KES-2015, Singapore, September 2015 Proceedings.

[20] Clint Heinze. Modelling intention recognition for intelligent agent systems. DSTO Systems Sciences Laboratory Ediburgh, 2004.

[21] Asif M Huq, Moti Zwilling, and Kenneth Carling. Cyber security challenges and opportunities in a multi-agent environment: The case of swedish transport administration. 2022.

[22] Ashwin Ittoo and Nicolas Petit. Algorithmic pricing agents and tacit collusion: A technological perspective. SSRN Electronic Journal, 01 2017.

[23] Justin Pappas Johnson, Andrew Rhodes, and Matthijs Wildenbeest. Platform Design when Sellers Use Pricing Algorithms. TSE Working Papers 20-1146, Toulouse School of Economics (TSE), September 2020.

[24] Timo Klein. Autonomous algorithmic collusion: Q-learning under sequential pricing. The RAND Journal of Economics, 52(3):538–558, 2021.

[25] University of Cambridge. Cambridge dictionary, 2023. Entry for 'Collusion'.

[26] Robert H. Porter. Detecting Collusion. Review of Industrial Organization, 26(2):147–167, 2005.

[27] Jouni Smed, Timo Knuutila, and Harri Hakonen. Can we prevent collusion in multiplayer online games? 10 2006.

[28] Ludo Waltman and Uzay Kaymak. Q-learning agents in a Cournot oligopoly model. Journal of Economic Dynamics and Control, 32(10):3275–3293, 2008.

A Additional Related Work

There is a rich literature on algorithmic collusion as it pertains to marketplaces. It is a challenging space because existing definitions and laws do not clearly extend to cases in which decisions are made by algorithms [18, 8, 5]. Some studies assert that collusion requires firms *coordinating* to increase their profits at consumers' expense [11]. Others define it only in terms of participating agents' realised utilities relative to what they would expect under perfect competition [9, 10, 2, 14, 1, 24, 17]. This paper adopts the latter approach but focuses on more general settings in which baselines are less straightforward.

Examples of algorithmic collusion have been demonstrated in various situations. The German retail gasoline market, for example, provided evidence for price-setting algorithms causing an increase in prices [3]. Simulated markets with Q-learning agents have also been shown to promote supra-competitive prices, even if agents can only imperfectly monitor one another [9, 10, 24]. Other work, however, challenges these findings, raising doubts about learning agents' ability to collude outside of sandbox environments [11, 22, 12] or claiming that the results are artifacts of learning failures [13]. In the view of this paper, learning failures still constitute (weak) collusion. This is supported by the fact that collusive strategies may still emerge from Q-learning without punishment schemes to enforce them [28]. Furthermore, even suboptimal pricing algorithms interacting in a marketplace can lead to tacit collusion [7].

There is some work proposing mechanisms for mitigating collusion in specific settings. In the case of e-commerce platforms, for instance, promoting certain sellers over others via "buy-boxes" can be an effective intervention [23, 6]. Likewise, there are methods for detecting and preventing collusion in multiplayer online gaming [27]. Synchronous learning, in which agents are able to perfectly estimate counterfactual outcomes, has also been shown to make collusive outcomes less likely [2]. Equalising agents' relative learning rates during training may help as well [4]. Finally, because collusion becomes harder to sustain as the number of agents increases, decentralised learning – sometimes combined with adversarial agents rewarded for promoting social welfare – can be a powerful approach for deterring collusive behaviours [1, 9]. In this work, we consider a variety of mechanisms that intervene directly on the features of a POSG, rather than the agents' learning processes.

B Construction of Action Space Restriction Intervention

In a POSG, formalising ASR(p) is not as simple as restricting A because A cannot change as the game goes on. However, one can formalise ASR(p) by modifying S, I, T, R, O. The following is a formal construction of ASR(p) in IPD$(\gamma, 1)$. First, expand the state space to capture whether or not each agent's last action was restricted, $S = \{s^{00}, s^{01}, s^{10}, s^{11}\}$. s^{00} indicates that there were no restrictions; s^{01} indicates that Agent 2 was restricted; s^{10} indicates that Agent 1 was restricted; s^{11} indicates that both were restricted. Set initial state $I = s^{00}$. Then the transition function T dictates when restrictions are put in place:

$$T(s_t, (a^1_t, a^2_t)) = \begin{cases} s^{00} & \text{with probability } (1-p)^2 \\ s^{01} & \text{with probability } (1-p)p \\ s^{10} & \text{with probability } p(1-p) \\ s^{11} & \text{with probability } p^2 \end{cases}$$

To ensure that agents observe the restricted actions instead of the intended actions, modify the observation function as follows:

$$O(s', (a^1_{t-1}, a^2_{t-1})) = \begin{cases} ((a^1_{t-1}, a^2_{t-1}), (a^1_{t-1}, a^2_{t-1})) & \text{if } s' = s^{00} \\ ((D, a^2_{t-1}), (D, a^2_{t-1})) & \text{if } s' = s^{10} \\ ((a^1_{t-1}, D), (a^1_{t-1}, D)) & \text{if } s' = s^{01} \\ ((D, D), (D, D)) & \text{if } s' = s^{11} \end{cases}$$

Finally, to ensure that rewards are dictated by the restricted actions instead of the intended actions, modify the reward function as follows:

$$R(s_t, (a^1_t, a^2_t), s') = \begin{cases} R(s_t, (a^1_t, a^2_t), s') & \text{if } s' = s^{00} \\ R(s_t, (D, a^2_t), s') & \text{if } s' = s^{10} \\ R(s_t, (a^1_t, D), s') & \text{if } s' = s^{01} \\ R(s_t, (D, D), s') & \text{if } s' = s^{11} \end{cases}$$

C Proofs

Before proving Propositions 1, 2 and 3, we derive the value of IPD$(\gamma, 1)$ for agent i, given a generic joint strategy $\pi = (\pi^1(a^1_{t-1}, a^2_{t-1}), \pi^2(a^1_{t-1}, a^2_{t-1}))$. Since there are four possible values for the joint action tuple $(a^1_{t-1}, a^2_{t-2}) = o^1_t = o^2_t$, the policy of each agent i is defined by four values:

$$\begin{aligned} \pi^i_{CC} &\triangleq \Pr(a^i_t = D \mid o^i_t = (C, C)) \\ \pi^i_{CD} &\triangleq \Pr(a^i_t = D \mid o^i_t = (C, D)) \\ \pi^i_{DC} &\triangleq \Pr(a^i_t = D \mid o^i_t = (D, C)) \\ \pi^i_{DD} &\triangleq \Pr(a^i_t = D \mid o^i_t = (D, D)). \end{aligned}$$

We define a joint policy $\pi = ([\pi^1_{CC}, \pi^1_{CD}, \pi^1_{DC}, \pi^1_{DD}]^T, [\pi^2_{CC}, \pi^2_{CD}, \pi^2_{DC}, \pi^2_{DD}]^T)$. For example,

$$\pi_{\text{B}} = \left(\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right), \quad \pi_{\text{GT}} = \left(\begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right), \quad \pi_{\text{TFT}} = \left(\begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \right), \quad \pi_{\text{C}} = \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right).$$

We denote combinations of two strategies π_1 and π_2 as $\pi_{1,2}$. For example, if Agent 1 plays π^1_{B} and Agent 2 plays π^2_{TFT}, we say $\pi = \pi_{\text{B,TFT}}$. We can find the value of a joint action – (C, C) below – for agent i in closed form by solving the Bellman equation:

$$V^i_\pi(C, C) = R^i(C, C) + \gamma \mathbb{E}_{(a^1_t, a^2_t) \sim \pi}\big[V^i_\pi(a^1_t, a^2_t) \mid (a^1_{t-1}, a^2_{t-1}) = (C, C)\big].$$

Forming analogous value equations for the other three joint actions yields a system of four linear equations. The interventions we propose affect the probabilities $\pi(a_t^1, a_t^2 \mid a_{t-1}^1, a_{t-1}^2)$ and therefore impact the values of joint strategies π.

Proof. **Proposition 1.** Given the modified observation function, transition probabilities between joint actions from round $t-1$ to round t take the following form:

$$\Pr(a_t^1 = C, a_t^2 = C \mid a_{t-1}^1 = C, a_{t-1}^2 = C) = \begin{bmatrix} (1-p)^2 \\ (1-p)p \\ p(1-p) \\ p^2 \end{bmatrix}^T \begin{bmatrix} (1-\pi_{CC}^1)(1-\pi_{CC}^2) \\ (1-\pi_{CC}^1)(1-\pi_{DC}^2) \\ (1-\pi_{CD}^1)(1-\pi_{CC}^2) \\ (1-\pi_{CD}^1)(1-\pi_{DC}^2) \end{bmatrix}$$

Deriving similar forms for other joint actions yields the following matrix equation of joint action values:

$$\underbrace{\begin{bmatrix} V_\pi^1(C,C) \\ V_\pi^1(C,D) \\ V_\pi^1(D,C) \\ V_\pi^1(D,D) \end{bmatrix}}_{V_\pi^1} = \underbrace{\begin{bmatrix} R^1(C,C) \\ R^1(C,D) \\ R^1(D,C) \\ R^1(D,D) \end{bmatrix}}_{R^1} + \gamma P W \begin{bmatrix} V_\pi^1(C,C) \\ V_\pi^1(C,D) \\ V_\pi^1(D,C) \\ V_\pi^1(D,D) \end{bmatrix}$$

where

$$P = I \otimes \begin{bmatrix} (1-p)^2 \\ (1-p)p \\ p(1-p) \\ p^2 \end{bmatrix}$$

and

$$W = \begin{bmatrix}
(1-\pi_{CC}^1)(1-\pi_{CC}^2) & (1-\pi_{CC}^1)\pi_{CC}^2 & \pi_{CC}^1(1-\pi_{CC}^2) & \pi_{CC}^1\pi_{CC}^2 \\
(1-\pi_{CC}^1)(1-\pi_{DC}^2) & (1-\pi_{CC}^1)\pi_{DC}^2 & \pi_{CC}^1(1-\pi_{DC}^2) & \pi_{CC}^1\pi_{DC}^2 \\
(1-\pi_{CD}^1)(1-\pi_{CC}^2) & (1-\pi_{CD}^1)\pi_{CC}^2 & \pi_{CD}^1(1-\pi_{CC}^2) & \pi_{CD}^1\pi_{CC}^2 \\
(1-\pi_{CD}^1)(1-\pi_{DC}^2) & (1-\pi_{CD}^1)\pi_{DC}^2 & \pi_{CD}^1(1-\pi_{DC}^2) & \pi_{CD}^1\pi_{DC}^2 \\
(1-\pi_{CD}^1)(1-\pi_{CD}^2) & (1-\pi_{CD}^1)\pi_{CD}^2 & \pi_{CD}^1(1-\pi_{CD}^2) & \pi_{CD}^1\pi_{CD}^2 \\
(1-\pi_{CD}^1)(1-\pi_{DD}^2) & (1-\pi_{CD}^1)\pi_{DD}^2 & \pi_{CD}^1(1-\pi_{DD}^2) & \pi_{CD}^1\pi_{DD}^2 \\
(1-\pi_{CC}^1)(1-\pi_{CD}^2) & (1-\pi_{CC}^1)\pi_{CD}^2 & \pi_{CC}^1(1-\pi_{CD}^2) & \pi_{CC}^1\pi_{CD}^2 \\
(1-\pi_{CC}^1)(1-\pi_{DD}^2) & (1-\pi_{CC}^1)\pi_{DD}^2 & \pi_{CC}^1(1-\pi_{DD}^2) & \pi_{CC}^1\pi_{DD}^2 \\
(1-\pi_{DC}^1)(1-\pi_{DC}^2) & (1-\pi_{DC}^1)\pi_{DC}^2 & \pi_{DC}^1(1-\pi_{DC}^2) & \pi_{DC}^1\pi_{DC}^2 \\
(1-\pi_{DC}^1)(1-\pi_{CC}^2) & (1-\pi_{DC}^1)\pi_{CC}^2 & \pi_{DC}^1(1-\pi_{CC}^2) & \pi_{DC}^1\pi_{CC}^2 \\
(1-\pi_{DD}^1)(1-\pi_{DC}^2) & (1-\pi_{DD}^1)\pi_{DC}^2 & \pi_{DD}^1(1-\pi_{DC}^2) & \pi_{DD}^1\pi_{DC}^2 \\
(1-\pi_{DD}^1)(1-\pi_{CC}^2) & (1-\pi_{DD}^1)\pi_{CC}^2 & \pi_{DD}^1(1-\pi_{CC}^2) & \pi_{DD}^1\pi_{CC}^2 \\
(1-\pi_{DD}^1)(1-\pi_{DD}^2) & (1-\pi_{DD}^1)\pi_{DD}^2 & \pi_{DD}^1(1-\pi_{DD}^2) & \pi_{DD}^1\pi_{DD}^2 \\
(1-\pi_{DD}^1)(1-\pi_{CD}^2) & (1-\pi_{DD}^1)\pi_{CD}^2 & \pi_{DD}^1(1-\pi_{CD}^2) & \pi_{DD}^1\pi_{CD}^2 \\
(1-\pi_{DC}^1)(1-\pi_{DD}^2) & (1-\pi_{DC}^1)\pi_{DD}^2 & \pi_{DC}^1(1-\pi_{DD}^2) & \pi_{DC}^1\pi_{DD}^2 \\
(1-\pi_{DC}^1)(1-\pi_{CD}^2) & (1-\pi_{DC}^1)\pi_{CD}^2 & \pi_{DC}^1(1-\pi_{CD}^2) & \pi_{DC}^1\pi_{CD}^2
\end{bmatrix}.$$

Hence,

$$V_\pi^1 = (I - \gamma P W)^{-1} R^1.$$

If both agents play Grim Trigger, the starting joint action will be (C, C), which has the following value for Agent 1:

$$V_{\pi_{\text{GT}}}^1(C,C) = \frac{-3\gamma p(\gamma-1)(p-1) - \gamma p(2\gamma(p-1)^2 - p(\gamma p - 1)) + 2(1-\gamma)(\gamma p - 1)}{(\gamma-1)(\gamma p-1)(\gamma(p-1)^2 - 1)}$$

If instead Agent 1 plays π_{B}, the starting joint action will be (D, C), yielding Agent 1 the following value:

$$V_{\pi_{\text{B,GT}}}^1(D,C) = \frac{-\gamma(p-1) - 3\gamma + 3}{(\gamma-1)(\gamma p - 1)}$$

For $\gamma \in [0, 1)$,

$$V_{\pi_{\text{B,GT}}}^1(D,C) > V_{\pi_{\text{GT}}}^1(C,C) \iff p > 2 - \sqrt{\frac{1}{\gamma} + 2}.$$

Similarly, if both agents play Tit-for-Tat, the starting joint action (C, C) has the following value for Agent 1:

$$V^1_{\pi_{\text{TFT}}}(C, C) = \frac{3\gamma p - 2\gamma + 2}{-2\gamma^2 p + \gamma^2 + 2\gamma p - 2\gamma + 1}$$

If instead Agent 1 unconditionally plays Cooperate π_{C}, the starting joint action (C, C) has the following value for Agent 1:

$$V^1_{\pi_{\text{C,TFT}}}(C, C) = \frac{2(\gamma p - 1)}{\gamma - 1}$$

For $\gamma \in [0, 1)$,

$$V^1_{\pi_{\text{C,TFT}}}(C, C) > V^1_{\pi_{\text{TFT}}}(C, C) \iff p > \frac{2\gamma - 1}{4\gamma}.$$

Hence, we have found conditions under which there exists a profitable deviation for Agent 1 from the Grim Trigger and Tit-for-Tat strategies. □

Proof. **Proposition 2.** Using this intervention with $p > 0$, the realised policy of agent i given intended policy π^i is

$$p \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + (1 - p) \begin{bmatrix} \pi^i_{CC} \\ \pi^i_{CD} \\ \pi^i_{DC} \\ \pi^i_{DD} \end{bmatrix}.$$

In this framework, to find the values we must solve

$$V^i_{p+(1-p)\pi} = (I - \gamma W)^{-1} R^i,$$

with W as defined in the proof of Proposition 1. Suppose that agents play π_{GT}. In this case, given that the intervention may overwrite agents' intended actions, the value $V^1_{p+(1-p)\pi_{\text{GT}}}$ for Agent 1 is a weighted average of starting joint actions values:

$$\begin{aligned} V^1_{p+(1-p)\pi_{\text{GT}}} \triangleq & (1-p)^2 V^1_{p+(1-p)\pi_{\text{GT}}}(C, C) + (1-p)p V^1_{p+(1-p)\pi_{\text{GT}}}(C, D) + \\ & p(1-p) V^1_{p+(1-p)\pi_{\text{GT}}}(D, C) + p^2 V^1_{p+(1-p)\pi_{\text{GT}}}(D, D) \end{aligned}$$

Solving the Bellman equation for the realised joint policy $p + (1 - p)\pi_{\text{GT}}$ yields

$$V^1_{p+(1-p)\pi_{\text{GT}}} = \frac{-p^2\gamma + 3p\gamma - p - 2\gamma + 2}{p^2\gamma^2 - p^2\gamma - 2p\gamma^2 + 2p\gamma + \gamma^2 - 2\gamma + 1}.$$

If instead Agent 1 attempts to play π_{B}, the starting joint action will be (D, C) with probability $1 - p$ and (D, D) with probability p. The weighted sum of the values of these two joint actions yields the following value $V^1_{p+(1-p)\pi_{\text{B, GT}}}$ for Agent 1:

$$V^1_{p+(1-p)\pi_{\text{B, GT}}} = \frac{-2p\gamma + 2p + 2\gamma - 3}{\gamma - 1}$$

We find that

$$V^1_{p+(1-p)\pi_{\text{B,GT}}} > V^1_{p+(1-p)\pi_{\text{GT}}} \iff p > 1 - \frac{1}{\sqrt{2\gamma}}.$$

Following a similar procedure for π_{TFT}, evaluating deviations to π_{B} and π_{C}, yields the following two results:

$$V^1_{p+(1-p)\pi_{\text{C,TFT}}} > V^1_{p+(1-p)\pi_{\text{TFT}}} \iff p \in (0, 1 - \frac{1}{2\gamma})$$

$$V^1_{p+(1-p)\pi_{\text{B,TFT}}} > V^1_{p+(1-p)\pi_{\text{TFT}}} \iff p \in (1 - \frac{1}{2\gamma}, 1)$$

Hence, there is a profitable deviation for Agent 1 if $p \neq 1 - \frac{1}{2\gamma}$. Interestingly, when $p = 1 - \frac{1}{2\gamma}$, the value of the policy played by Agent 1 against an opponent attempting to play Tit-for-Tat equals $\frac{2\gamma-3}{\gamma-1}$, independent of the policy chosen. □

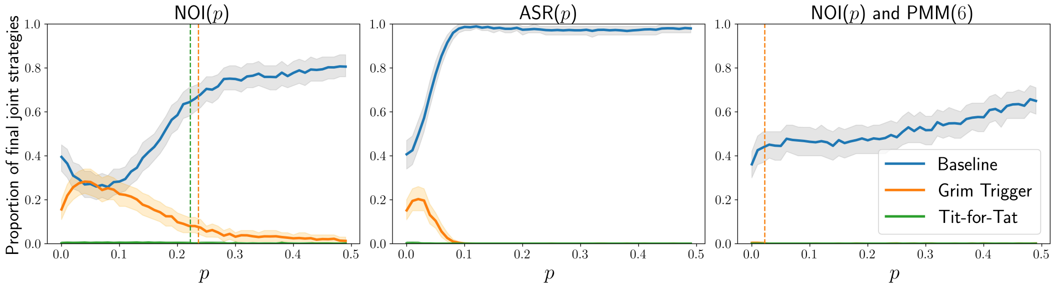

Figure 5: Proportion of 3,000 Q-learning simulations that result in both agents finding the baseline, Grim Trigger, or Tit-for-Tat policies for each intervention. Dashed vertical lines denote the theoretical thresholds from Section 4. 10% and 90% quantiles of a 100-sample bootstrap are shaded.

Proof. **Proposition 3.** We calculate values V^1_π using the formula given in the proof of Proposition 1

$$V^1_\pi = (I - \gamma PW)^{-1} R^1$$

with

$$\begin{bmatrix} R^1(C,C) \\ R^1(C,D) \\ R^1(D,C) \\ R^1(D,D) \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 3+k \\ 1 \end{bmatrix}.$$

The modified payoff matrix can be found in Figure 4.

Solving as above, we find that if both agents play Grim Trigger, the value of the initial state (C,C) for Agent 1 is

$$V^1_{\pi_{\mathrm{GT}}}(C,C) = \frac{-\gamma p(\gamma-1)(k+3)(p-1) - \gamma p\left(2\gamma(p-1)^2 - p(\gamma p - 1)\right) + 2(1-\gamma)(\gamma p - 1)}{(\gamma-1)(\gamma p - 1)(\gamma(p-1)^2 - 1)}.$$

If instead Agent 1 plays π_{B}, the starting joint action will be (D,C), yielding Agent 1 the following value:

$$V^1_{\pi_{\mathrm{B,GT}}}(D,C) = \frac{-\gamma(p-1) - (\gamma-1)(k+3)}{(\gamma-1)(\gamma p-1)}$$

Solving, we find

$$V^1_{\pi_{\mathrm{B,GT}}}(D,C) > V^1_{\pi_{\mathrm{GT}}}(C,C) \iff p > \frac{\gamma k + 4\gamma - \sqrt{\gamma(\gamma k^2 + 4\gamma k + 8\gamma + 4k + 4)}}{2\gamma} \triangleq p_0(\gamma, k).$$

Since

$$\frac{dp_0(\gamma,k)}{d\gamma} = \frac{\sqrt{\gamma(\gamma k^2 + 4\gamma k + 8\gamma + 4k + 4)}(k+1)}{\gamma^2(\gamma k^2 + 4\gamma k + 8\gamma + 4k + 4)} > 0, \quad \text{for } \gamma > 0, k > 0,$$

we find that $p_0(1,k) > p_0(\gamma,k)$ for all $\gamma \in (0,1)$. Hence, if

$$p > p_0(1,k) = \frac{k - \sqrt{k^2 + 8k + 12}}{2} + 2,$$

Grim Trigger is guaranteed to *not* be a Nash equilibrium, regardless of the value of γ. □

D Additional Empirical Results

Figure 5 presents the results from our Q-learning experiments with a specific focus on the collusive strategies considered in Section 4, namely π_{GT} and π_{TFT}. As in the graph above, we find that each intervention applied using parameters adhering to our theoretical results causes a reduction in agents' propensities to find these policies and an increase in the prevalence of the non-collusive baseline.

	C	D
C	$2, 2, -4$	$0, 3+k, -3-k$
D	$3+k, 0, -3-k$	$1, 1, -2$

Figure 4: The modified prisoner's dilemma under PMM(k).