Emergence of Punishment in Social Dilemma with Environmental Feedback

Zhen Wang[1,2,*], **Zhao Song** [1,2,*], **Chen Shen**[3], **Shuyue Hu**[4]

[1]School of Mechanical Engineering, Northwestern Polytechnical University
[2]School of Artifcial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University
[3]Faculty of Engineering Sciences, Kyushu University
[4]Shanghai Artifcial Intelligence Laboratory
songzhao@mail.nwpu.edu.cn, hushuyue@pjlab.org.cn

Abstract

Altruistic punishment (or punishment) has been extensively shown as an important mechanism for promoting cooperation in human societies. In AI, the emergence of punishment has received much recent interest. In this paper, we contribute with a novel evolutionary game theoretic model to study the impacts of environmental feedback. Whereas a population of agents plays public goods games, there exists a third-party population whose payoffs depend not only on whether to punish or not, but also on the state of the environment (e.g., how cooperative the agents in a social dilemma are). Focusing on one-shot public goods games, we show that environmental feedback, by itself, can lead to the emergence of punishment. We analyze the co-evolution of punishment and cooperation, and derive conditions for their co-presence, co-dominance and co-extinction. Moreover, we show that the system can exhibit bistability as well as cyclic dynamics. Our findings provide a new explanation for the emergence of punishment. On the other hand, our results also alert the need for careful design of implementing punishment in multi-agent systems, as the resulting evolutionary dynamics can be somewhat complex.

Introduction

Altruistic punishment (or punishment) has been proved to be an important mechanism for promoting cooperation in human societies as well as in AI systems (Dreber et al. 2008; Bou, López-Sánchez, and Rodríguez-Aguilar 2006). Laboratory and field data show that human subjects across different cultures are eager to punish non-cooperators in social dilemma games even if this incurs a cost (Fehr and Gächter 2002). In AI, punishment has also gained much interest (Morris-Martin, De Vos, and Padget 2019; Giardini et al. 2014; Pereira et al. 2017). On one hand, punishment is a conceptually appealing tool to promote cooperative behaviours and to enforce norm compliance among self-regarding agents (Mahmoud et al. 2015; Villatoro et al. 2011). On the other hand, it has been implemented in various multi-agent systems, such as e-marketplace (Liu et al. 2016), online virtual agent societies (Savarimuthu, Padget, and Purvis 2013), and smart grids (Du et al. 2021); this renders the study of punishment to be of practical interest.

Although common in both human societies and AI systems, punishment, by itself, creates an evolutionary puzzle—how punishment could possibly emerge given that punishment is costly (i.e. punishing others generally incurs a cost to punishers themselves). To address this puzzle, a number of models that leverage *reciprocity* (Raihani and Bshary 2015; Helbing et al. 2010; Nowak 2006), *voluntary participation* (Semmann, Krambeck, and Milinski 2003), and *prior commitment* (Han, Pereira, and Lenaerts 2017; Han 2016) have been proposed to account for the emergence of punishment. However, these models typically assume that *environmental feedback* does not exist, let alone influence punishment; put differently, whether to punish or not is independent of the state of the *endogenous* environment.

Different from the absence of environmental feedback in most existing models, real-world systems, however, often feature environmental feedback (Tilman, Plotkin, and Akçay 2020; Weitz et al. 2016): while strategic decisions of individuals may change the environment, the environment may the other way around shape their decision making. For example, while global climate change is under the influence of whether nations ratify an (e.g., Paris) agreement and be fined when failing to meet the goals, long-term ecological damage may conversely alter their strategic incentives (Tilman, Plotkin, and Akçay 2020). Likewise, as the well-known *broken window experiments* (Zimbardo 1969) hint, any visible sign of disorder (e.g., norm violation, anti-social behaviours) which goes untended encourages further disorder; this may, in turn, silence social punishment (e.g., anger, social exclusion) in the environment (Wanders et al. 2021). Examples from these diverse fields suggest that the effect of environmental feedback on punishment is non-trivial and perhaps crucial under some scenarios as they may fundamentally change dynamical predictions. This drives our motivating questions:

Can environmental feedback, by itself, lead to the emergence of punishment in absence of any other mechanisms previously studied? Under what conditions does environmental feedback promote or preclude punishment? How much does environmental feedback affect the interplay between punishment and cooperation?

To address these questions, we propose a novel model based on evolutionary game theory. We focus on the one-shot public goods game (PGG) with third-party punish-

[*]These authors contributed equally.

ment. More specifically, there co-exist a game-playing population and a third-party population. For every time step, anonymous agents in the game-playing population are randomly drawn to play a PGG game and choose to cooperate or not. In the meantime, the third-party agents oversee the game plays and can choose to punish non-cooperators with a cost. One-shot PGG is a widely adopted strategic interaction framework for social dilemmas, and also arguably stands for the most challenging benchmark for punishment as well as cooperation to emerge (Archetti and Scheuring 2012). Third-party punishment, which can promote cooperation by allowing third-party agents to punish defectors (Fehr and Fischbacher 2004), brings the question of why punishment can emerge when bearing the invariable cost.

Different from prior studies on PPG with third-party punishment, we consider the effects of environmental feedback on third-party agents and that such effects jointly depend (i) on the amount of cooperation in the environment, and (ii) on the endogenous incentives for punishment in the environment. Intuitively, in a cooperative environment, third-party agents are incentivized to punish non-cooperative behaviors if they exist, as sustaining cooperation requires only few efforts. On the other hand, the strength of such incentives generally varies in different environments; for example, experimental works have shown that there are marked differences in the willingness to engage in costly punishment across difficult cultures (Henrich et al. 2006; Herrmann, Thoni, and Gachter 2008).

We analyze the effects of environmental feedback on punishment and on the interplay between punishment and cooperation, assuming that the system evolves according to the replicator dynamics in evolutionary game theory. We provide an affirmative answer to our key motivating question — environmental feedback, which is beyond the scope of previous mechanisms, provides a pathway to achieve the emergence of punishment. Moreover, we show that under the sole influence of environmental feedback, punishment and cooperation generally co-emerge. In particular, it is possible that a system will always eventually evolve into the state of complete punishment (all third-party agents choosing punishment) and complete cooperation (all game players choosing cooperation), regardless of its initial state (Theorem 1). Complete punishment, which represents the most stringent punishment system, however, is *not* a necessary condition for complete cooperation (Theorem 2). Conversely, complete cooperation will never occur in absence of punishment, whereas cooperation will never go extinct as long as there still exists punishment (Theorem 3). It is also interesting to note that taking into account environmental feedback expands the suite of dynamical possibilities in the co-evolution of punishment and cooperation. We showcase that it is possible for the phenomena of bistability and cyclic dynamics exist and analyze their conditions (Theorems 5 and 6). From the perspective of system design, this alerts the need for careful design of multi-agent sanctioning systems as the arising evolutionary dynamics can be somewhat complex. Last but not least, we corroborate our theoretical findings with numerical simulations on finite populations that evolve according to the Fermi process.

Our key contributions are summarized as follows.

- We propose a novel evolutionary game theoretic framework to model the effects of environmental feedback.
- We show that environmental feedback, by itself, can lead to the emergence of punishment. This paves a new way to understand the source of punishment.
- We analyze the co-evolution of punishment and cooperation, deriving the conditions for their co-presence, codominance and co-extinction along with the conditions for bistability and persistent cycles.

In the following, we discuss related work in Section 2. We present our evolutionary game theoretic model that takes into account environmental feedback in Section 3. In Section 4, we analyze the resulting evolutionary dynamics. We conclude this paper and discuss directions for future work in Section 5. The numerical simulations on finite populations and detailed proofs of our theoretical claims are presented in the supplementary[1] due to the lack of space.

Related Work

Research on the emergence of punishment involves a significant literature (Fehr and Gächter 2002; Dreber et al. 2008; Han 2016). A major approach is to leverage typical reciprocity mechanisms (e.g., indirect reciprocity or reputation, direct reciprocity, network reciprocity) (Raihani and Bshary 2015; Helbing et al. 2010; Campos et al. 2008), with few notable exceptions (Semmann, Krambeck, and Milinski 2003; Han, Pereira, and Lenaerts 2017). Among these, prior works typically assume that third-party punishment is invariably costly and must co-exist with other mechanisms (Jordan, McAuliffe, and Rand 2016; Mathew and Boyd 2011), such as strong reciprocity (Fehr and Fischbacher 2004) and reputation (Jordan et al. 2016). However, these works have not considered nor analyzed the influence of environmental feedback. To our knowledge, our paper is the first work that theoretically shows environmental feedback contributes to the emergence of punishment.

On the other hand, the study of environmental feedback is not entirely new. In their pioneering work, Weitz et al. (2016) propose a unified approach to analyzing feedback-evolving games and show that environmental feedback can cause an oscillatory tragedy of the commons. Tilman, Plotkin and Akçay (2020) suggest that the joint dynamics of strategies and the environment rely (i) on the incentives for individuals to lead or follow behavioral changes, and (ii) on the relative speed of environmental versus strategic change. Levy and Griffiths (2021) propose a framework where agents' reward exogenously depend on the strategy of others and the environment, enabling the emergence of social norms. While environmental feedback has recently received increasing interest, it has never been explored in the co-presence with punishment.

Last but not least, it is important to note that punishment has also been extensively studied in the literature of normative multi-agent systems (Lenaerts et al. 2015; Pereira et al. 2017; Bench-Capon and Modgil 2017; Santos et al. 2019;

[1] https://github.com/yt-songz/AAAI2023SI

Pynadath and Marsella 2005; Koppol, Admoni, and Simmons 2021; Mahmoud et al. 2012). These works typically aim at using punishment to regulate individual and collective behaviors, formalizing different relevant aspects of these mechanisms (e.g., norms and conventions) in a multi-agent system. Our work complements these studies, addressing the evolutionary puzzle of punishment and providing a new explanation for the source of punishment.

An Evolutionary Game Theoretic Model

In this section, we recall public goods games and present our evolutionary game theoretic model that characterizes environmental feedback in the context of third-party punishment.

Public Goods Game

A public good is a common resource shared among individuals regardless of their contributions (Kollock 1998). The management of these goods naturally results in a social dilemma — whereas all individuals would be better off cooperating (i.e. making joint contributions to a public pool), each individual has an incentive to free-ride on the contributions of others. Public Goods Games (PGG) is a typical model for capturing this social dilemma.

In a standard PGG with G players, each agent has two strategies: to cooperate (C) or to defect (D). Playing C results in a contribution c to the public pool, while playing D means a free rider and makes no contribution. After playing a PGG, the total contributions in the public pool are multiplied by a synergy factor r and are then equally distributed among the G players. Let π_C and π_D denote the payoffs of cooperation and defection, respectively. They are given by

$$\begin{aligned} \pi_C &= rc\frac{n_c}{G} - c \\ \pi_D &= rc\frac{n_c}{G} \end{aligned} \tag{1}$$

where n_c is the number of cooperators. Evidently, defection always yields a higher payoff for each individual. However, the group's total payoffs are maximized when every individual contributes to the public pool.

Population Setup and Punishment

In this paper, we consider two infinitely large, well-mixed, anonymous agent populations: a game-playing population and a third-party population. In the game-playing population, for each time step, G agents are randomly chosen to play a PGG. Note that agents generally do not meet with the same opponents twice nor have prior knowledge (e.g., reputation) about other agents at a given time step. Hence, this is PPG under one-shot setting (Archetti and Scheuring 2012). In the third-party population, agents do not play games against one another, nor directly participate in any PGG in the game-playing population. Rather, for each time step, G agents are randomly chosen from the third-party population, and each of them oversees a player of the PGG in the game-playing population.

Specifically, before the PGG is played in the game-playing population, each chosen third-party agent decides to punish (P) or not (N). After the PGG is played, the chosen third-party agents will execute their strategies decided prior to the game play. Put differently, for each time step, there are two stages: third-party agents choose to punish or not at the first stage, and agents in the game-playing population play the PGG actually at the second stage.

Punishment is costly. If a third-party agent decides to punish, the punisher needs to pay a cost α if the PPG player whom it oversees defects. Meanwhile, such punishment will take an effect and cause a fine β to the PPG player only if the PPG player defects.

Environmental Feedback

We now extend the model with environmental feedback. Note that typically, evolutionary game theoretic analysis assumes that the nature of the strategic interaction is fixed in time, or that it depends on the state of an independent, exogenous environment. This assumption has been critiqued that there often exists co-evolution of the environment and individual strategic decisions in many real-world systems (Sigdel, Anand, and Bauch 2019; Hilbe et al. 2018). The concept of environmental feedback, which is proposed to address this limitation, features bi-direction feedback — while the strategies of individuals may alter the state of the environment, the evolution of the environment may conversely feedback to change the incentive structure of strategic decisions (Tilman, Plotkin, and Akçay 2020; Weitz et al. 2016).

In the context of punishment, evidently, whether third-party agents punish or not will affect the endogenous environment of PPG, as defectors will be fined if the third-party agents choose to punish. Taking into account environmental feedback, we consider that the endogenous environment of PPG — how cooperative the players are — will conversely influence the payoffs of third-party agents. Specifically, we define the environmental payoff of punishment (or not) to be s_{PC} (or s_{NC}), given that the game player whom it oversees cooperates. Likewise, we define the environmental payoff of punishment (or not) to be s_{PD} (or s_{ND}), given that the game player whom it oversees defects. We summarize the payoffs of game players and third-party agents in each case as follows:

- the game-player cooperates and the third-party agent punishes: $\pi_C = \frac{rcn_c}{G} - c$, $\pi_P = s_{PC}$
- the game-player cooperates and the third-party agent not punish: $\pi_C = \frac{rcn_c}{G} - c$, $\pi_N = s_{NC}$
- the game-player defects and the third-party agent punishes: $\pi_D = \frac{rcn_c}{G} - \beta$, $\pi_P = s_{PD} - \alpha$
- the game-player defects and the third-party agent not punishes: $\pi_D = \frac{rcn_c}{G}$, $\pi_N = s_{ND}$

Compared with Equation 1, the payoff of cooperation remains unchanged even if third-party agents choose punishment. On the other hand, the payoff of defection will be reduced by the fine β if third-party agents choose punishment.

We remark that we only require the values of $s_{PC}, s_{NC}, s_{PD}, s_{ND}$ to be real values; in other words, the third-party agents can receive negative environmental payoffs. It is interesting to note that the difference in these values characterizes the endogenous incentive of punishment in the environment. To see this, let us define

$$\delta_C := s_{PC} - s_{NC}, \qquad \delta_D := s_{PD} - s_{ND}. \tag{2}$$

The value of δ_C measures the difference in environmental payoff between punishment and non-punishment, when the game player cooperates; put simply, it is the gain of punishment given cooperation. Intuitively, a positive value of δ_C suggests that punishment is encouraged and more incentivized in a cooperative environment. On the contrary, a negative value of δ_C suggests that punishment is discouraged in a cooperative environment. The same reasoning can be applied to δ_D which represents the gain of punishment given defection.

Replicator Dynamics

In this paper, we assume that the game-playing population and the third-party population evolve according to the replicator dynamics. The replicator dynamics is a widely used model in evolutionary game theory to express how the frequencies of strategies in a population evolve over time (Hofbauer, Sigmund et al. 1998). It is based on the basic idea that the proportion of agents of a given strategy increases when the strategy achieves expected payoffs higher than the average payoff, and decreases when achieving expected payoffs lower than the average payoff. Formally, the replicator dynamics is given by the differential equation

$$\dot{\rho}_i = \rho_i(E[\pi_i] - \sum_{i \in \{C,D,P,N\}} \rho_i E[\pi_i]), \quad i \in \{C,D,P,N\} \tag{3}$$

where ρ_i is the proportion of agents using strategy i and $E[\pi_i]$ is the expected payoff of strategy i. Note that we have $\rho_C + \rho_D = 1$ and $\rho_P + \rho_N = 1$ given our population setup.

To calculate the expected payoff of cooperation, we observe that given the proportion of cooperators (denoted by ρ_C), having m cooperators in a PPG follows a binomial distribution. Therefore, the expected payoff of cooperation is

$$\begin{aligned} E[\pi_C] &= \sum_{m=0}^{G-1} \binom{G-1}{m} \rho_C^m (1-\rho_C)^{G-1-m} (rc\frac{m+1}{G} - c) \\ &= (G-1)\rho_C \frac{rc}{G} + (\frac{rc}{G} - c) \end{aligned} \tag{4}$$

Likewise, the expected payoff of defection without punishment is

$$\begin{aligned} E[\pi_D]' &= \sum_{m=0}^{G-1} \binom{G-1}{m} \rho_C^m (1-\rho_C)^{G-1-m} rc\frac{m}{G} \\ &= E[\pi_C] - (\frac{rc}{G} - c) \end{aligned} \tag{5}$$

and the expected payoff of defection with punishment is

$$E[\pi_D] = E[\pi_D]' - \beta\rho_P. \tag{6}$$

Regarding the expected payoff of punishment or not, recall that the payoff of a third-party agent depends on its strategy and on the strategy of the game player whom it oversees. Thus, the expected payoffs are given by

$$\begin{aligned} E[\pi_P] &= \rho_C s_{PC} + (1-\rho_C)s_{PD} - (1-\rho_C)\alpha, \\ E[\pi_N] &= \rho_C s_{NC} + (1-\rho_C)s_{ND}. \end{aligned} \tag{7}$$

Overall, the system of differential equations that express the evolutionary dynamics of the game-playing population and the third-party population is

$$\begin{aligned} \dot{\rho}_C &= \rho_C(1-\rho_C)(\frac{rc}{G} - c + \beta\rho_P) \\ \dot{\rho}_P &= \rho_P(1-\rho_P)\left[(s_{PC} - s_{NC})\rho_C + (s_{PD} - s_{ND} - \alpha)(1-\rho_C)\right]. \end{aligned} \tag{8}$$

Co-Evolution of Punishment and Cooperation

In this section, we present our study on the system evolution by analyzing the replicator dynamics presented in Equation (8). Throughout our analysis, we assume that $\frac{r}{G} < 1$ and $\frac{rc}{G} - c + \beta > 0$. The former is standard in theoretical models for PPG (Archetti and Scheuring 2012). If $\frac{r}{G} \geq 1$, as shown in Equation (5), the expected payoff of cooperation would be at least as high as that of defection, suggesting that social dilemmas would no longer exist. Regarding the latter, observe from Equation (6) that the difference in the expected payoff between cooperation and defection is $E[\pi_C] - E[\pi_D] = \frac{rc}{G} - c + \beta\rho_P$. Imagine that $E[\pi_C] - E[\pi_D] \leq 0$ given $\rho_P = 1$. This means that even if all the third-party agents punish, the expected payoff of defection would still be higher than or equal to that of cooperation; in other words, punishment has no effect on promoting cooperation. Evidently, this assumption contradicts numerous empirical findings. Hence, we assume that $E[\pi_C] - E[\pi_D] > 0$ given $\rho_P = 1$, which leads to $\frac{rc}{G} - c + \beta > 0$.

To start with, we derive the equilibrium points (or states) of the system based on the replicator dynamics (Equation (8)). An equilibrium state should satisfy one of the five following cases:

- $\rho_C = 0, \rho_P = 0$, i.e. co-extinction of C and P,
- $\rho_C = 0$, $\rho_P = 1$, i.e. extinction of C but dominance of P,
- $\rho_C = 1$, $\rho_P = 0$, i.e. dominance of C but extinction of P,
- $\rho_C = 1$, $\rho_P = 1$, i.e. co-dominance of C and P,
- $\rho_C = \frac{\delta_D - \alpha}{\delta_D - \alpha - \delta_C}$, $\rho_P = -\frac{1}{\beta}(\frac{rc}{G} - c)$, i.e. co-existence of C, D, P, N,

where the last case exists if and only if $\delta_C \neq \delta_D - \alpha$ and $r/G \leq 1$.

If a system starts at an equilibrium state, it will remain there thereafter. Yet we are also interested in the system evolution when it is off equilibrium. To see this, we analyze the stability of these equilibrium points using Lyapunov's indirect method. We visualize a schematic representation of our results in Figure 1. In the following subsections, we elaborate on and discuss our key findings.

Co-Emergence of Punishment and Cooperation

We provide an affirmative answer to our key motivating question "*can environmental feedback, by itself, lead to the emergence of punishment in absence of any other mechanism previously studied?*" In the following theorem, we show that even the state of complete punishment and complete cooperation can be achieved.

Theorem 1 *The equilibrium state $\rho_P = 1, \rho_C = 1$ is the unique asymptotically stable state if $\delta_C > 0$, $\delta_D > \alpha > 0$.*

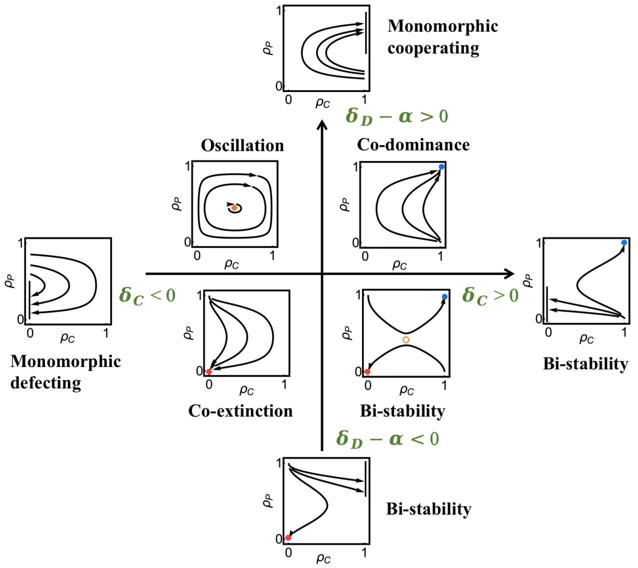

Figure 1: A schematic representation of the co-evolution of punishment and cooperation with environmental feedback.

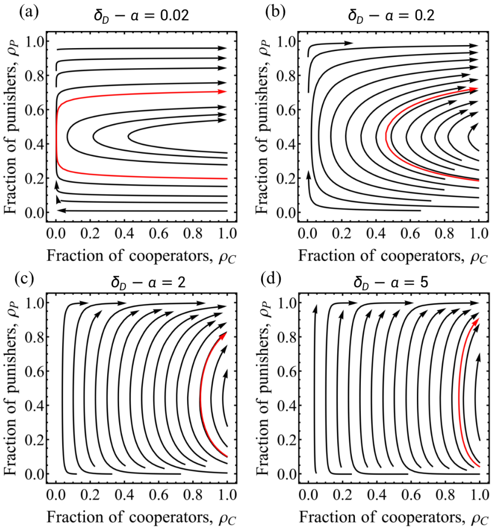

Figure 2: Phase portrait with different values of $\delta_D - \alpha$, given $\delta_C = 0$.

This theorem makes it clear that given positive values of δ_C and $\delta_D - \alpha$, the system will always evolve into the state in which all the third-party agents implement punishment and all the game players cooperate, regardless of initial conditions. Note that this condition makes *no* assumption that punishment should be rewarding (the environmental payoffs of punishment or not can be negative). Rather, it requires (i) that the environmental payoff of punishment is always better than that of non-punishment ($\delta_C, \delta_D > 0$), and (ii) that if game players defect, the environmental payoff of punishment is sufficiently better than that of non-punishment so that the gain of punishment can offset the cost ($\delta_D > \alpha$).

The rationale behind this theorem can be explained as follows. As punishment is always more incentivized in the environment ($\delta_C, \delta_D > 0$) and its gain can even offset its cost ($\delta_D > \alpha$), intuitively, punishment will dominate the third-party population. The increase in the amount of punishment will, in turn, significantly decrease the payoff of defection; consequently, cooperation will prevail and dominate the game-playing population.

Intuitively, the result of Theorem 1 implies the most stringent punishment system — if a defection occurs, the defector would be punished with absolute certainty (since all the third-party agents are punishers). It is then natural to ask: *to achieve complete cooperation within the game-playing population, is it necessary to have punishment imposed with absolute certainty?* We show in the following theorem that this is indeed *unnecessary*.

Theorem 2 *There only exists a continuum of stable equilibrium state* $\rho_C = 1$, $\rho_P = x$ *with* $x \in (-\frac{1}{\beta}(\frac{rc}{G} - c), 1]$ *if* $\delta_C = 0$, $\delta_D > \alpha > 0$.

This theorem states that under the condition $\delta_C = 0, \delta_D > \alpha > 0$, the system will eventually evolve into a state in which all the agents in the game-playing population cooperate whereas a proportion of third-party agents actually do *not* punish. Hence, complete punishment is not a necessary condition for complete cooperation.

Compared to the condition of Theorem 1, Theorem 2 differs in a zero value of δ_C, i.e. third-party agents receive exactly the same payoff by punishment and by non-punishment if game players cooperate. Because of this difference, third-party agents have no incentive to punish in a cooperative environment, and punishment can flourish only if defection exists. As a result, the amount of punishment will remain unchanged once complete cooperation is reached.

In Figure 2, we visualize the phase portrait with different values of $\delta_D - \alpha$ given $\delta_C = 0$. We gradually increase the value of $\delta_D - \alpha$ from 0.02 to 5. The trajectories that start with the same initial state $\rho_C = 0.9, \rho_P = 0.2$ are marked in red. Comparing Figure 2 (a-d), although converging to a similar state eventually, these trajectories are very distinct. A high value of $\delta_D - \alpha$, which represents a strong incentive of punishment in a non-cooperative environment, leads to an abrupt change in the amount of punishment. On the other hand, with a small value of $\delta_D - \alpha$, the amount of punishment slowly evolves most of the time.

Putting Theorems 1 and 2 together, we observe the following corollary about complete cooperation.

Corollary 1 *Complete cooperation* $\rho_C = 1$ *will be always achieved regardless of initial system states if* $\delta_C \geq 0, \delta_D > \alpha > 0$.

The above results have shown that complete cooperation is bound to occur under certain conditions of environmental feedback. Theorem 2 even suggests that a significant proportion of punishers suffices to achieve complete cooperation. Naturally, one may ask: *taking one step further, can complete cooperation be achieved in absence of punishment?* In the following theorem, We show that this is impossible.

Theorem 3 *The equilibrium states* $\rho_C = 1, \rho_P = 0$, *and*

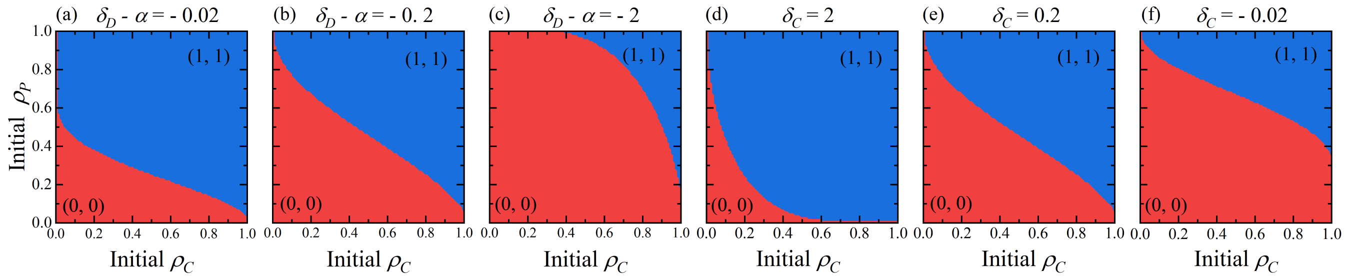

Figure 3: Region of attraction to the equilibrium states ($\rho_C = 0, \rho_P = 0$) and ($\rho_C = 1, \rho_P = 1$) given different values of $\delta_C > 0$ and $\delta_D - \alpha < 0$. In (a)-(c), $\delta_C = 0.2$, and in (d)-(f), $\delta_D - \alpha = -0.2$.

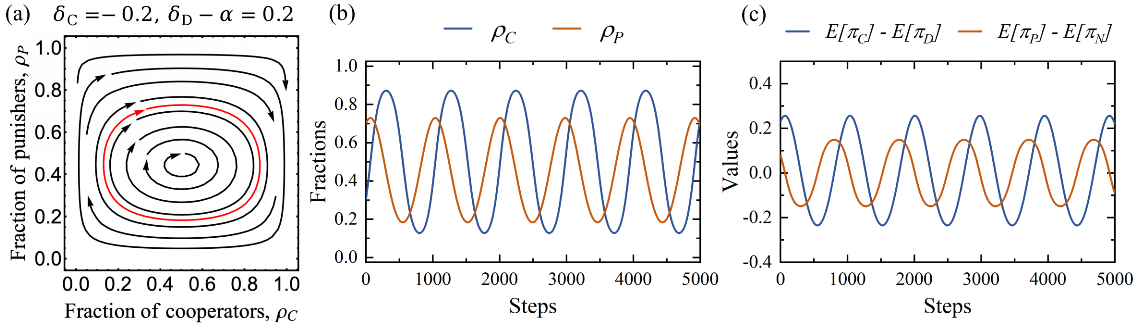

Figure 4: Illustration of cyclic dynamics. Panel (a) depicts a phase portrait given $\delta_C = -0.2$ and $\delta_D - \alpha = 0.2$. Panels (b) and (c) depict the dynamics of strategy frequency and payoff, respectively, corresponding to the cycle that marked in red in panel (a).

$\rho_C = 0, \rho_P = 1$ are always unstable.

We emphasize that the instability of these two equilibrium points stands, no matter what environmental feedback and initial system states are. Therefore, this theorem indicates that complete cooperation will never occur if there is no punishment in the system; on the other hand, cooperation will never become extinct as long as there still exists punishment in the system. Put differently, through environmental feedback, the emergence of cooperation is closely tied with the emergence of punishment.

We evidence more on this close connection through environmental feedback in the following theorem.

Theorem 4 *It is possible for a system to rest in the co-presence of punishment and cooperation, $\rho_C > 0, \rho_P > 0$, if $\delta_C \geq 0$ (except $\delta_C = 0, \delta_D = \alpha$) or $\delta_D > \alpha > 0$. Conversely, the equilibrium state $\rho_C = 0, \rho_P = 0$ is the unique asymptotically stable state if $\delta_C < 0$, $\delta_D < \alpha$.*

We remark that the condition for the co-presence of punishment and cooperation is not stringent, as it only requires either a positive gain (δ_C) of punishment in a cooperative environment, *or* a sufficiently high gain (δ_D) of punishment in a non-cooperative environment which can offset the cost.

Richness in Evolutionary Dynamics: Bistability and Persistent Cycles

We have shown that environmental feedback can cause the co-emergence of punishment and cooperation. In this subsection, we show that environmental feedback also expands the suite of dynamical possibilities. In particular, we observe that bistability and persistent cycles are able to arise from the system evolution, as a result of environmental feedback. These phenomena are non-trivial, especially when taking into account that our mechanism of environmental feedback is not complex.

In the following theorem, we derive the conditions that permit bistability.

Theorem 5 *There co-exist two stable equilibrium states $\rho_C = 0, \rho_P = 0$ and $\rho_C = 1, \rho_P = 1$, along with a saddle point $\rho_C = \frac{\delta_D - \alpha}{\delta_D - \alpha - \delta_C}$, $\rho_P = -\frac{1}{\beta}\left(\frac{rc}{G} - c\right)$ if $\delta_C > 0, \delta_D < \alpha, \alpha > 0$.*

Here the saddle point is an interior fixed point representing that cooperation and defection co-exist in the game-playing population, while punishment and non-punishment also co-exist in the third-party population. This theorem shows that given $\delta_C > 0$ and $\delta_D < \alpha$, the system will rest in either the state of complete cooperation and complete punishment or the state of complete defection and complete non-punishment. Yet which particular state that a system will eventually evolve into generally depends on the initial state.

We visualize the bistability phenomenon in Figure 3. In particular, we color the region of attraction to the state $\rho_C = 0, \rho_P = 0$ in red, and the counterpart to the state $\rho_C = 1, \rho_P = 1$ in blue. It is shown that as the values of δ_C and δ_D increase, a larger range of initial system states will eventually converge to the state of co-dominance of punishment and cooperation. This can be expected, since the increase in δ_C and δ_D indicates that agents are more incentivized to punish.

Next, we turn to cyclic dynamics. The conditions that permit persistent cycles are presented in the following theorem.

Theorem 6 *All the boundary equilibrium states are unstable, the interior equilibrium states is neutrally stable, and cyclic dynamics exist if* $\delta_C < 0, \delta_D > \alpha > 0$.

We visualize such cyclic dynamics in Figure 4. As shown in the phase portrait (Figure 4 (a)), there exists numerous clockwise cycles surrounding the neutrally stable point $\rho_C = \frac{\delta_D - \alpha}{\delta_D - \alpha - \delta_C}, \rho_P = -\frac{1}{\beta}(\frac{rc}{G} - c)$. Consider the cycle marked in red. What happens in the system along this cycle is illustrated in Figure 4 (b). Clearly, the proportion of cooperators and punishers persistently oscillate. Moreover, the trends of change in the proportion of punishers (ρ_P) as well as in the proportion of cooperators (ρ_C) are generally consistent, though the change in ρ_C slightly lags behind.

Taking the system evolution shown in Figure 4 as an example, we explain the cyclic dynamics as follows. Initially, since there is a significant amount of punishment in the system, which reduces the payoff of defection, the proportion of cooperators increases. However, given a negative gain of punishment in a cooperative environment ($\delta_C < 0$), punishers receive less environmental payoff and hence the proportion of punishers decreases. This further leads to an increase in the proportion of defectors. As the gain of punishment in a non-cooperative environment is sufficiently high ($\delta_D > \alpha > 0$), punishment revives, which starts the next cycle.

Environmental Feedback of No Effect

In previous subsections, we generally considered that $\delta_C \neq 0, \delta_D \neq \alpha$ such that environmental feedback favors either punishment or non-punishment. For comparison, in this subsection, we analyze the case that $\delta_C = 0, \delta_D = \alpha$, suggesting that environmental feedback takes no effect. We establish the following result.

Theorem 7 *Given that* $\delta_C = 0, \delta_D - \alpha = 0$. *Over time,* ρ_P *remains unchanged, and* ρ_C *converges to* 1 *if* $\rho_P > -(\frac{rc}{G} - c)/\beta$ *but to* 0 *if* $\rho_P < -(\frac{rc}{G} - c)/\beta$.

Therefore, under this scenario, there will be no change in the proportion of punishers, however, the initial proportion of punishers determines whether complete cooperation or complete defection occurs.

Conclusions

In this paper, we propose a novel evolutionary game theoretic framework to address the puzzle of the evolution of punishment. Different from the traditional setting of one-shot PGG, we consider two populations. Agents in a game-playing population choose whether to cooperate with their opponents, while agents in a third-party population choose whether to punish the non-cooperators among the game players. We consider that the decision of punishment or not jointly depends on (i) the endogenous incentives for punishment in the environment (characterized by δ_C and δ_D), and (ii) on the amount of cooperation in the endogenous environment, whereby the two populations can influence each other.

We find that in absence of typical mechanisms previously studied, environmental feedback can, by itself, establish punishment and cooperation. Moreover, we find that with environmental feedback, complete punishment is not necessary to complete cooperation. On the other hand, complete cooperation will never go extinct if there still exists punishment in the system. We derive and analyze the conditions under which punishment and cooperation will co-present, co-dominate, or co-extinct. We notice that these conditions are closely related to the values of δ_C and $\delta_D - \alpha$, where δ_C and δ_D measure the gain of punishment in a cooperative and non-cooperative environment, respectively, and α is the cost of punishment. Interestingly, given certain conditions of these values, the system will eventually evolve into the states of bistability, oscillation, or monomorphic cooperating/defecting. Besides, through agent-based simulations, we find that the co-dominance, bi-stability, oscillation of punishment and cooperation can be reproduced in small, finite populations that evolve according to the Fermi process. The simulation results and details are summarized in our supplementary (the URL is provided in footnote 1).

To our knowledge, our paper is the first theoretical work that studies the effects of environmental feedback on punishment. Our most important takeaway message is that environmental feedback can promote the emergence of punishment. This provides a new pathway to address the evolutionary puzzle of punishment. From the perspective of implementing punishment in multi-agent systems, our results suggest some new insights. Considering the possibility that the environment can affect individual decision making sounds intuitive though, the effects of environmental feedback on punishment are non-trivial. Our analysis reveals that there is great richness in the resulting evolutionary dynamics, and particularly that persistent oscillation may potentially arise. This discovery of somewhat complex phenomena alerts the need for careful implementation of punishment, especially when taking into account that our environmental feedback mechanism is far from complex.

As future work, there are several interesting and fertile avenues. In this paper, we focus on altruistic punishment; however, previous studies have shown that the effectiveness of altruistic punishment is not only challenged by second-order free-riders (those who cooperate but do not punish) but also by antisocial punishment. The existence of antisocial punishment can destabilize altruistic punishment and even preclude the co-emergence of cooperation and punishment (Rand, Ohtsuki, and Nowak 2009; Rand et al. 2010). Therefore, allowing the possibility of antisocial punishment, whether and how environmental feedback still promotes the emergence of punishment requires further investigations. Moreover, as our framework can accommodate other social dilemmas, it would be interesting to see if our theoretical findings on public goods games can be carried over to those scenarios. Last but not least, testing our findings in human behavioural experiments would also be relevant and interesting.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work was supported in part by the National Science Fund for Distinguished Young Scholarship of China (No. 62025602), the National Natural Science Foundation of China (Nos. U22B2036, 11931015, U1803263), Fok Ying-Tong Education Foundationm China (No. 171105), Technological Innovation Team of Shaanxi Province (No. 2020TD-013), the Tencent Foundation and XPLORER PRIZE, and a JSPS Postdoctoral Fellowship Program for Foreign Researchers (grant no. P21374), and an accompanying Grant-in-Aid for Scientific Research from JSPS KAKENHI (grant no. JP 22F31374).

References

Archetti, M.; and Scheuring, I. 2012. Game theory of public goods in one-shot social dilemmas without assortment. *Journal of theoretical biology*, 299: 9–20.

Bench-Capon, T.; and Modgil, S. 2017. Norms and value based reasoning: justifying compliance and violation. *Artificial Intelligence and Law*, 25(1): 29–64.

Bou, E.; López-Sánchez, M.; and Rodríguez-Aguilar, J. A. 2006. Adaptation of autonomic electronic institutions through norms and institutional agents. In *International Workshop on Engineering Societies in the Agents World*, 300–319. Springer.

Campos, J.; López-Sánchez, M.; Rodríguez-Aguilar, J. A.; and Esteva, M. 2008. Formalising situatedness and adaptation in electronic institutions. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*, 126–139. Springer.

Dreber, A.; Rand, D. G.; Fudenberg, D.; and Nowak, M. A. 2008. Winners don't punish. *Nature*, 452(7185): 348–351.

Du, X.; Qi, Y.; Chen, B.; Shan, B.; and Liu, X. 2021. The integration of blockchain technology and smart grid: framework and application. *Mathematical Problems in Engineering*, 2021.

Fehr, E.; and Fischbacher, U. 2004. Third-party punishment and social norms. *Evolution and human behavior*, 25(2): 63–87.

Fehr, E.; and Gächter, S. 2002. Altruistic punishment in humans. *Nature*, 415(6868): 137–140.

Giardini, F.; Paolucci, M.; Villatoro, D.; and Conte, R. 2014. Punishment and gossip: sustaining cooperation in a public goods game. In *Advances in social simulation*, 107–118. Springer.

Han, T. A. 2016. Emergence of social punishment and cooperation through prior commitments. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2494–2500.

Han, T. A.; Pereira, L. M.; and Lenaerts, T. 2017. Evolution of commitment and level of participation in public goods games. *Autonomous Agents and Multi-Agent Systems*, 31(3): 561–583.

Helbing, D.; Szolnoki, A.; Perc, M.; and Szabó, G. 2010. Punish, but not too hard: how costly punishment spreads in the spatial public goods game. *New Journal of Physics*, 12(8): 083005.

Henrich, J.; McElreath, R.; Barr, A.; Ensminger, J.; Barrett, C.; Bolyanatz, A.; Cardenas, J. C.; Gurven, M.; Gwako, E.; Henrich, N.; et al. 2006. Costly punishment across human societies. *Science*, 312(5781): 1767–1770.

Herrmann, B.; Thoni, C.; and Gachter, S. 2008. Antisocial punishment across societies. *Science*, 319(5868): 1362–1367.

Hilbe, C.; Šimsa, Š.; Chatterjee, K.; and Nowak, M. A. 2018. Evolution of cooperation in stochastic games. *Nature*, 559(7713): 246–249.

Hofbauer, J.; Sigmund, K.; et al. 1998. *Evolutionary games and population dynamics*. Cambridge university press.

Jordan, J.; McAuliffe, K.; and Rand, D. 2016. The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, 19(4): 741–763.

Jordan, J. J.; Hoffman, M.; Bloom, P.; and Rand, D. G. 2016. Third-party punishment as a costly signal of trustworthiness. *Nature*, 530(7591): 473–476.

Kollock, P. 1998. Social dilemmas: The anatomy of cooperation. *Annual review of sociology*, 183–214.

Koppol, P.; Admoni, H.; and Simmons, R. G. 2021. Interaction Considerations in Learning from Humans. In *IJCAI*, 283–291.

Lenaerts, T.; et al. 2015. The efficient interaction of costly punishment and commitment. In *14th International Conference on Autonomous Agents and Multiagent Systems*.

Levy, P.; and Griffiths, N. 2021. Convention Emergence with Congested Resources. In *European Conference on Multi-Agent Systems*, 126–143. Springer.

Liu, Y.; Zhang, J.; An, B.; and Sen, S. 2016. A simulation framework for measuring robustness of incentive mechanisms and its implementation in reputation systems. *Autonomous Agents and Multi-Agent Systems*, 30(4): 581–600.

Mahmoud, S.; Griffiths, N.; Keppens, J.; Taweel, A.; Bench-Capon, T. J.; and Luck, M. 2015. Establishing norms with metanorms in distributed computational systems. *Artificial Intelligence and Law*, 23(4): 367–407.

Mahmoud, S.; Villatoro, D.; Keppens, J.; and Luck, M. 2012. Optimised reputation-based adaptive punishment for limited observability. In *2012 IEEE Sixth International Conference on Self-Adaptive and Self-Organizing Systems*, 129–138. IEEE.

Mathew, S.; and Boyd, R. 2011. Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences*, 108(28): 11375–11380.

Morris-Martin, A.; De Vos, M.; and Padget, J. 2019. Norm emergence in multiagent systems: a viewpoint paper. *Autonomous Agents and Multi-Agent Systems*, 33(6): 706–749.

Nowak, M. A. 2006. Five rules for the evolution of cooperation. *science*, 314(5805): 1560–1563.

Pereira, L. M.; Lenaerts, T.; Martinez-Vaquero, L. A.; and Han, T. A. 2017. Social Manifestation of Guilt Leads to Stable Cooperation in Multi-Agent Systems. In *AAMAS*, 1422–1430.

Pynadath, D. V.; and Marsella, S. C. 2005. PsychSim: Modeling theory of mind with decision-theoretic agents. In *IJCAI*, volume 5, 1181–1186.

Raihani, N. J.; and Bshary, R. 2015. The reputation of punishers. *Trends in ecology & evolution*, 30(2): 98–103.

Rand, D. G.; Armao IV, J. J.; Nakamaru, M.; and Ohtsuki, H. 2010. Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of theoretical biology*, 265(4): 624–632.

Rand, D. G.; Ohtsuki, H.; and Nowak, M. A. 2009. Direct reciprocity with costly punishment: Generous tit-for-tat prevails. *Journal of theoretical biology*, 256(1): 45–57.

Santos, F. P.; Mascarenhas, S. F.; Santos, F. C.; Correia, F.; Gomes, S.; and Paiva, A. 2019. Outcome-based Partner Selection in Collective Risk Dilemmas. In *AAMAS*, 1556–1564.

Savarimuthu, B. T. R.; Padget, J.; and Purvis, M. A. 2013. Social norm recommendation for virtual agent societies. In *International Conference on Principles and Practice of Multi-Agent Systems*, 308–323. Springer.

Semmann, D.; Krambeck, H.-J.; and Milinski, M. 2003. Volunteering leads to rock–paper–scissors dynamics in a public goods game. *Nature*, 425(6956): 390–393.

Sigdel, R.; Anand, M.; and Bauch, C. T. 2019. Convergence of socio-ecological dynamics in disparate ecological systems under strong coupling to human social systems. *Theoretical Ecology*, 12(3): 285–296.

Tilman, A. R.; Plotkin, J. B.; and Akçay, E. 2020. Evolutionary games with environmental feedbacks. *Nature communications*, 11(1): 1–11.

Villatoro, D.; Andrighetto, G.; Sabater-Mir, J.; and Conte, R. 2011. Dynamic sanctioning for robust and cost-efficient norm compliance. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Wanders, F.; Homan, A. C.; van Vianen, A. E.; Rahal, R.-M.; and Van Kleef, G. A. 2021. How norm violators rise and fall in the eyes of others: The role of sanctions. *PLoS one*, 16(7): e0254574.

Weitz, J. S.; Eksin, C.; Paarporn, K.; Brown, S. P.; and Ratcliff, W. C. 2016. An oscillating tragedy of the commons in replicator dynamics with game-environment feedback. *Proceedings of the National Academy of Sciences*, 113(47): E7518–E7525.

Zimbardo, P. G. 1969. The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In *Nebraska symposium on motivation*. University of Nebraska press.