PNAS RESEARCH ARTICLE | EVOLUTION APPLIED MATHEMATICS OPEN ACCESS

# Unilateral incentive alignment in two-agent stochastic games

Alex McAvoy[a,b,1], Udari Madhushani Sehwag[c,d,2], Christian Hilbe[e], Krishnendu Chatterjee[f], Wolfram Barfuss[g,h], Qi Su[i,j,k], Naomi Ehrich Leonard[l], and Joshua B. Plotkin[m,n,2]

Edited by Yoh Iwasa, Kyushu Daigaku, Fukuoka, Japan; received November 22, 2023; accepted April 16, 2024, by Editorial Board Member Elke U. Weber

**Multiagent learning is challenging when agents face mixed-motivation interactions, where conflicts of interest arise as agents independently try to optimize their respective outcomes. Recent advancements in evolutionary game theory have identified a class of "zero-determinant" strategies, which confer an agent with significant unilateral control over outcomes in repeated games. Building on these insights, we present a comprehensive generalization of zero-determinant strategies to stochastic games, encompassing dynamic environments. We propose an algorithm that allows an agent to discover strategies enforcing predetermined linear (or approximately linear) payoff relationships. Of particular interest is the relationship in which both payoffs are equal, which serves as a proxy for fairness in symmetric games. We demonstrate that an agent can discover strategies enforcing such relationships through experience alone, without coordinating with an opponent. In finding and using such a strategy, an agent ("enforcer") can incentivize optimal and equitable outcomes, circumventing potential exploitation. In particular, from the opponent's viewpoint, the enforcer transforms a mixed-motivation problem into a cooperative problem, paving the way for more collaboration and fairness in multiagent systems.**

fairness | mixed-motivation interaction | stochastic game | zero-determinant strategy

## Significance

Conflicts of interest permeate a variety of domains, including biology, the social sciences, and artificial intelligence. In many cases, self-interested agents fail to navigate these conflicts effectively, often resulting in socially inefficient outcomes. Building on recent advances in evolutionary game theory, we demonstrate how an agent can discover strategies to single-handedly align incentives with an opponent, thereby promoting fair and optimal outcomes when the opponent optimizes its own payoff. This approach does not rely on agent coordination or the presence of a centralized authority, and so it offers a decentralized solution to conflicts of interests.

Game theory has proven indispensable for studying the dynamics of collective behavior in groups of interacting agents. The games typically analyzed by researchers are not tailored to precisely depict a particular empirical system. Rather, they aim to capture essential aspects of the underlying system while retaining a level of simplicity that allows rigorous mathematical analysis, controlled experimentation, and numerical simulation. This reductionist approach enables researchers to explore the fundamental principles governing interactions among individuals in various contexts, fostering understanding of a system's dynamics and uncovering potential avenues for interventions.

To explore the tension between individual and group incentives, one useful model involves two strategic types: $C$ ("cooperate") or $D$ ("defect"). If the payoff to an individual when all agents cooperate exceeds the payoff when they all defect, yet any individual cooperator could improve their payoff by switching to defection, then the game is a cooperative social dilemma (1). For groups of two individuals ("agents"), a game with these properties is called a "prisoner's dilemma" and can be described by the payoff matrix

$$\begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} a_{CC} & a_{CD} \\ a_{DC} & a_{DD} \end{pmatrix} \end{array}, \qquad [1]$$

where the parameters $a_{CC}$, $a_{CD}$, $a_{DC}$, and $a_{DD}$ satisfy $a_{DC} > a_{CC} > a_{DD} > a_{CD}$. This ranking ensures that, although both agents would prefer mutual cooperation to mutual defection ($a_{CC} > a_{DD}$), they each have a temptation to defect, regardless of what the opponent does ($a_{DC} > a_{CC}$ and $a_{DD} > a_{CD}$). As a result, any rational agent will choose to defect since defection is, in game-theoretic terminology, the "dominant" strategy.

Outcomes are much more interesting and complicated for the iterated prisoner's dilemma. Agents can choose to defect or cooperate in each round (or stage) conditioned on the actions taken by both agents in all prior rounds. When iterated for a random (but sufficiently large) number of rounds, there are many Nash equilibria and possible long-term outcomes (2), depending upon how agents learn. Importantly, the iterated game allows for the threat of punishment and the promise of future reciprocation (3, 4), what Axelrod incisively calls the "shadow of the future" (5).

Approximately a decade ago, Press and Dyson (6) surprised the game theory community with the discovery of a class of strategies for the iterated prisoner's dilemma (IPD) that allow one agent to exert substantial unilateral control on mean payoffs for

Author contributions: A.M., U.M.S., C.H., K.C., W.B., Q.S., N.E.L., and J.B.P. designed research; performed research; contributed new reagents/analytic tools; analyzed data; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. Y.I. is a guest editor invited by the Editorial Board.



[1]To whom correspondence may be addressed. Email: amcavoy@unc.edu.

[2]U.M.S. and J.B.P. are organizers of this Special Feature.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2319927121/-/DCSupplemental.

Published June 16, 2025.

the pair. These "zero-determinant" strategies allow, for example, an agent to ensure that their payoff is exactly twice that of the opponent's in donation games (a kind of prisoner's dilemma in which cooperators pay $c$ to donate $b$, and defectors do nothing). One special case of a zero-determinant strategy is tit-for-tat (TFT), which begins by cooperating and subsequently copies the action the opponent used in the previous round. Tit-for-tat is famous for having won Axelrod's IPD tournaments, and it produces identical payoffs for both agents, over an infinite number of rounds, regardless of what the opponent does. Another way of thinking about this property is that the TFT agent effectively removes itself from the game, ensuring that no matter what the opponent does, both agents get the same score. Thus, if the opponent is motivated to improve their score (e.g., through an adaptation/optimization process), then both agents will benefit equally. Indeed, this property is one reason why TFT was so successful in IPD tournaments, despite apparent drawbacks such as sensitivity to errors in implementation (7).

On the applied side, recent studies have shown that zero-determinant strategies can effectively address cybersecurity challenges and optimize resource allocation in wireless communication and crowdsensing systems (8–11). For instance, in the case of the vulnerable open architecture of the Internet of Things, the use of a moving-target defense approach based on zero-determinant strategies can enhance counterattack performance (8). Likewise, zero-determinant strategies employed by information requestors can incentivize mobile-device owners to share high-quality data and increase social welfare in mobile crowdsensing (9). Zero-determinant strategies have even been applied to design controllers that mitigate cyber switching attacks in smart grid systems, which can stabilize the power system regardless of attack strategies (10).

Many studies of zero-determinant strategies, including the original discovery (6), are based on a linear-algebraic technique that exploits the $2 \times 2$ structure of the game as well as its infinite time horizon. Subsequent work extends this approach to other settings, e.g., games with discounting and finite time horizons (12, 13), larger action spaces (14, 15), larger groups (16, 17), longer memories (17, 18), or observation errors (19). These extensions provide more a general perspective on what conditions allow an agent to unilaterally enforce a given payoff relationship among agents, as well as the implications of that relationship for groups (20–22) and populations (23–25).

An important aspect of zero-determinant strategies is that they encourage a geometric view of strategies in payoff space. When $X$ plays a fixed strategy, one can envision the resulting space of all possible payoffs (to both agents) within $\mathbb{R}^2$, as the opponent varies their strategy. If $X$ plays a zero-determinant strategy, then the resulting space of possible payoffs falls along a line. In this way, $X$ unilaterally enforces a linear payoff relationship between the two agents. In general, the space of possible payoffs, given the strategy of $X$, can have the same or smaller dimension as the ambient space (two); and this space defines the geometry of incentives that $X$ dictates to their opponent by choosing a particular strategy.

One useful perspective that has emerged involves mapping a condition on short-term rewards to a linear relationship on long-term payoffs. Specifically, suppose that $A_X$ and $A_Y$ are the action spaces available to $X$ and $Y$, respectively; $r_X$ and $r_Y$ are the reward functions for the stage game; $\lambda \in [0, 1)$ is a discount factor (or, equivalently, a continuation probability); $\pi_X^0$ is a probability distribution on $A_X$ representing the initial mixed action of $X$; and, for every $x \in A_X$ and $y \in A_Y$, $\pi_X(x, y)$ a distribution on $A_X$ representing the mixed action of $X$ conditioned on observing $X$ play $x$ and $Y$ play $y$ in the previous round. For shape parameters $(\alpha, \beta, \gamma) \in \mathbb{R}^3$, let $\Phi(x, y) = \alpha r_X(x, y) + \beta r_Y(x, y) + \gamma$. For repeated games, it is known (14) that if there exists a function $\psi : A_X \to \mathbb{R}$ such that the equation

$$\Phi(x, y) = \psi(x) - \lambda \mathbb{E}\left[\psi(x') \middle| x' \sim \pi_X(x, y)\right] - (1 - \lambda)\,\mathbb{E}\left[\psi(x^0) \middle| x^0 \sim \pi_X^0\right], \quad [2]$$

holds for all $x \in A_X$ and $y \in A_Y$, then $X$ can ensure a linear relationship between the long-term mean payoffs to $X$ and $Y$, $\alpha V_X + \beta V_Y + \gamma = 0$, regardless of $Y$'s strategy $\pi_Y$ and regardless of agent $Y$'s memory length. This result was motivated by a desire to understand zero-determinant strategies in games with continuous action spaces (14), where the linear-algebraic approach of Press and Dyson (6) is not applicable.

Inspired by this perspective, our goal here is to establish a result for the existence of zero-determinant [or "autocratic" (14)] strategies for two-agent stochastic games, in which there is an external state that can change over time, thereby altering the payoff structure of interactions. However, unlike in previous settings where the goal is to solve directly for zero-determinant strategies, here we take a learning perspective, where "learn" is relative to an objective and is interpreted more broadly than simple payoff-maximization (as is common in evolutionary game theory). We study how agent $X$ can discover, over time, a style of play that resembles (or coincides with) a zero-determinant strategy. In doing so, this agent learns to incentivize an opponent in a desired manner and thereby mitigate conflicts of interest.

## Model and Goals

The most general setting we consider is a partially observable stochastic game (POSG) (26), which is defined by a tuple $\{n, S, A, r, \mu^0, T\}$, where $n$ is the number of agents, $S$ is the state space of the game, $A = A_1 \times \cdots \times A_n$ is the joint action space available to agents, $r : S \times A \to \mathbb{R}^n$ is the reward function, $\mu^0 \in \Delta(S)$ is the distribution over initial states, and $T : S \times A \to \Delta(S)$ is the state-transition probability map, where $\Delta(S)$ denotes the space of probability distributions on $S$. We denote by $r_i : S \times A \to \mathbb{R}$ the reward function of agent $i$.

In principle, the action spaces depend on the state, $s \in S$, but by augmenting the action spaces to have rewards of zero on infeasible actions in a state, if necessary, we may assume that $A_i$ is independent of state. Although agent $i$ need not directly observe state $s \in S$ and action $a \in A$ of oneself or others in order to receive the reward $r_i(s, a)$, the policy $i$ uses to play the game depends on at least a partial observation of the state. In state $s \in S$, we assume that there is a signal $O(s) \in \Delta(\mathcal{O})$ in some observation space, $\mathcal{O}$. Each agent, $i$, has a private view of this signal, taking values in some space $\mathcal{O}_i$, via a map $O_i : \mathcal{O} \to \mathcal{O}_i$. Note that if $\mathcal{O} = \mathcal{O}_1 \times \cdots \times \mathcal{O}_n$ and the image of $O$ is in $\Delta(\mathcal{O}_1) \times \cdots \times \Delta(\mathcal{O}_n) \subseteq \Delta(\mathcal{O})$, then by projecting onto the $i$th coordinate we get a map $O_i : S \to \Delta(\mathcal{O}_i)$, so independent observations constitute a special case. A strategy for agent $i$ is then a map $\pi_i : \mathcal{O}_i \to \Delta(A_i)$. Once observations are made, the agents randomize independently based on their respective strategies; however, the observations themselves are allowed to be correlated across individuals (for example, visibility of the environment based on weather). In summary, the setting of partially observable stochastic games is a vast generalization of simple repeated games, allowing much greater flexibility for applications.

There are different ways to incorporate the "memory" of a policy into this framework. One way is for the history of past state-action pairs to be included directly into the state space, augmenting $S$. From this perspective, the state of a stochastic

game consists of an environmental component, which affects rewards, and a memory component, which does not affect rewards. Even matrix games in this setting, where the interaction is the same in every time step, may be viewed as finite state machines (27, 28) and would not be considered "stateless" ($|S| = 1$). For example, some authors model strategies as maps from observations of the state to mixed actions (29, 30), as we do in the model above, in which case a strategy for a stateless game is just an unconditional mixed action, which cannot leverage repetition. Other authors explicitly incorporate the history of past states and actions into the strategy (31), and in this case, stateless games do correspond to the repeated matrix games used extensively in behavioral economics.

Here, we take a hybrid approach to memory. For the sake of generality and notation, we state our main theoretical results in terms of strategies conditioning on only the observations of the state, which could include information about both the environment and the history of actions. Later, when discussing specializations of the model and concrete examples, we make explicit which components of a strategy come from the environment and which are related to history. There, we follow the tradition in evolutionary game theory of treating strategies as stochastic with fixed memory (32). We focus on memory-one strategies, which use the environment and actions in the previous time step, in addition to the current environment, when devising actions.

In this study, we focus on $n = 2$ agents, denoted by $X$ and $Y$, respectively. The game starts in some state, $s \in S$. Each agent $i$ observes this state via the map $O_i$, chooses an action, and receives a reward. We adopt a probabilistic interpretation of discounting, which is common in evolutionary game theory. With probability $\lambda \in [0, 1]$, a new state is chosen and the game continues to another round; with probability $1 - \lambda$, the game ends. The mean game length is $1/(1-\lambda)$. Suppose that $\Phi : S \times A \to \mathbb{R}$ is a function of the state and action profile. For $s \in S$ and strategies $\pi_X$ and $\pi_Y$, the expected value of $\Phi$, given that the initial state $s$ is drawn from $\mu^0$, is

$$V_{\Phi}^{\pi_X,\pi_Y}\left(\mu^0\right) = \mathbb{E}\left[(1-\lambda)\sum_{t=0}^{\infty}\lambda^t\Phi\left(s^t,a^t\right)\middle|\begin{array}{c} s^0\sim\mu^0 \\ s^{t+1}\sim T\left(s^t,a_X^t,a_Y^t\right) \\ a_X^{t+1}\sim\pi_X\circ O_X\left(s^t\right) \\ a_Y^{t+1}\sim\pi_Y\circ O_Y\left(s^t\right)\end{array}\right]. \quad [3]$$

When $\Phi = r_X$ or $\Phi = r_Y$, we write $V_X^{\pi_X,\pi_Y}$ and $V_Y^{\pi_X,\pi_Y}$, respectively, for $V_{\Phi}^{\pi_X,\pi_Y}$. To clarify the two timescales, we refer to stage scores as "rewards" and long-run discounted scores as "payoffs."

Note that, when viewing $\lambda$ as a continuation probability, the probability that the game ends immediately after round $t$ is $(1-\lambda)\lambda^t$, so the summation $(1-\lambda)\sum_{t=0}^{\infty}\lambda^t\Phi_t$ can be interpreted as the expected value of $\Phi$ in the final round of the stochastic game. Alternatively, since $\sum_{t=0}^{\infty}\lambda^t\Phi_t = (1-\lambda)\sum_{t=0}^{\infty}\lambda^t\sum_{t'=0}^{t}\Phi_{t'}$, the summation $(1-\lambda)\sum_{t=0}^{\infty}\lambda^t\Phi_t$ can be seen as the normalized expected sum of all values of $\Phi$ over the duration of the stochastic game. The latter interpretation is a hybrid of classical and evolutionary approaches, since (often) the former treats $\lambda$ as a discounting factor and the latter treats $\lambda$ as a continuation probability. Practically speaking, this distinction is important because, with $\lambda$ interpreted probabilistically, there may be no guarantee of a minimum number of rounds in an episode.

Ultimately, we are concerned with comparing the long-term payoffs of $X$ and $Y$, so we need a baseline notion of what it means for agents to be interchangeable. Of particular interest is when equality of payoffs can be interpreted as a proxy for fairness. To this end, we use a notion of exchangeability for the game, which roughly means that although agents can hold different positions in different states of the game, there are no intrinsic differences among the agents. When there is only a single game environment, an exchangeable stochastic game is simply a symmetric normal-form game (33). We relegate the formal technical definition of exchangeability for partially observable stochastic games to *SI Appendix*, and here we only briefly describe what this notion requires for two-state stochastic games. We begin with a simple example of an extended prisoner's dilemma:

***Example (a two-state prisoner's dilemma):*** Inspired by the "coin game" of Lerer and Peysakhovich (34), we can define a two-state version of a prisoner's dilemma as follows: The state of the game is defined by the color of a coin, blue (state $s_1$) or red (state $s_2$), which can be picked up by either agent. In addition, one agent is blue and the other is red. If an agent picks up a coin of any color, then this agent receives a benefit of $b$. If the coin matches the agent's color, then the opponent receives nothing (neither benefit nor harm). However, if the coin does not match the agent's color, then the opponent receives $-c$, representing harm. We take the action $C$ to mean "pick up the coin only if it is the same color as yourself" and $D$ to mean "pick up the coin regardless of color." If both agents attempt to pick up the coin at the same time, then a fair coin is flipped for which one gets it. The reward matrices in the blue and red states, $s_1$, and $s_2$, are thus

$$r(s_1,-,-) = \begin{array}{c} \\ C \\ D \end{array}\begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} b,0 & \frac{1}{2}(b-c),\frac{1}{2}b \\ b,0 & \frac{1}{2}(b-c),\frac{1}{2}b \end{pmatrix} \end{array}; \quad [4a]$$

$$r(s_2,-,-) = \begin{array}{c} \\ C \\ D \end{array}\begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} 0,b & 0,b \\ \frac{1}{2}b,\frac{1}{2}(b-c) & \frac{1}{2}b,\frac{1}{2}(b-c) \end{pmatrix} \end{array}, \quad [4b]$$

where we use the convention that the row agent is blue and the column agent is red. The initial state is chosen uniformly at random, and all subsequent transitions are periodic, passing from $s_1$ to $s_2$ and from $s_2$ to $s_1$.

We note that by averaging over trajectories, the empirical payoffs are given by a matrix game with $a_{CC} = b/2$, $a_{CD} = (b-c)/4$, $a_{DC} = 3b/4$, and $a_{DD} = b/2 - c/4$, which satisfies the ranking of a traditional prisoner's dilemma, $a_{DC} > a_{CC} > a_{DD} > a_{CD}$. Although both matrix games in this two-state prisoner's dilemma are highly asymmetric, the stochastic game is exchangeable in the sense that there is no bias in the initial state, the roles of the red and blue agents are swapped in the two states, and the transitions are symmetric with respect to the states. There are no intrinsic differences between the agents. This idea is formalized in *SI Appendix* for general $n$-agent POSGs.

The focus of our study is on exchangeable, two-agent stochastic games. Although nonexchangeable games are common in many natural populations and certainly warrant further study, the assumption of exchangeability is a natural starting point. It is also not overly restrictive from the point of view of multiagent reinforcement learning. While not fully general (e.g., different autonomous vehicles might be made by different manufacturers with distinct specifications), it is reasonable as a first approximation that learning takes place among comparable agents (e.g.,

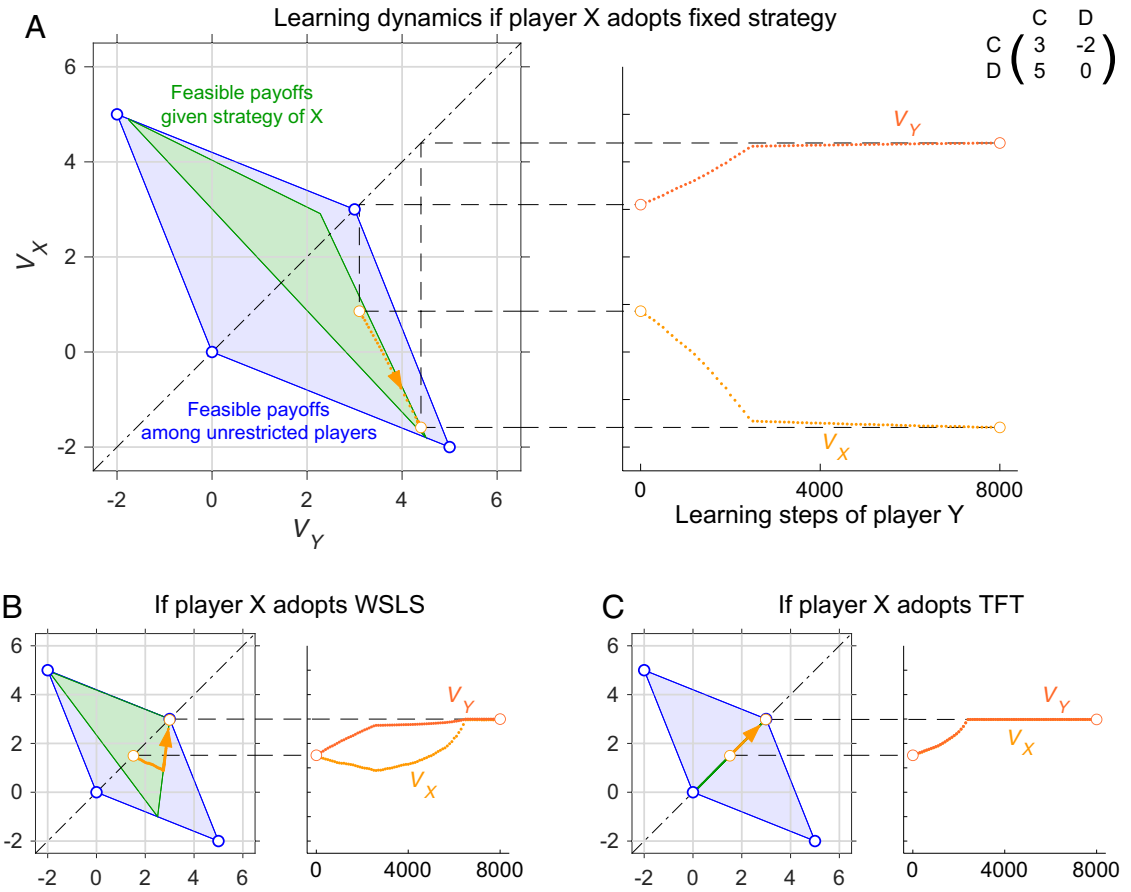


**Fig. 1.** Adaptation/learning against fixed strategies. For a classical prisoner's dilemma game, we illustrate how agent *X* can guide *Y*'s learning process. In each panel, the blue area depicts the set of all feasible payoffs. These are the payoffs, $(V_X, V_Y) \in \mathbb{R}^2$, that are possible in principle, when both *X* and *Y* can choose their strategies arbitrarily. The green area represents the payoffs that are feasible given that agent *X* uses some fixed strategy. (*A*) For most generic strategies of agent *X*, the best response for Y is to always defect (35). In particular, if *Y* learns to adopt more profitable strategies over time according to a selfish learning process (36), *Y* ends up with a high payoff whereas *X* obtains a low payoff. (*B*) If agent *X* instead adopts the strategy "win-stay, lose-shift" (4), *Y* eventually learns to adopt a fully cooperative strategy. However, the learning process can be slow and agent *X*'s payoff may temporarily decrease along *Y*'s learning trajectory. (*C*) If *X* enforces equal payoffs by using the strategy tit-for-tat, *Y* quickly learns to cooperate. (The green region collapses onto the line $V_X = V_Y$ in tit-for-tat.) Here, both agents' payoffs increase monotonically. In all panels, the orange dots represent the trajectories of both agents' payoffs when *Y* optimizes its strategy to attain better payoffs against *X* (whose strategy is fixed).

identical vehicles from one company). In this initial study of zero-determinant strategies in stochastic games, we are focused on idealized scenarios, and we do not seek to capture the reality of agents nearly as complex as self-driving cars.

Given this model, we seek to explore the following questions:

**Question 1.** *How can an agent incentivize a selfish learner to lead both agents toward a fixed target? That is, for a target* $(V_X^*, V_Y^*) \in \mathbb{R}^2$*, which is chosen unilaterally by X, how can X find a strategy with the property that Y's best response gives* $V_X^*$ *to X and* $V_Y^*$ *to Y?*

**Question 2.** *For a target* $(V_X^*, V_Y^*) \in \mathbb{R}^2$*, which is chosen unilaterally by X, how can X find a strategy such that whenever Y changes its strategy to improve its payoff, the payoffs of both agents move closer to the target (e.g., based on the Euclidean distance in* $\mathbb{R}^2$*)?*

In a sense, Question 1 is sufficient. As long as one can incentivize an opponent in a certain manner, the learner might not care what the trajectory looks like on the way to the final outcome (Fig. 1*B*). But transient behavior can also be important, and zero-determinant strategies impose even stricter conditions than those needed to answer Question 1, so we also consider when the feasible region can be further collapsed onto a space of codimension one (Fig. 1*C*). Such a strict condition is useful, for instance, when the number of learning steps is uncertain and an agent cares about the payoffs being approximately equal whenever the interaction terminates.

## Results

**Autocratic Strategies for Stochastic Games.** Our main question is when *X* can unilaterally enforce a linear relationship between payoffs,

$$\alpha V_X^{\pi_X,\pi_Y}(\mu^0) + \beta V_Y^{\pi_X,\pi_Y}(\mu^0) + \gamma = 0, \qquad [5]$$

regardless of the strategy $\pi_Y$ of *Y*. Specifically, if $\alpha = -\beta = 1$ and $\gamma = 0$, then this relationship is simply $V_X^{\pi_X,\pi_Y}(\mu^0) = V_Y^{\pi_X,\pi_Y}(\mu^0)$. We take equality of the two value functions as a proxy for fairness in the stochastic game, provided the game is exchangeable.

For general parameters $\alpha, \beta, \gamma \in \mathbb{R}$, we have the following theoretical result:

***Theorem* (autocratic strategies for stochastic games).** *Let*

$$\Phi(s,x,y) := \alpha r_X(s,x,y) + \beta r_Y(s,x,y) + \gamma, \qquad [6]$$

**Box 1.**

**Intuition behind unilateral payoff control.**

By considering the simple case of a repeated game, we can gain intuition for the condition (Eq. **7**) guaranteeing $X$ the ability to enforce a linear payoff relationship (Eq. **5**). In this case, there is no external environmental state, and rewards depend on only the actions taken. When $\Phi$ is additive, meaning $\Phi(x,y) = f(x) + g(y)$, letting $\psi(x) = f(x)$ gives

$$g(y) = -\lambda \mathbb{E}\left[f\left(x'\right)\right] - (1-\lambda)\,\mathbb{E}\left[f\left(x^0\right)\right].$$

(We suppress the conditioning in these expectations to simplify notation.) This equation says that $Y$'s contribution to $\Phi$ in one round, $g(y)$, can be offset (on average) by $X$'s contribution to $\Phi$ in the next round, $\mathbb{E}\left[f\left(x'\right)\right]$, with an initial-round correction. In the additive case, $\psi(x)$ is exactly $X$'s contribution to $\Phi(x,y)$. In the nonadditive case, this interpretation suggests thinking of $\psi(x)$ as an approximation to $X$'s contribution to $\Phi(x,y)$.

Stochastic games are more nuanced than repeated games because corrections cannot always take place in subsequent rounds, owing to the dynamic nature of the game. Therefore, although we can think of $\psi(x)$ as a one-sided approximation to $\Phi(x,y)$ in repeated games, we caution that the same interpretation does not necessarily extend to $\psi(s,x)$ and $\Phi(s,x,y)$ in stochastic games. One reason is that $\psi$ can carry (accumulate) corrections through several rounds, which can allow memory-one strategies to implement longer-range reciprocity (Fig. 4 and *SI Appendix*).

*and suppose that there exists a function $\psi : \mathcal{O}_X \times A_X \to \mathbb{R}$ such that*

$$\begin{aligned}
&\mathbb{E}\left[\Phi(s,x,y)\,\middle|\,\substack{o_X\sim O_X(s)\\ x\sim\pi_X(o_X)}\right]\\
&\quad= \mathbb{E}\left[\psi(o_X,x)\,\middle|\,\substack{o_X\sim O_X(s)\\ x\sim\pi_X(o_X)}\right]\\
&\qquad-\lambda\mathbb{E}\left[\psi\left(o_X',x'\right)\,\middle|\,\substack{o_X\sim O_X(s)\\ x\sim\pi_X(o_X)\\ s'\sim T(s,x,y)\\ o_X'\sim O_X(s')\\ x'\sim\pi_X\left(o_X'\right)}\right]\\
&\qquad-(1-\lambda)\,\mathbb{E}\left[\psi\left(o_X^0,x^0\right)\,\middle|\,\substack{s^0\sim\mu^0\\ o_X^0\sim O_X\left(s^0\right)\\ x^0\sim\pi_X\left(o_X^0\right)}\right],
\end{aligned}\qquad [7]$$

*for all $s \in S$ and $y \in A_Y$. Then, $\pi_X$ ensures that the linear payoff relationship of Eq. **5** holds for all strategies, $\pi_Y$, of $Y$.*

This result provides a way to reduce the problem of enforcing a linear relationship on long-run average payoffs (Eq. **5**) to a condition over pairs of interactions in adjacent time steps (Eq. **7**). This result suggests that we need not consider an objective defined over long-run payoffs if the goal is to enforce a linear (or approximately linear) payoff relationship. Instead, a learner need only consider a more myopic objective function, but the cost of doing so is finding an appropriate function, $\psi : \mathcal{O}_X \times A_X \to \mathbb{R}$.

In addition to generalizing the main condition of McAvoy and Hauert (14), which holds for repeated games, the condition above also extends the corresponding condition for alternating games (37), which is a special case of a stochastic game in which the agents move in a strictly- or randomly alternating fashion. In both synchronous and alternating games, autocratic strategies have been studied in the context of memory-one strategies, meaning an agent conditions its play on only the most recent action(s). To put this kind of strategy into the present context, suppose that the state is perfectly observable (an assumption we will retain for most of our presentation). We define a memory-one strategy for $X$ to be a function $\pi_X$ that takes in the previous state, $s \in S$, the previous actions $(x,y) \in A_X \times A_Y$, and the current state, $s' \in S$, and returns a distribution $\pi_X(s,x,y,s') \in \Delta(A_X)$. Such a strategy must also specify the initial play, which is simply a map $\pi_X^0 : S \to \Delta(A_X)$. A sufficient condition for Eq. **5** to hold for all strategies of $Y$ is then

$$\begin{aligned}
\Phi(s,x,y) = \psi(s,x) &- \lambda\mathbb{E}\left[\psi\left(s',x'\right)\,\middle|\,\substack{s'\sim T(s,x,y)\\ x'\sim\pi_X(s,x,y,s')}\right]\\
&-(1-\lambda)\,\mathbb{E}\left[\psi\left(s^0,x^0\right)\,\middle|\,\substack{s^0\sim\mu^0\\ x^0\sim\pi_X\left(s^0\right)}\right],
\end{aligned}\qquad [8]$$

for all $s \in S$ and $(x,y) \in A_X \times A_Y$. By explicitly conditioning on the recent history of play, this expression allows for the interpretation of $s \in S$ as the environmental component of the game, i.e., the part that affects rewards. Box 1 shows the intuition behind this condition when $|S| = 1$.

**Learning Autocratic Strategies.** We now have a condition that extends autocratic strategies from repeated games to much more general stochastic games. One key difference between this study and previous works is how we will use this condition. Rather than attempting to use this (implicit) condition to solve for autocratic strategies directly, we view its components as functions that can be learned. To this end, suppose that an initial state $s^0 \in S$ is sampled from $\mu^0$, and let $s' \sim T(s,x,y)$ following actions $x \in A_X$ and $y \in A_Y$ in state $s \in S$. Since we wish for Eq. **8** to hold, we associate to the memory-one strategy $\pi_X$ and the function $\psi : S \times A_X \to \mathbb{R}$ an effective cost of

$$\left|\begin{aligned}
&\Phi(s,x,y) - \psi(s,x) + \lambda\mathbb{E}\left[\psi\left(s',x'\right)\,\middle|\,x' \sim \pi_X\left(s,x,y,s'\right)\right]\\
&\qquad+(1-\lambda)\,\mathbb{E}\left[\psi\left(s^0,x^0\right)\,\middle|\,x^0 \sim \pi_X\left(s^0\right)\right]
\end{aligned}\right|.\qquad [9]$$

In practice, we think of both $\pi_X$ and $\psi$ as being modeled by neural networks with parameters $\theta_X$ and $w_X$, in which case gradient descent can be used to minimize the cost (Algorithm 1). Of course, one can calculate gradients exactly in simple games like the prisoner's dilemma, but the use of neural networks allows for flexible function approximation in Algorithm 1, applicable to a variety of game environments. We call a learner who optimizes based on this objective an enforcer with shape parameters $\alpha$, $\beta$, and $\gamma$ (which influence the observed values of $\Phi$).

***Algorithm 1:*** An $(\alpha, \beta, \gamma)$-enforcer, whose goal is to enforce the linear equation $\alpha V_X + \beta V_Y + \gamma = 0$. Since the enforcer's objective is defined over pairs of rounds within an episode, we opt for an approach in which optimization steps are taken at each time step rather than only at the conclusion of an episode. We also assume that the learning rates, $\eta_{\theta_X}$ and $\eta_{w_X}$, are equal in our implementation. Other variations are possible. Note that an agent's action is conditioned on the state, which could include information about the history of past play in addition to information about the game (environment).

1: $\theta_X \leftarrow$ Random
2: $w_X \leftarrow$ Random
3: **for** episode $= 1$ to $E$ **do**
4: sample game length, which is $t_{\max} \in \{0, 1, \dots\}$ with probability $\lambda^{t_{\max}}(1-\lambda)$
5: sample state-action-action pairs, $(s_0, x_0, y_0), \dots, (s_{t_{\max}}, x_{t_{\max}}, y_{t_{\max}})$
6: **for** $t = 0$ to $t_{\max} - 1$ **do**
7: let $J_t(\theta_X, w_X) := \left| \begin{array}{c} \alpha r_X(s_t, x_t, y_t) + \beta r_Y(s_t, x_t, y_t) + \gamma - \psi_{w_X}(s_t, x_t) \\ +\lambda \mathbb{E}_{x' \sim \pi_{\theta_X}(s_{t+1})}\left[\psi_{w_X}(s_{t+1}, x')\right] + (1-\lambda)\,\mathbb{E}_{x' \sim \pi_{\theta_X}(s_0)}\left[\psi_{w_X}(s_0, x')\right] \end{array} \right|$
8: $\theta_X \leftarrow \theta_X - \eta_{\theta_X} \nabla_{\theta_X} J_t(\theta_X, w_X)$
9: $w_X \leftarrow w_X - \eta_{w_X} \nabla_{w_X} J_t(\theta_X, w_X)$

Note that both $\pi_X$ and $\psi$ are subject to optimization here, with $\psi$ playing an indirect role on the strategy itself (but a direct role in how this strategy is learned). At a high level, this framing is reminiscent of actor-critic methods (38), with $\pi_X$ playing the role of the actor and $\psi$ the role of the critic. However, we note that use of the "critic" in Algorithm 1 is significantly different from traditional approaches, which view the critic as an approximation to a value function. Box 1 gives some intuition for how to think about $\psi$ as a one-sided approximation to $\Phi$, at least in the case of repeated games. The interpretation of $\psi$ can be substantially more complicated in multistate stochastic games.

In Algorithm 1, the parameters $\theta_X$ and $w_X$ are updated at each round within an episode. While this approach is not strictly necessary, it is motivated by the fact that we desire Eq. **9** to hold for all $s \in S$, $x \in A_X$, and $y \in A_Y$. Thus, with sufficiently expressive representations of $\pi_X$ and $\psi$, the idea is that we ought to be able to leverage each round in the search for an autocratic strategy (if one exists). One could also update parameters only at the end of each episode, which could also include shuffling the state-action-reward histories in order to reduce correlations in this sequential data (39). For our examples, below, we find that autocratic and autocratic-like strategies can be found using the round-by-round optimization in Algorithm 1.

Using this algorithm for the enforcer, we consider interactions that unfold in two different ways. In the first treatment, the opponent, $Y$, is given a random memory-one strategy, where each coordinate is chosen independently from a Dirichlet distribution whose $|A_Y|$ parameters are all $1/2$. This choice of randomization generalizes the arcsine distribution, which itself explores the corners of $[0, 1]$ better than the uniform distribution (4). A game length $t_{\max}$ is then sampled from a geometric distribution based on the parameter $\lambda$, with an average game length of $1/(1-\lambda)$ rounds. Starting from an initial state chosen from $\mu^0$, the game then proceeds for $t_{\max}$ rounds, and between each pair of successive rounds, the enforcer takes a step to minimize its objective (Eq. **9**). At the end of the iterated game, the opponent is discarded, a new one is sampled, and the enforcer brings forth the strategy it learned against the previous opponent. Notably, in this setting, the opponent is not learning and exists solely to generate experience for the enforcer. Later, we do consider enforcers against selfishly optimizing opponents, but the motivation for pairing enforcers with random agents is embedded in the nature of zero-determinant strategies themselves. When fixed, these strategies enforce restricted outcomes against any opponent, so it is natural that one might be able to learn these strategies even against agents who devise behaviors randomly (or are shuffled around frequently).

**Geometry of the Feasible Region.** Before illustrating the behavior of enforcers in several examples, we first emphasize one aspect of repeated games and "simple" stochastic games that clarifies the nature of zero-determinant strategies and the learning procedure we propose: the feasible region induced by a strategy. The structure of a game puts a constraint on all possible payoffs, $(V_X, V_Y) \in \mathbb{R}^2$, as both agents vary their strategies, $\pi_X$ and $\pi_Y$ (not necessarily memory-one or even limited memory). For a repeated prisoner's dilemma game, Fig. 1 depicts this region in blue. But when one agent (e.g., $X$) fixes its strategy, this imposes a further constraint on attainable payoffs, as $Y$ varies its strategy. This region is shown in green in Fig. 1. In *SI Appendix*, we show that when $X$ plays a memory-one strategy, the feasible region generated by this strategy can be calculated by allowing $Y$ to vary over all memory-one strategies (as opposed to all strategies of arbitrary complexity). In other words, in order to understand the feasible region generated by a memory-one strategy, it suffices to explore opponents who also use memory-one strategies. This result extends a finding of Press and Dyson (6) to multistate games with discounting.

Although this result holds for all stochastic games, it is often not practical to evaluate the feasible payoff region, given $\pi_X$, in most games, due to the size of the space of opponent memory-one strategies. Even the number of "boundary" memory-one strategies, with each conditional action being deterministic, grows exponentially in the size of the game. Nonetheless, this geometric perspective of the feasible region provides a clear picture of what enforcers are doing in simple games, so we primarily consider examples in which this region can be visualized, including the prisoner's dilemma in Eq. **1**, a nonlinear variant of the donation game, and several stochastic games transitioning between two states.

**Examples.**

***Prisoner's dilemma, classical and multistate.*** For a classical variant of the prisoner's dilemma, Fig. 2 *A* and *B* show the results of an enforcer against 100 fixed, randomly chosen (Dirichlet) opponents, in succession. Generically, this kind of random strategy for the IPD has the property that it is readily exploited (Fig. 2*A*).

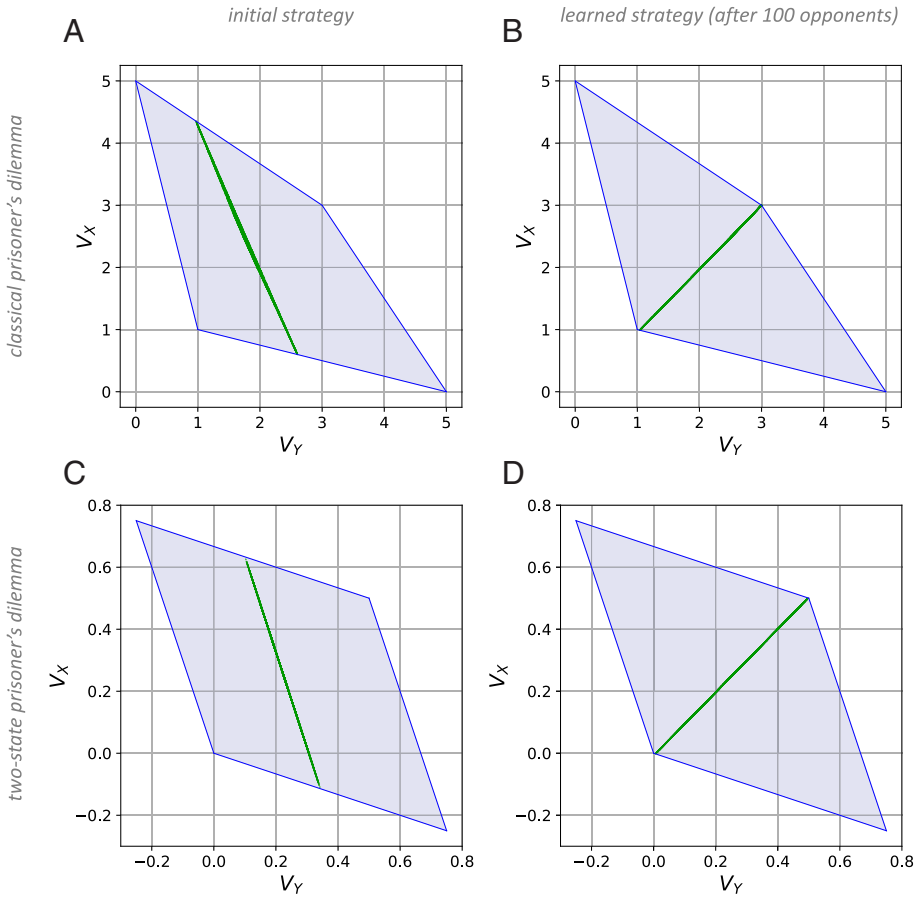


**Fig. 2.** Enforcer in a classical prisoner's dilemma (*A* and *B*) and a two-state analogue (*C* and *D*). The feasible region initially generated by agent *X* is depicted in *A*. After performing learning steps between every adjacent pair of time steps against 100 different random (Dirichlet) opponents, the resulting feasible region is shown in *B*. Note that this feasible region represents all possible payoffs that can be achieved when *X* uses the strategy it learned, regardless of the memory of the opponent. At this point, the incentives of the two agents are aligned, as any attempt by *Y* to increase its payoff leads to an increase in the payoff of *X*. Furthermore, because this relationship is approximately $V_X = V_Y$, once *Y* optimizes its strategy (by whatever means necessary), the outcome will be equal payoffs for both *X* and *Y*, with each getting the socially optimal payoff for mutual cooperation. The green dots give payoffs against $10^4$ random (Dirichlet) memory-one opponents. Parameters: $(a_{CC}, a_{CD}, a_{DC}, a_{DD}) = (3, 0, 5, 1)$ in (*A* and *B*) (Eq. **1**) and $(b, c) = (1, 2)$ in (*C* and *D*) (Eq. **4**). In all panels, $\lambda = 0.99$ and $(\alpha, \beta, \gamma) = (1, -1, 0)$.

Attempting to enforce equality among payoffs using $(\alpha, \beta, \gamma) = (1, -1, 0)$ results in a feasible region that collapses onto a line connecting the payoffs for mutual defection and mutual cooperation. This strategy resembles the geometry generated by TFT, and indeed the learning process has effectively recovered the mechanics of TFT using a purely geometric objective. Fig. 2 *C* and *D* show similar outcomes in a two-state analogue of the prisoner's dilemma, in which both stage games (Eq. **4**) are asymmetric.

***Nonlinear donation game.*** While the IPD is an interesting and well-studied game, it is too simple to capture the tension between multiple investment levels with differing levels of efficiency. To study this we consider a three-action donation game, which allows for defection ($D$), mild cooperation ($C_1$), and full cooperation ($C_2$). A defector provides no benefits and pays no costs. Mild cooperation entails a cost of $c_1$ to provide a benefit of $b_1$ to another agent. Full cooperation is similar, except the cost is $c_2 > c_1$ and the benefit is $b_2 > b_1$. The payoff matrix for this interaction is

$$\begin{array}{c} \\ C_1 \\ C_2 \\ D \end{array} \begin{array}{c} \begin{array}{ccc} C_1 & C_2 & D \end{array} \\ \begin{pmatrix} b_1 - c_1 & b_2 - c_1 & -c_1 \\ b_1 - c_2 & b_2 - c_2 & -c_2 \\ b_1 & b_2 & 0 \end{pmatrix} \end{array}. \qquad [10]$$

Although full cooperation is more altruistic in the sense that it provides a greater benefit than do defection and mild cooperation, we assume that it is also less efficient than mild cooperation in that $b_2 - c_2 < b_1 - c_1$. This three-action game is "nonlinear" in the sense that the benefit is not a linear function of the cost. If it were, it would have to be true that $b_2/c_2 = b_1/c_1$, which would contradict the inequality $b_2 - c_2 < b_1 - c_1$. Thus, if both agents agree to coordinate on an action, they prefer mild cooperation to defection and full cooperation. Fig. 3 shows the results of an enforcer aiming for equality in this game. The results are remarkably similar to those in Fig. 2: notably, the enforcer automatically incentivizes mutual mild cooperation (the socially efficient outcome) over full cooperation.

In both of the cases considered so far, the desired outcome (a linear payoff relationship) was found via the learning process, which is at the heart of Question 2. However, to answer Question 1, a strict payoff line is not necessary. Instead, one can relax this condition and ask for only a statistical relationship between the two agents' payoffs. To explore the behavior of an enforcer when strict linear relationships cannot be enforced on the nose, we consider once again the nonlinear variant of the donation game. This time, however, we use the parameters $(\alpha, \beta, \gamma) = (1, -2, 2)$. For these parameters, no strategy exists that enforces $\alpha V_X + \beta V_Y + \gamma = 0$ for all strategies $\pi_Y$ of the

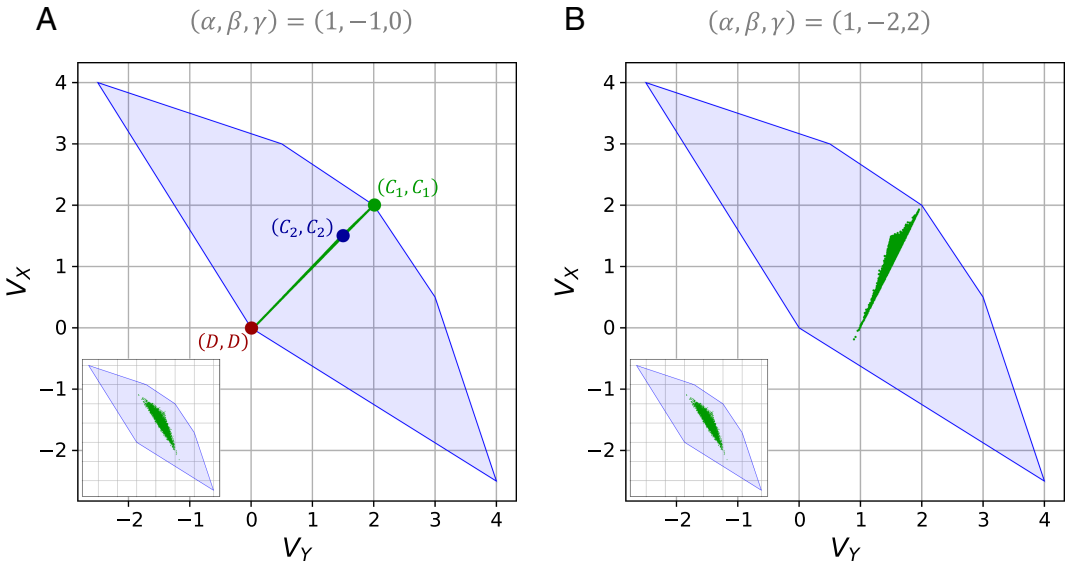


**Fig. 3.** Two kinds of enforcers in a nonlinear donation game. In this game, the optimal outcome is not mutual full investment ($C_2$; blue circle), even though this outcome is superior to mutual defection (red circle). Instead, both agents fare better by each investing a smaller amount, $C_1$ (green circle), an outcome that the enforcer can incentivize after learning. (*A*) Aiming for a fair strategy, meaning $V_X = V_Y$, which incentivizes a selfish opponent to lead both agents toward $(C_1, C_1)$. (*B*) Aiming for a generous (or "compliant") (23, 40) strategy, which enforces the equation $2 - V_X = 2\,(2 - V_Y)$. For this game, one can show that it is impossible to attain this goal in *B*, as no such strategy exists for *X*. Nevertheless, optimizing the associated objective function leads to a strategy with a two-dimensional feasible region that still leads a selfish opponent toward $(C_1, C_1)$. The small green dots give payoffs against $10^4$ random (Dirichlet) memory-one opponents. The insets depict the initial strategies of the enforcer (which are the same in both panels). Parameters: $(b_1, c_1, b_2, c_2) = (3, 1, 4, 2.5)$ (Eq. **10**) and $\lambda = 0.99$.

opponent (*SI Appendix*). Nonetheless, if we use the same objective function for the enforcer, we obtain the results shown in Fig. 3*B*. This figure depicts a two-dimensional feasible region following the learning process, which still incentivizes the socially optimal behavior of mutual mild cooperation. When facing a sufficiently clever self-interested opponent, there might be no meaningful distinction between *A* and *B* in Fig. 3 since they both incentivize the same outcome (just via different paths). This example also shows that it need not matter whether an enforcer can reach its goal of a global minimum of zero for its cost function.

There is, however, one important distinction between the two cases depicted in Fig. 3. If the opponent is selfish and the learning horizon is uncertain, then the enforcer might wish to ensure that the payoffs are approximately equal (a proxy for a "fair" outcome) when learning terminates. In this case *A* would be preferred over *B* in Fig. 3 because it exhibits better control of the learning trajectories. In this sense, the restrictive nature of autocratic strategies can be useful for the transient portions of a learning process.

***Prisoner's dilemma with sparse rewards.*** Autocratic strategies can control longer-range punishment and reward, even when the game changes state over time, despite the apparent limitations of using memory-one strategies. Fig. 4 illustrates a striking example of this, when an enforcer faces a selfish agent [using policy gradient optimization (41)] in an environment with sparse rewards. Here, there is a single "rare" state with a standard matrix game with nonzero rewards, which is then followed by many "null" states in which rewards are zero for both agents. Eventually, the state cycles back to the nontrivial game (Fig. 4*A*). It might seem impossible for *X* to learn a useful memory-one strategy in such a setting (where there are many intervening rounds with no payoff effects, between rounds that matter), but our general theorem still holds, and we can run the algorithm to see the outcomes of the two learners.

The enforcer leads both agents toward higher payoffs on average (Fig. 4*B*). In fact, even when individual learning trajectories lead to lower payoffs, the enforcer still learns strategies that align incentives; in this case, it is the selfish agent that does not always efficiently navigate the resulting landscape (Fig. 4 *C*–*F*). This performance could be improved by facing a more sophisticated selfish agent, but this is a secondary concern: For the sake of aligning incentives and effectively transforming the nature of the interaction, the goal is already attained. The distinction between *E* and *F* in Fig. 4 highlights the utility of visualizing the feasible region enforced by a strategy, as opposed to just a time series of payoffs. (Although feasible regions can be visualized only in relatively simple settings, in general.)

Note that, in Fig. 4, we still allow both agents to choose from two actions in the null states, even though all action pairs lead to a reward of zero. Masking the opponent's behavior in the null states would limit the amount of information *X* could carry forward into the next nontrivial game. However, even if *X* cannot observe *Y*'s actions in the sparse states, *X* can still augment its action space with "informational" actions that serve as memory. For example, if there are four possible outcomes in the nonnull state, then *X* can play one of four informational actions in the following null state as a way to encode the previous outcome, even though these actions all lead to a payoff of zero in the null state. An important takeaway from this example is that memory-one strategies can propagate payoff corrections across many time steps, resulting in behaviors that effectively have longer-range memories. This observation is also why the interpretation of the "critic" function, $\psi$, in repeated games (Box 1) does not necessarily extend to multistate interactions, where payoff corrections from one point in time cannot be guaranteed in the next.

## Discussion

Learning problems in stochastic games can be broadly characterized as fully competitive, fully cooperative, or mixed-motivation (42). Phrased in these terms, we have studied the following question: can an agent unilaterally transform a mixed-motivation

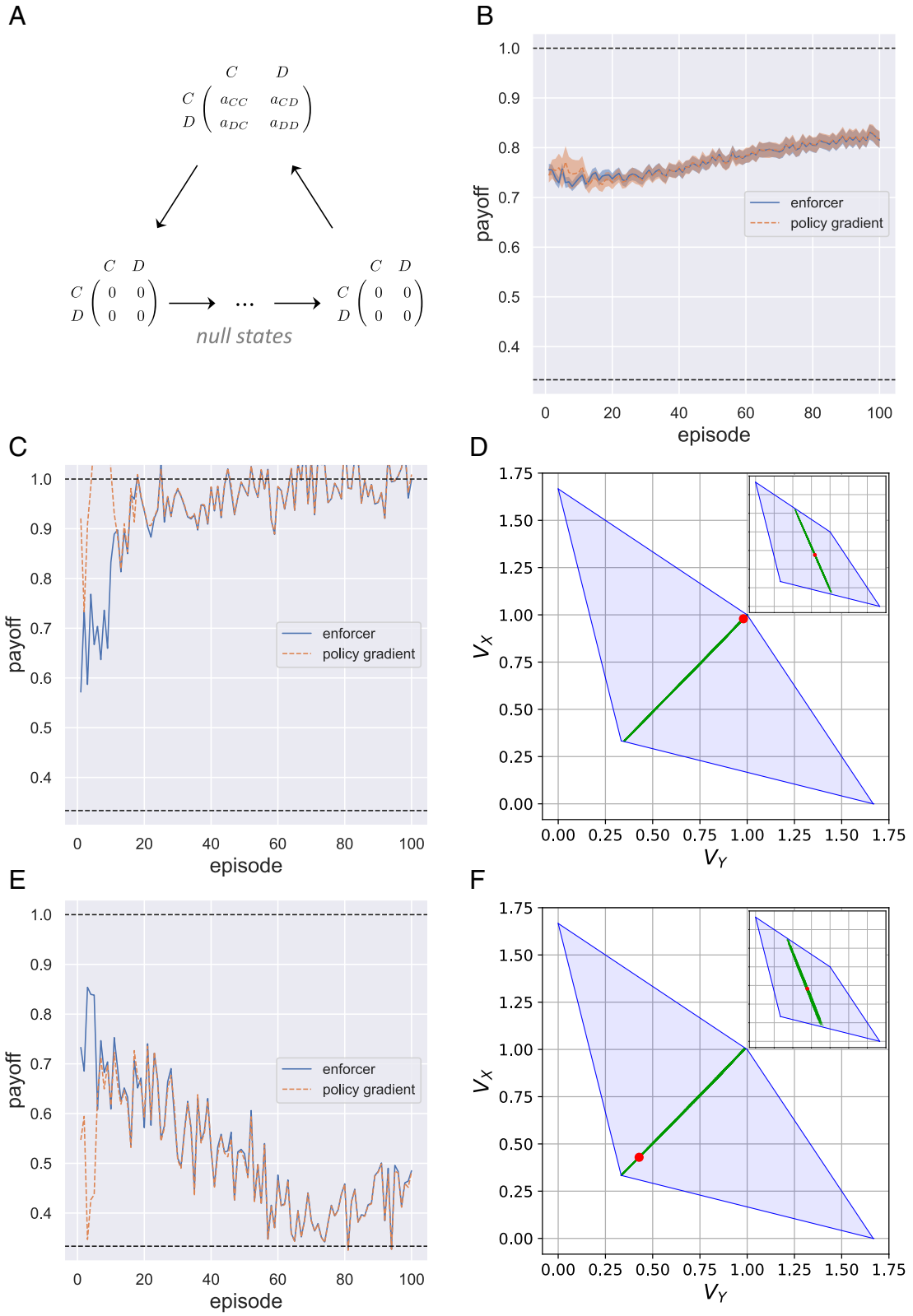


**Fig. 4.** Enforcer against policy gradient in a prisoner's dilemma with sparse rewards. (*A*) Following a standard prisoner's dilemma interaction, the agents transition through $m$ "null" states in which the reward is zero, regardless of the actions taken. When paired against a selfish agent optimizing using policy gradient, an enforcer tends to lead both agents toward higher payoffs, on average. (*B*) The mean payoff and SE over 100 runs. The slow growth in mean payoffs is due to some learning trajectories converging to the optimal outcome rapidly (*C* and *D*), while others flounder (*E* and *F*). However, this failure to quickly ascend to the optimal payoff is due to the nature of the selfish learner rather than the enforcer: (*D* and *F*) show the corresponding feasible regions after 100 episodes, which reveal that in both cases the enforcer has effectively aligned incentives. In these cases, policy gradient is not always able to efficiently optimize in the resulting payoff landscape. The *Insets* in *D* and *F* show the initial strategies of the enforcer. The green regions show $10^4$ random (Dirichlet) memory-one opponents. Parameters: $(a_{CC}, a_{CD}, a_{DC}, a_{DD}) = (3, 0, 5, 1)$ and $\lambda = \sqrt[3]{0.99}$.

game into a fully cooperative game? Once an agent arrives at a strategy that enforces approximately equal payoffs for the two agents, the game appears to be fully cooperative from the standpoint of payoffs. This agent must then stick to its enforcer strategy and rely on the (selfish) opponent to move both agents' payoffs toward larger payoffs. In doing so, the enforcer has transformed the nature of the interaction at a cost of effectively removing itself from the game.

Our approach is deliberately asymmetric with respect to the learning algorithms of the agents. Our goal was to understand the amount of unilateral control an agent can have on the joint incentives, not how various combinations of learning rules perform against one another. Requiring two agents to have the same kind of learning rule (e.g., both enforcers) would be a strong assumption and would require coordination, a central institution, or some form of imitation process between learners. By contrast, we assume only that the opponent uses a fixed policy (Figs. 2 and 3) or optimizes in a self-interested manner (Fig. 4), which are much weaker assumptions. Nevertheless, our focus on fairness is motivated, in part, by what happens when two extortioners face one another in the iterated prisoner's dilemma. In attempting to extort one another, both agents end up defecting, whereas fair and generous strategies can support much more cooperative when facing themselves.

Importantly, our approach is opponent-agnostic. The opponent is there primarily to generate and simulate play, but we are not necessarily concerned with the actual score the learner obtains against any one individual. Because the result is a region that shows payoffs against all possible opponents, once this region begins to resemble the learned strategies shown in each of our examples, we know the ultimate goal (in payoff space, at least) of any other opponent, provided the opponent is motivated to improve its own score. However, simple reinforcement learning algorithms like policy gradient do not always efficiently carve out a path to this shared goal. This behavior is expected, in both reinforcement learning (43) and evolutionary game theory (35, 44), in the sense that "bad" initial conditions can lead to suboptimal outcomes due to the nonconvex nature of the payoff landscape. But it is interesting to note that even bad initial conditions (Fig. 4*E*) still often allow an enforcer to successfully align incentives in the desired way (Fig. 4*F*). This is the main reason we have focused on pairing Algorithm 1 against fixed strategies in most of the examples. (Pairing an enforcer against a policy gradient agent in Fig. 2, for example, also leads to increasing payoffs for both agents, just as it does in Fig. 4*B*.) The goal of an enforcer is simply to align incentives; it is up to the other agent to figure out how to get to the optimal point (which is then optimal for both agents).

The explicit conditioning on past histories used in memory-one strategies differs from the memory implicit in the weights of a neural network [e.g., those arising in deep Q-learning (45)]. For example, if a policy network takes as input only the environmental component of a state, then parameter updates in this network are influenced by the trajectory of past state-action pairs, but whatever policy results from such a process ultimately still conditions on just the environmental state. When there is a single environmental state, such a policy is again just an unconditional mixed action, and such an unconditional behavior cannot generally enforce linear payoff relationship. With that said, one could apply our main theorem to memory-zero strategies, but the perspective would no longer necessarily be one of obtaining a "trained" strategy that can then be used against various kinds of selfish agents. Instead, we would have a learning process that would have to be carried out against every opponent. In *SI Appendix*, we discuss further how memory-zero strategies fit into the context of our main results.

There are many interesting open extensions related to learning zero-determinant strategies in spatially and temporally complex social dilemmas. Our focus is deliberately on "enumerated" stochastic games with a small number of states, where we can visualize the feasible region of a strategy. Even though simple, enumerated stochastic games naturally fail to capture all aspects of multiagent interactions, they are still sufficiently rich as testbeds for learning. For example, our approach has been based on attaining a target $\left(V_X^*, V_Y^*\right)$, which is chosen unilaterally. Although repeated games have "folk theorems" that characterize when such a target is attainable in equilibrium (2), the situation is more complicated in stochastic games—where the question of whether there exists a Nash equilibrium in which one agent wins is undecidable in general (46). Furthermore, in line with empirical game-theoretic analysis (30, 47), downsampling complex stochastic games into those with a small number of salient states may be a fruitful approach to understanding qualitative aspects of interactions in complex environments.

Although this study illustrates a productive case of using ideas from evolutionary game theory combined with multiagent reinforcement learning, there are still many unresolved problems related to zero-determinant strategies and cooperative AI more generally, especially for general stochastic games. Even just for repeated games without an external state, there is much that remains to be understood about the scope of enforcing zero-determinant strategies. Our contribution here involves three parts. First, we provide a theorem with a condition for the existence of autocratic strategies in stochastic games. This allows us to reduce a relationship on long-term payoffs to a condition on short-term rewards. Second, we use this theoretical result to define a learning objective, which can easily make use of standard function approximation techniques in machine learning. Finally, we demonstrate the performance of this learning algorithm on several representative examples of social dilemmas, showing that it can align incentives and effectively transform mixed-motivation settings into more cooperative interactions. Of course, there is no unique goal of multiagent learning (48), but aligning incentives in this way is a fruitful way to mitigate conflicts of interest.

There remains a lot of work to be done on the problem of incentive alignment in social dilemmas. In more complex environments, there are nuances to what even constitutes a social dilemma (30), and it remains to be seen when collapsing payoffs onto a smaller-dimensional space is an attainable objective (including in many-agent games). Even in the simplest possible setting of a repeated, two-agent, two-action game, the fact that a single agent can enforce a linear payoff relationship at all was one of the main surprises of Press and Dyson (6). We have stated our main theorem to allow for substantial flexibility, but there are engineering and implementation challenges that will arise in complex environments, such as controlling the relative learning rates and architectures of the policy and $\psi$ networks for the enforcer, as well as dealing with sparse rewards, credit-assignment problems, nondeterministic state transitions, and correlations between states as episodes transpire. We expect that existing techniques, including temporal reward smoothing (49) and experience replay (39), will be useful for extending our approach to more complex environments.

**Data, Materials, and Software Availability.** The environments used to study these games may be found at https://github.com/alexmcavoy/evoenv (50).

**ACKNOWLEDGMENTS.** We gratefully acknowledge the support from the European Research Council (Starting Grant 850529: E-DIRECT) and the Max Planck Society (C.H.), the European Research Council (Consolidator Grant 863818: ForM-SMArt) (K.C.), the Shanghai Pujiang Program (No. 23PJ1405500) (Q.S.), the Army Research Office (Grant No. W911NF-18-1-0325) (N.E.L.), and the John Templeton Foundation (Grant No. 62281) (J.B.P.).

Author affiliations: [a]School of Data Science and Society, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [b]Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [c]Department of Computer Science, Stanford University, Stanford, CA 94305; [d]JP Morgan AI Research, New York, NY 10017; [e]Max Planck Research Group: Dynamics of Social Behavior, Max Planck Institute for Evolutionary Biology, Plön 24306, Germany; [f]Institute of Science and Technology Austria, Klosterneuburg 3400, Austria; [g]Transdisciplinary Research Area: Sustainable Futures, University of Bonn, 53113 Bonn, Germany; [h]Center for Development Research (ZEF), University of Bonn, 53113 Bonn, Germany; [i]Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China; [j]Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China; [k]Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai 200240, China; [l]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544; [m]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104; and [n]Center for Mathematical Biology, University of Pennsylvania, Philadelphia, PA 19104

1. R. M. Dawes, Social dilemmas. *Annu. Rev. Psychol.* **31**, 169–193 (1980).
2. D. Fudenberg, E. Maskin, The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **54**, 533 (1986).
3. R. L. Trivers, The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57 (1971).
4. M. Nowak, K. Sigmund, A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* **364**, 56–58 (1993).
5. R. Axelrod, *The Evolution of Cooperation* (Basic Books, 1984).
6. W. H. Press, F. J. Dyson, Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10409–10413 (2012).
7. M. A. Nowak, *Evolutionary Dynamics: Exploring the Equations of Life* (Belknap Press, 2006).
8. S. Wang, H. Shi, Q. Hu, B. Lin, X. Cheng, Moving target defense for internet of things based on the zero-determinant theory. *IEEE Internet Things J.* **7**, 661–668 (2020).
9. Q. Hu *et al.*, Quality control in crowdsourcing using sequential zero-determinant strategies. *IEEE Trans. Knowl. Data Eng.* **32**, 998–1009 (2020).
10. A. K. Farraj, E. M. Hammad, A. Al Daoud, D. Kundur, "A game-theoretic control approach to mitigate cyber switching attacks in Smart Grid systems" in *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)* (IEEE, 2014).
11. H. Zhang, D. Niyato, L. Song, T. Jiang, Z. Han, Zero-determinant strategy for resource sharing in wireless cooperations. *IEEE Trans. Wireless Commun.* **15**, 2179–2192 (2016).
12. C. Hilbe, A. Traulsen, K. Sigmund, Partners or rivals? Strategies for the iterated prisoner's dilemma. *Games Econ. Behav.* **92**, 41–52 (2015).
13. G. Ichinose, N. Masuda, Zero-determinant strategies in finitely repeated games. *J. Theor. Biol.* **438**, 61–77 (2018).
14. A. McAvoy, C. Hauert, Autocratic strategies for iterated games with arbitrary action spaces. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 3573–3578 (2016).
15. A. J. Stewart, T. L. Parsons, J. B. Plotkin, Evolutionary consequences of behavioral diversity. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E7003–E7009 (2016).
16. C. Hilbe, B. Wu, A. Traulsen, M. A. Nowak, Cooperation and control in multiplayer social dilemmas. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 16425–16430 (2014).
17. A. J. Stewart, J. B. Plotkin, Small groups and long memories promote cooperation. *Sci. Rep.* **6**, 26889 (2016).
18. M. Ueda, Memory-two zero-determinant strategies in repeated games. *R. Soc. Open Sci.* **8**, 202186 (2021).
19. A. Mamiya, G. Ichinose, Zero-determinant strategies under observation errors in repeated games. *Phys. Rev. E* **102**, 032115 (2020).
20. E. Akin, What you gotta know to play good in the iterated prisoner's dilemma. *Games* **6**, 175–190 (2015).
21. C. Hilbe, K. Chatterjee, M. A. Nowak, Partners and rivals in direct reciprocity. *Nat. Hum. Behav.* **2**, 469–477 (2018).
22. X. Chen, F. Fu, Outlearning extortioners: Unbending strategies can foster reciprocal fairness and cooperation. *PNAS Nexus* **2**, pgad176 (2023).
23. C. Hilbe, M. A. Nowak, A. Traulsen, Adaptive dynamics of extortion and compliance. *PLoS ONE* **8**, e77886 (2013).
24. C. Adami, A. Hintze, Evolutionary instability of zero-determinant strategies demonstrates that winning is not everything. *Nat. Commun.* **4**, 2193 (2013).
25. X. Chen, L. Wang, F. Fu, The intricate geometry of zero-determinant strategies underlying evolutionary adaptation from extortion to generosity. *N. J. Phys.* **24**, 103001 (2022).
26. L. S. Shapley, Stochastic games. *Proc. Natl. Acad. Sci. U.S.A.* **39**, 1095–1100 (1953).
27. A. Rubinstein, Finite automata play the repeated prisoner's dilemma. *J. Econ. Theory* **39**, 83–96 (1986).
28. N. Masuda, H. Ohtsuki, A theoretical analysis of temporal difference learning in the iterated prisoner's dilemma game. *Bull. Math. Biol.* **71**, 1818–1850 (2009).
29. E. Hughes *et al.*, "Inequity aversion improves cooperation in intertemporal social dilemmas" in *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18* (Curran Associates Inc., Red Hook, NY, 2018), pp. 3330–3340.
30. J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, T. Graepel, "Multi-agent reinforcement learning in sequential social dilemmas" in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS '17* (2017), pp. 464–473.
31. J. Stastny *et al.*, Normative disagreement as a challenge for cooperative AI. arXiv [Preprint] (2021). https://arxiv.org/abs/2111.13872 (Accessed 1 October 2023).
32. C. Hilbe, L. A. Martinez-Vaquero, K. Chatterjee, M. A. Nowak, Memory-*n* strategies of direct reciprocity. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4715–4720 (2017).
33. J. Nash, Non-Cooperative Games. *Ann. Math.* **54**, 286 (1951).
34. A. Lerer, A. Peysakhovich, Maintaining cooperation in complex social dilemmas using deep reinforcement learning. arXiv [Preprint] (2017). https://arxiv.org/abs/1707.01068 (Accessed 1 October 2023).
35. A. Mirani, A. McAvoy, Payoff landscapes and the robustness of selfish optimization in iterated games. *J. Math. Biol.* **84**, 55 (2022).
36. A. McAvoy, J. Kates-Harbeck, K. Chatterjee, C. Hilbe, Evolutionary instability of selfish learning in repeated games. *PNAS Nexus* **1**, pgac141 (2022).
37. A. McAvoy, C. Hauert, Autocratic strategies for alternating games. *Theor. Popul. Biol.* **113**, 13–22 (2017).
38. V. Konda, J. Tsitsiklis, "Actor–critic algorithms" in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, K. Müller, Eds. (MIT Press, 1999), vol. 12.
39. V. Mnih *et al.*, Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
40. A. J. Stewart, J. B. Plotkin, From extortion to generosity, evolution in the Iterated Prisoner's Dilemma. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15348–15353 (2013).
41. R. S. Sutton, D. McAllester, S. Singh, Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation" in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, K. Müller, Eds. (MIT Press, 2000), vol. 12.
42. K. Zhang, Z. Yang, T. Başar, Multi-agent reinforcement learning: A selective overview of theories and algorithms. arXiv [Preprint] (2019). https://arxiv.org/abs/1911.10635 (Accessed 1 October 2023).
43. P. Henderson *et al.*, "Deep reinforcement learning that matters" in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18* (AAAI Press, 2018).
44. J. Chen, A. Zinger, The robustness of zero-determinant strategies in iterated prisoner's dilemma games. *J. Theor. Biol.* **357**, 46–54 (2014).
45. S. Gronauer, K. Diepold, Multi-agent deep reinforcement learning: A survey. *Artif. Intell. Rev.* **55**, 895–943 (2021).
46. M. Ummels, D. Wojtczak, The complexity of Nash equilibria in stochastic multiplayer games. *Log. Methods Comput. Sci.* **7**, 20 (2011).
47. W. E. Walsh, R. Das, G. Tesauro, J. O. Kephart, "Analyzing complex strategic interactions in multi-agent systems" in *AAAI-02 Workshop on Game-Theoretic and Decision-Theoretic Agents, AAAI '02* (2002), pp. 109–118.
48. Y. Shoham, R. Powers, T. Grenager, If multi-agent learning is the answer, what is the question? *Artif. Intell.* **171**, 365–377 (2007).
49. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (The MIT Press, ed. 2, 2018).
50. A. McAvoy, Evoenv: Environments for evolutionary games. GitHub. https://github.com/alexmcavoy/evoenv. Deposited 12 May 2024.