Supplementary information: Emergence of Punishment in Social Dilemma with Environmental Feedback

1. APPENDIX A: STABILITY ANALYSIS OF REPLICATOR DYNAMICS

In this appendix, we analyze the stability conditions for different equilibrium states of the following system of differential equations:

$$\begin{aligned}\dot{\rho}_C =& \rho_C(1-\rho_C)(\frac{rc}{G}-c+\beta\rho_P)\\ \dot{\rho}_P =& \rho_P(1-\rho_P)\left[(s_{PC}-s_{NC})\rho_C+(s_{PD}-s_{ND}-\alpha)(1-\rho_C)\right].\end{aligned} \tag{S1}$$

These two equations describe how the frequencies of playing cooperation (ρ_C) and playing punishment (ρ_P) evolve over time. To see the incentive of punishment in the environment, we define

$$\delta_C \coloneqq s_{PC}-s_{NC} \tag{S2}$$

$$\delta_D \coloneqq s_{PD}-s_{ND} \tag{S3}$$

Therefore, δ_C measures the difference in environmental payoff between punishment and non-punishment when the game player is a cooperator, and δ_D measures the difference in environmental payoff between punishment and non-punishment when the game player is a defector. Throughout our analysis, we assume $\frac{r}{G}<1$ and $\frac{rc}{G}-c+\beta>0$. The former is the standard in the theoretical model for PGG, and the latter aims to ensure the effectiveness of punishment.

Lemma 1 *The system (S1) has five fixed points,*

- $\rho_C=0,\rho_P=0$, *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.*

Proof *Let $\dot{\rho}_C=0$ and $\dot{\rho}_P=0$, we obtain the fixed points of system (S1).*

Lemma 2 *The Jacobian matrix of a fixed point $\rho_C=\rho_C^*$, $\rho_P=\rho_P^*$ is*

$$J_{\rho_C^*,\rho_P^*}=\begin{bmatrix}(1-2\rho_C^*)(\frac{rc}{G}-c+\beta\rho_P^*) & \rho_C^*(1-\rho_C^*)\beta\rho_P^*\\ \rho_P^*(1-\rho_P^*)[\delta_C-(\delta_D-\alpha)] & (1-2\rho_P^*)[\delta_C\rho_C^*+(\delta_D-\alpha)(1-\rho_C^*)]\end{bmatrix}. \tag{S4}$$

The eigenvalues λ_1, λ_2 of the Jacobian matrix evaluated at each fixed point are

- $\rho_C=0,\rho_P=0$: $\lambda_1=\delta_D-\alpha$, $\lambda_2=\frac{rc}{G}-c$,
- $\rho_C=0$, $\rho_P=1$: $\lambda_1=-(\delta_D-\alpha)$, $\lambda_2=\frac{rc}{G}-c+\beta$,
- $\rho_C=1$, $\rho_P=0$: $\lambda_1=\delta_C$, $\lambda_2=-(\frac{rc}{G}-c)$,
- $\rho_C=1$, $\rho_P=1$: $\lambda_1=-\delta_C$, $\lambda_2=-(\frac{rc}{G}-c+\beta)$,
- $\rho_C=\frac{\delta_D-\alpha}{\delta_D-\alpha-\delta_C}$, $\rho_P=-\frac{1}{\beta}(\frac{rc}{G}-c)$: $\lambda_1=-\lambda$, $\lambda_2=\lambda$, where

$$\lambda^2=-\frac{(\frac{rc}{G}-c)(\frac{rc}{G}-c+\beta)(\delta_D-\alpha)\delta_C}{\beta(\delta_D-\alpha-\delta_C)}. \tag{S5}$$

We see that the interior fixed point exists, $\rho_C=\frac{\delta_D-\alpha}{\delta_D-\alpha-\delta_C}\in[0,1]$, only when δ_C and $\delta_D-\alpha$ have different signs.

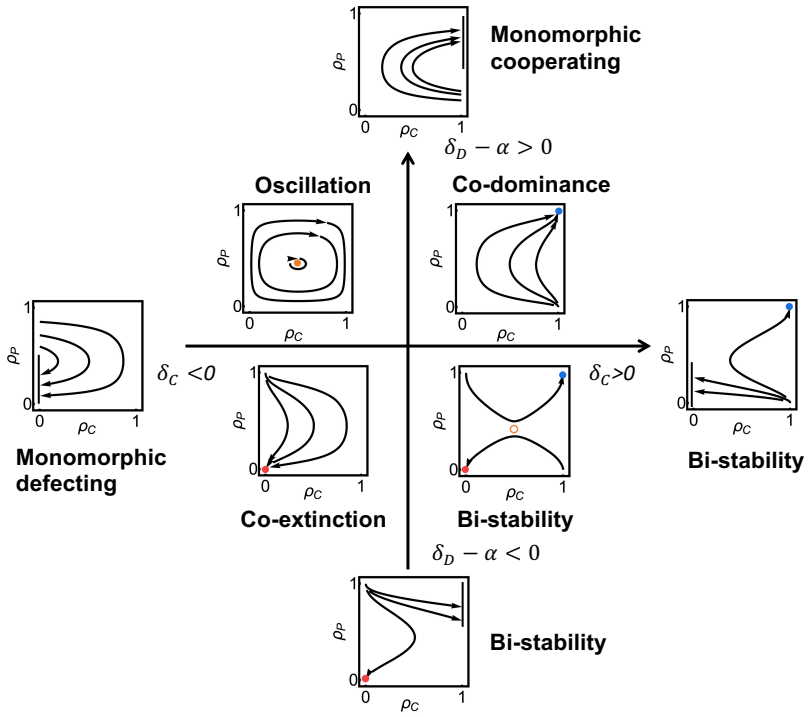

Fig. S1. A schematic representation of the co-evolution of punishment and cooperation with environmental feedback.

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

Proof *When $\delta_C > 0, \delta_D > 0$, according to Lemma 2, the eigenvalues at $\rho_C = 0, \rho_P = 0$ are $\lambda_1 = \frac{rc}{G} - c < 0$, $\lambda_2 = \delta_D - \alpha > 0$. Thus $\rho_C = 0, \rho_P = 0$ is unstable. The eigenvalues at $\rho_C = 1, \rho_P = 0$ are $\lambda_1 = -(\frac{rc}{G} - c) > 0$, $\lambda_2 = \delta_C > 0$. Thus $\rho_C = 1, \rho_P = 0$ is unstable. The eigenvalues at $\rho_C = 0, \rho_P = 1$ are $\lambda_1 = \frac{rc}{G} - c + \beta > 0$, $\lambda_2 = -(\delta_D - \alpha) < 0$. Thus $\rho_C = 0, \rho_P = 1$ is unstable. The eigenvalues at $\rho_C = 1, \rho_D = 1$ are $\lambda_1 = -(\frac{rc}{G} - c + \beta) < 0$, $\lambda_2 = -\delta_C < 0$. Thus $\rho_C = 1, \rho_P = 1$ is stable. The fixed point $\frac{\delta_D - \alpha}{\delta_D - \alpha - \delta_C}$, $\rho_P = -\frac{1}{\beta}(\frac{rc}{G} - c)$ is do not exist as $|\frac{\delta_D - \alpha}{\delta_D - \alpha - \delta_C}| > 1$.*

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 $\delta_C = 0$, $\delta_D > \alpha > 0$.*

Proof *When $\delta_C = 0$, $\delta_D > \alpha > 0$, according to Lemma 2, the eigenvalues at $\rho_C = 0$, $\rho_P = 0$ are $\lambda_1 = \delta_D - \alpha > 0$, $\lambda_2 = \frac{rc}{G} - c < 0$. Thus $\rho_C = 0$, $\rho_P = 0$ is unstable. The the eigenvalues at $\rho_C = 0$, $\rho_P = 1$ are $\lambda_1 = -(\delta_D - \alpha) < 0$, $\lambda_2 = \frac{rc}{G} - c + \beta > 0$. Thus $\rho_C = 0$, $\rho_P = 1$ is unstable. The eigenvalues at $\rho_C = 1$, $\rho_P = 0$ are $\lambda_1 = \delta_C = 0$, $\lambda_2 = -(\frac{rc}{G} - c) > 0$. Thus $\rho_C = 1$, $\rho_P = 0$ is unstable. The interior fixed point do not exist since $\frac{\delta_D - \alpha}{\delta_D - \alpha - \delta_C} = 1$. There are a continuum of fixed points $\rho_C = 1$, $\rho_D = x$, $x \in [0, 1]$. The eigenvalues are $\lambda_1 = 0$, $\lambda_2 = -(\frac{rc}{G} - c + \beta x)$. The fixed point can be stable if $\lambda_2 < 0$, namely, $x < -\frac{1}{\beta}(\frac{rc}{G} - c)$, otherwise unstable since $\lambda_2 \geq 0$.*

In the condition $\lambda_1 = 0$, $\lambda_2 < 0$, we analyse the stability of the fixed point by centre manifold theorem [1]. We construct a matrix T whose column vectors are the eigenvectors of Jacobi matrix J, given as:

$$T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{S6}$$

We obtain

$$T^{-1}JT = \begin{bmatrix} 0 & 0 \\ 0 & \frac{rc}{G} - c + \beta x \end{bmatrix}. \tag{S7}$$

Using variable substitution, we obtain:

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = T^{-1} \begin{bmatrix} \rho_C \\ \rho_P \end{bmatrix} = \begin{bmatrix} \rho_P \\ \rho_C \end{bmatrix}. \tag{S8}$$

Then the system (S1) becomes:

$$\begin{aligned} \dot{u'} &= u'(1-u')(\delta_D - \alpha)(1-v') \\ \dot{v'} &= v'(1-v')(\frac{rc}{G} - c + \beta u'). \end{aligned} \tag{S9}$$

Let $u' = u = x$ and $v' = v+1$, the system becomes:

$$\begin{aligned} \dot{u} &= (u+x)(1-u-x)(\delta_D - \alpha)(1-v-1) \\ \dot{v} &= -v(v+1)(\frac{rc}{G} - c + \beta(u+x)). \end{aligned} \tag{S10}$$

Using the centre manifold theorem, we have that $u = h(v)$ is the center manifold for the system. The dynamics on the centre manifold is given by:

$$\dot{v} = -v(v+1)(\frac{rc}{G} - c + \beta x + \beta h(v)). \tag{S11}$$

We start to try $h(v) = O(v^2)$, and system (S11) becomes:

$$\dot{v} = -v(v+1)(\frac{rc}{G} - c + \beta x) + O(v^3). \tag{S12}$$

By defining $g(v) = -v(v+1)(\frac{rc}{G} - c + \beta x)$, we obtain $g'(v) = -(1-2v)(\frac{rc}{G} - c + \beta x)$. As $g(0)' < 0$, the system is stable. Accordingly, $\rho_C = 1$, $\rho_P = x$, $x \in (-\frac{1}{\beta}(\frac{rc}{G} - c), 1]$ is stable.

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

Proof *Theorem 1 indicates that $\rho_C = 1$, $\rho_P = 1$ is stable when $\delta_C > 0, \delta_D > \alpha > 0$. Theorem 2 indicates that $\rho_C = 1$, $\rho_P = x$ is stable when $\delta_C = 0, \delta_D > \alpha > 0$. Thus, when $\delta_C \geq 0$, $\delta_D > \alpha > 0$, $\rho_C = 1$ will always be achieved.*

Theorem 3 *The equilibrium states $\rho_C = 1, \rho_P = 0$, and $\rho_C = 0, \rho_P = 1$ are always unstable.*

Proof *The eigenvalues at $\rho_C = 1, \rho_P = 0$ are $\lambda_1 = \delta_C$, $\lambda_2 = -(\frac{rc}{G} - c) > 0$, thus $\rho_C = 1, \rho_P = 0$ is always unstable. The eigenvalues at $\rho_C = 0, \rho_P = 1$ are $\lambda_1 = -(\delta_D - \alpha)$, $\lambda_2 = \frac{rc}{G} - c + \beta > 0$, thus $\rho_C = 0, \rho_P = 1$ is always unstable.*

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 $\delta_C < 0$, $\delta_D < \alpha$.*

Proof *According to theorem 1, only $\rho_C = 1$, $\rho_P = 1$ is stable when $\delta_C > 0, \delta_D > \alpha > 0$. According to theorem 2, only $\rho_C = 1$, $\rho_P = x$, $x \in (-\frac{1}{\beta}(\frac{rc}{G} - c), 1]$ is stable when $\delta_C = 0$, $\delta_D > \alpha > 0$. According to theorem 5, $\rho_C = 1$, $\rho_P = 1$ is stable when $\delta_C > 0, \delta_D < \alpha, \alpha > 0$. According to theorem 8, $\rho_C = 1$, $\rho_P = x$, $x \in (-\frac{1}{\beta}(\frac{rc}{G} - c), 1]$ is stable when $\delta_C = 0, \delta_D < \alpha, \alpha > 0$, and $\rho_C = 1$, $\rho_P = 1$ is stable when $\delta_C > 0, \delta_D = \alpha, \alpha > 0$. Thus, when $\delta_C \geq 0$ (except $\delta_C = 0$, $\delta_D = \alpha$), it is possible for a system to rest in the co-present of punishment and cooperation.*

According to theorem 1, $\rho_C = 1$, $\rho_P = 1$ is stable when $\delta_C > 0, \delta_D > \alpha > 0$. According to theorem 2, only $\rho_C = 1$, $\rho_P = x$, $x \in (-\frac{1}{\beta}(\frac{rc}{G} - c), 1]$ is stable when $\delta_C = 0$, $\delta_D > \alpha > 0$. According to theorem 6, cycle dynamics exist when $\delta_C < 0$, $\delta_D > \alpha > 0$. Thus, when $\delta_D > \alpha > 0$, it is possible for a system to rest in the co-present of punishment and cooperation.

When $\delta_C < 0$, $\delta_D < \alpha$, the eigenvalues at fixed point $\rho_C = 0, \rho_P = 0$ are $\lambda_1 = \delta_D - \alpha < 0$, $\lambda_2 = \frac{rc}{G} - c < 0$. Thus, $\rho_C = 0, \rho_P = 0$ is stable. The eigenvalues at fixed point $\rho_C = 0, \rho_P = 1$ are $\lambda_1 = -(\delta_D - \alpha) > 0$, $\lambda_2 = \frac{rc}{G} - c + \beta > 0$. Thus, $\rho_C = 0, \rho_P = 1$ is unstable. The eigenvalues at fixed point $\rho_C = 1, \rho_P = 0$ are $\lambda_1 = \delta_C < 0$, $\lambda_2 = -(\frac{rc}{G} - c) > 0$. Thus, $\rho_C = 1, \rho_P = 0$ is unstable. The eigenvalues at fixed point $\rho_C = 1, \rho_P = 1$ are $\lambda_1 = -\delta_C > 0$, $\lambda_2 = -(\frac{rc}{G} - c + \beta) < 0$. Thus, $\rho_C = 1, \rho_P = 1$ is unstable. The interior fixed point do not exits as δ_C and $\delta_D - \alpha$ have the same sign. In summary, $\rho_C = 0, \rho_P = 0$ is the unique asymptotically stable state.

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 $\rho_C = \frac{\delta_D - \alpha}{\delta_D - \alpha - \delta_C}$, $\rho_P = -\frac{1}{\beta}(\frac{rc}{G} - c)$ if $\delta_C > 0, \delta_D < \alpha, \alpha > 0$.*

Proof *When $\delta_C > 0, \delta_D < \alpha, \alpha > 0$, the eigenvalues at fixed point $\rho_C = 0$, $\rho_P = 0$ are $\lambda_1 = \delta_D - \alpha < 0$, $\lambda_2 = \frac{rc}{G} - c < 0$. Thus, $\rho_C = 0$, $\rho_P = 0$ is stable. The eigenvalues at fixed point $\rho_C = 0$, $\rho_P = 1$ are $\lambda_1 = -(\delta_D - \alpha) > 0$, $\lambda_2 = \frac{rc}{G} - c + \beta > 0$. Thus, $\rho_C = 0$, $\rho_P = 1$ is unstable. The eigenvalues at fixed point $\rho_C = 1$, $\rho_P = 0$ are $\lambda_1 = \delta_C < 0$, $\lambda_2 = -(\frac{rc}{G} - c) > 0$. Thus, $\rho_C = 1$, $\rho_P = 0$ is unstable. The eigenvalues at fixed point $\rho_C = 1$, $\rho_P = 1$ are $\lambda_1 = -\delta_C < 0$, $\lambda_2 = -(\frac{rc}{G} - c + \beta) < 0$. Thus, $\rho_C = 1$, $\rho_P = 1$ is stable. The eigenvalues at fixed point $\rho_C = \frac{\delta_D - \alpha}{\delta_D - \alpha - \delta_C}$, $\rho_P = -\frac{1}{\beta}(\frac{rc}{c} - c)$ have the same absolute values but different signs. Thus, $\rho_C = \frac{\delta_D - \alpha}{\delta_D - \alpha - \delta_C}$, $\rho_P = -\frac{1}{\beta}(\frac{rc}{c} - c)$ is the saddle point.*

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

Proof *When $\delta_C < 0, \delta_D > \alpha, \alpha > 0$, the eigenvalues at fixed point $\rho_C = 0$, $\rho_P = 0$ are $\lambda_1 = \delta_D - \alpha > 0$, $\lambda_2 = \frac{rc}{G} - c < 0$. Thus, $\rho_C = 0$, $\rho_P = 0$ is unstable. The eigenvalues at fixed point $\rho_C = 0$, $\rho_P = 1$ are $\lambda_1 = -(\delta_D - \alpha) < 0$, $\lambda_2 = \frac{rc}{G} - c + \beta > 0$. Thus, $\rho_C = 0$, $\rho_P = 1$ is unstable. The eigenvalues at fixed point $\rho_C = 1$, $\rho_P = 0$ are $\lambda_1 = \delta_C < 0$, $\lambda_2 = -(\frac{rc}{G} - c) > 0$. Thus, $\rho_C = 1$, $\rho_P = 0$ is unstable. The eigenvalues at fixed point $\rho_C = 1$, $\rho_P = 1$ are $\lambda_1 = -\delta_C > 0$, $\lambda_2 = -(\frac{rc}{G} - c + \beta) < 0$. Thus, $\rho_C = 1$, $\rho_P = 1$ is unstable. The eigenvalues at fixed point $\rho_C = \frac{\delta_D - \alpha}{\delta_D - \alpha - \delta_C}$, $\rho_P = -\frac{1}{\beta}(\frac{rc}{c} - c)$ are pure imaginary numbers. They have the same absolute values but different signs. Thus, $\rho_C = \frac{\delta_D - \alpha}{\delta_D - \alpha - \delta_C}$, $\rho_P = -\frac{1}{\beta}(\frac{rc}{c} - c)$ is neutrally stable.*

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 $\rho_P < -(\frac{rc}{G} - c)/\beta$.*

Proof *When $\delta_C = 0, \delta_D - \alpha = 0$, the system (S1) becomes:*

$$\begin{aligned} \dot{\rho}_C =& \rho_C(1 - \rho_C)(\frac{rc}{G} - c + \beta\rho_P) \\ \dot{\rho}_P =& 0. \end{aligned} \tag{S13}$$

ρ_P is constant after initialization. The evolution of ρ_C is monotonic and depends on $\frac{rc}{G} - c + \beta\rho_P$, which measures the difference of expected payoffs between cooperators and defectors. If $\frac{rc}{G} - c + \beta\rho_P > 0$, $\dot{\rho_C} > 0$ and ρ_C increases until equals 1. While if $\frac{rc}{G} - c + \beta\rho_P < 0$, $\dot{\rho_C} < 0$ and ρ_C decreases until equals 0.

Theorem 8 *There exists the bi-stability of $\rho_C = 1$, $\rho_P = 1$ and $\rho_C = 0$, $\rho_P = x$, $x \in [0, -\frac{1}{\beta}(\frac{rc}{G} - c))$ when $\delta_C > 0$, $\delta_D = \alpha$, and $\rho_C = 0$, $\rho_P = 0$ and $\rho_C = 1$, $\rho_P = x$, $x \in (-\frac{1}{\beta}(\frac{rc}{G} - c), 1]$ when $\delta_C = 0$, $\delta_D < \alpha$.*

Proof *When $\delta_C > 0, \delta_D = \alpha, \alpha > 0$, the eigenvalues at fixed point $\rho_C = 0$, $\rho_P = 1$ are $\lambda_1 = -(\delta_D - \alpha) = 0$, $\lambda_2 = \frac{rc}{G} - c + \beta > 0$. Thus, $\rho_C = 0$, $\rho_P = 1$ is unstable. The eigenvalues at fixed point $\rho_C = 1$, $\rho_P = 0$ are $\lambda_1 = \delta_C > 0$, $\lambda_2 = -(\frac{rc}{G} - c) > 0$. Thus, $\rho_C = 1$, $\rho_P = 0$ is unstable. The eigenvalues at fixed point*

$\rho_C = 1$, $\rho_P = 1$ *are* $\lambda_1 = -\delta_C < 0$, $\lambda_2 = -(\frac{rc}{G} - c + \beta) < 0$. *Thus,* $\rho_C = 1$, $\rho_P = 1$ *is stable. The interior fixed point do not exists as* $\frac{1}{-\delta_C} < 0$. *There a continuum of fixed points* $\rho_C = 0$, $\rho_P = x$, $x \in [0,1]$. *The eigenvalues at* $\rho_C = 0$, $\rho_P = x$ *are* $\lambda_1 = 0$, $\lambda_2 = \frac{rc}{G} - c + \beta x$. *The continuum of fixed points* $x < -\frac{1}{\beta}(\frac{rc}{G} - c)$ *can be stable since* $\lambda_2 = \frac{rc}{G} - c + \beta x \leq 0$, *others are unstable. In this condition, we analyse the stability by using the centre manifold theorem. We construct a matrix* T *whose column vectors are the eigenvectors of Jacobian matrix* J, *given as:*

$$T = \begin{bmatrix} 0 & \frac{\frac{rc}{G} - c + \beta x}{x(1-x)\delta_C} \\ 1 & 1 \end{bmatrix}. \tag{S14}$$

We obtain

$$T^{-1}JT = \begin{bmatrix} 0 & 0 \\ 0 & \frac{rc}{G} - c + \beta x \end{bmatrix}. \tag{S15}$$

Using variable substitution, we obtain:

$$\begin{bmatrix} u' \\ v \end{bmatrix} = T^{-1} \begin{bmatrix} \rho_C \\ \rho_P \end{bmatrix} = \begin{bmatrix} \rho_P \\ \rho_C \end{bmatrix}. \tag{S16}$$

Then the system (S1) becomes:

$$\begin{aligned} \dot{u'} &= u'(1-u')\delta_C v \\ \dot{v} &= v(1-v)(\frac{rc}{G} - c + \beta u'). \end{aligned} \tag{S17}$$

Let $u' = u + x$, *and the system becomes:*

$$\begin{aligned} \dot{u} &= (u+x)(1-u-x)\delta_C v \\ \dot{v} &= v(1-v)(\frac{rc}{G} - c + \beta(u+x)). \end{aligned} \tag{S18}$$

Using the centre manifold theorem, we have that $u = h(v)$ *is the center manifold for the system. The dynamics on the centre manifold is given by:*

$$\dot{v} = v(1-v)(\frac{rc}{G} - c + \beta x + \beta h(v)). \tag{S19}$$

We start to try $h(v) = O(v^2)$, *and system (S19) becomes:*

$$\dot{v} = v(1-v)(\frac{rc}{G} - c + \beta x) + O(v^3). \tag{S20}$$

By defining $g(v) = v(1-v)(\frac{rc}{G} - c + \beta x)$, *we obtain* $g'(v) = (1-2v)(\frac{rc}{G} - c + \beta x)$. *As* $g(0)' < 0$, *the system is stable. Accordingly,* $\rho_C = 0$, $\rho_P = x$, $x \in [0, -\frac{1}{\beta}(\frac{rc}{G} - c))$ *is stable.*

The proof of the bi-stability of $\rho_C = 0$, $\rho_P = 0$ *and* $\rho_C = 1$, $\rho_P = x$, $x \in (-\frac{1}{\beta}(\frac{rc}{G} - c), 1]$ *can be obtained by the same theorem.*

Theorem 9 *There only exists a continuum of stable equilibrium state* $\rho_C = 0$, $\rho_P = x$ *with* $x \in [0, -\frac{1}{\beta}(\frac{rc}{G} - c))$ *if* $\delta_C < 0$, $\delta_D = \alpha$, $\alpha > 0$.

Proof *When* $\delta_C < 0, \delta_D = \alpha, \alpha > 0$, *the eigenvalues at fixed point* $\rho_C = 1$, $\rho_P = 0$ *are* $\lambda_1 = \delta_C < 0$, $\lambda_2 = -(\frac{rc}{G} - c) > 0$. *Thus,* $\rho_C = 1$, $\rho_P = 0$ *is unstable. The eigenvalues at fixed point* $\rho_C = 0$, $\rho_P = 1$ *are* $\lambda_1 = -(\delta_D - \alpha) = 0$, $\lambda_2 = \frac{rc}{G} - c + \beta > 0$. *Thus,* $\rho_C = 0$, $\rho_P = 1$ *is unstable. The eigenvalues at fixed point* $\rho_C = 1$, $\rho_P = 1$ *are* $\lambda_1 = -\delta_C > 0$, $\lambda_2 = -(\frac{rc}{G} - c + \beta) < 0$. *Thus,* $\rho_C = 1$, $\rho_P = 1$ *is unstable. The interior fixed point do not exists and there a continuum of fixed points* $\rho_C = 0$, $\rho_P = x$, $x \in [0,1]$. *The eigenvalues at* $\rho_C = 0$, $\rho_P = x$ *are* $\lambda_1 = 0$, $\lambda_2 = \frac{rc}{G} - c + \beta x$. *The continuum of fixed points* $x < -\frac{1}{\beta}(\frac{rc}{G} - c)$ *can be stable since* $\lambda_2 = \frac{rc}{G} - c + \beta x \leq 0$, *others are unstable. In this condition, we analyse the stability by using the centre manifold theorem. We construct a matrix* T *whose column vectors are the eigenvectors of Jacobian matrix* J, *given as:*

$$T = \begin{bmatrix} 0 & \frac{\frac{rc}{G} - c + \beta x}{x(1-x)\delta_C} \\ 1 & 1 \end{bmatrix}. \tag{S21}$$

We obtain

$$T^{-1}JT = \begin{bmatrix} 0 & 0 \\ 0 & \frac{rc}{G} - c + \beta x \end{bmatrix}. \tag{S22}$$

Using variable substitution, we obtain:

$$\begin{bmatrix} u' \\ v \end{bmatrix} = T^{-1} \begin{bmatrix} \rho_C \\ \rho_P \end{bmatrix} = \begin{bmatrix} \rho_P \\ \rho_C \end{bmatrix}. \tag{S23}$$

Then the system (S1) becomes:

$$\begin{aligned} \dot{u'} &= u'(1-u')\delta_C v \\ \dot{v} &= v(1-v)(\frac{rc}{G} - c + \beta u'). \end{aligned} \tag{S24}$$

Let $u' = u + x$, the system becomes:

$$\begin{aligned} \dot{u} &= (u+x)(1-u-x)\delta_C v \\ \dot{v} &= v(1-v)(\frac{rc}{G} - c + \beta(u+x)). \end{aligned} \tag{S25}$$

Using the centre manifold theorem, we have that $u = h(v)$ is the center manifold for the system. The dynamics on the centre manifold is given by:

$$\dot{v} = v(1-v)(\frac{rc}{G} - c + \beta x + \beta h(v)). \tag{S26}$$

We start to try $h(v) = O(v^2)$, and system (S26) becomes:

$$\dot{v} = v(1-v)(\frac{rc}{G} - c + \beta x) + O(v^3). \tag{S27}$$

By defining $g(v) = v(1-v)(\frac{rc}{G} - c + \beta x)$, we obtain $g'(v) = (1-2v)(\frac{rc}{G} - c + \beta x)$. As $g(0)' < 0$, the system is stable. Accordingly, $\rho_C = 0$, $\rho_P = x$, $x \in [0, -\frac{1}{\beta}(\frac{rc}{G} - c))$ is stable.

2. APPENDIX B: AGENT-BASED SIMULATIONS ON FINITE POPULATIONS

We consider a game-playing population and a third-party population, each of which consists of $n = 100$ agents. The contribution c to the public pool of each game player is randomly drew from the normal distribution $c \sim N(1, 0.5)$ and $c \sim N(1, 0.8)$. Agents update their strategies according to the Fermi process (a standard imitative process for simulating replicator dynamics in finite populations) such that at every time step, an agent explores a strategy with probability 0.01 or otherwise imitates the strategy of another randomly selected agent with the probability following the Fermi function. Under each condition of δ_C and $\delta_D - \alpha$ (the condition predicted by our theory), we consider several different initialization of agent strategies. For each setting, we conduct 50 independent simulation runs and the presented results are the averaged results.

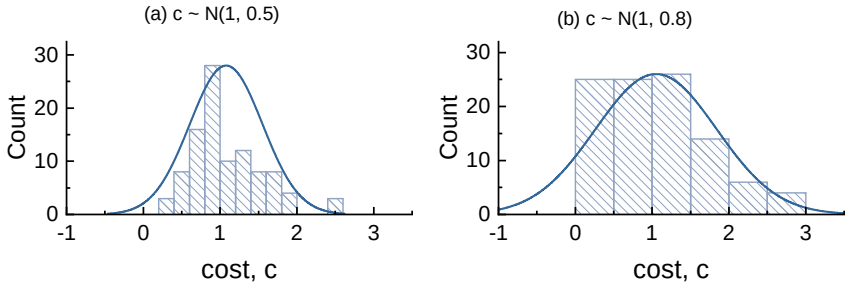

Fig. S2. Distributions of the cost, c. The curve represents the probability distribution, and the histogram represents the frequency distribution. To ensure cooperators contribute positively to the public pool, we limit the $c < 0$ to 0 when generating random values.

REFERENCES

1. J. Carr, *Applications of centre manifold theory*, vol. 35 (Springer Science & Business Media, 2012).

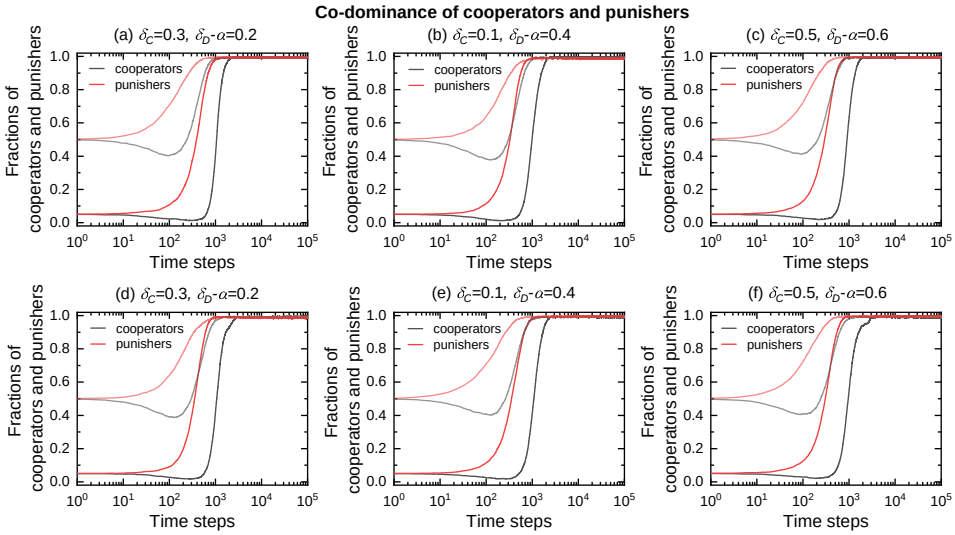

Fig. S3. Co-dominance of cooperators and punishers in small, finite, heterogeneous populations, given the conditions of Theorem 1 are met. $c \sim N(1,0.5)$ in the top row, $c \sim N(1,0.8)$ in the bottom row. Light and dark curves represent different initial states, respectively.

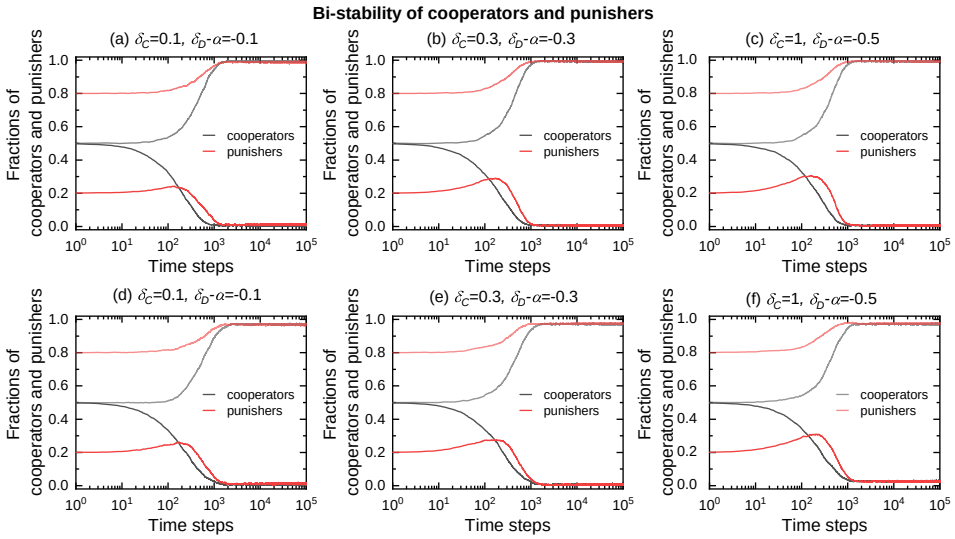

Fig. S4. Bi-stability of cooperators and punishers in small, finite, heterogeneous populations, given the conditions of Theorem 5 are met. $c \sim N(1,0.5)$ in the top row, $c \sim N(1,0.8)$ in the bottom row. Light and dark curves represent different initial states, respectively.

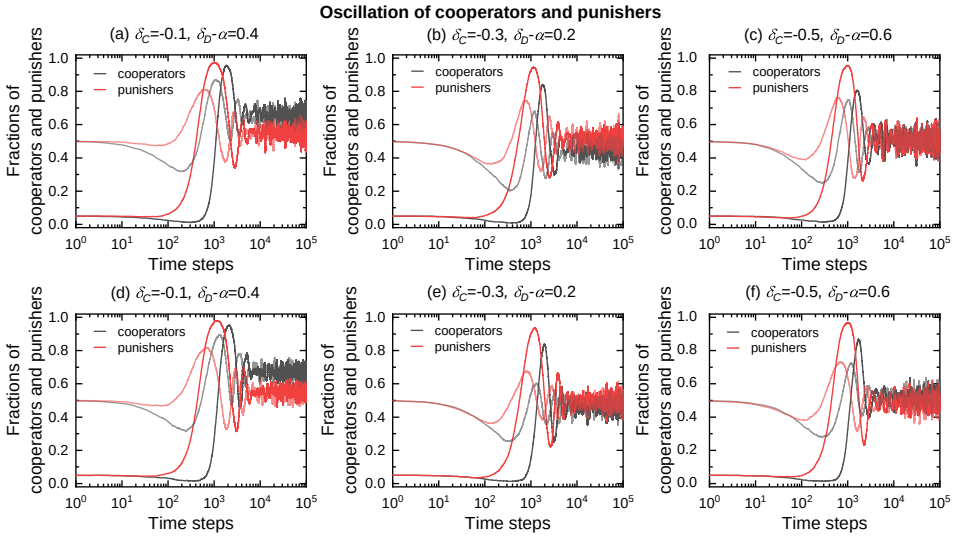

Fig. S5. Oscillation of cooperators and punishers in small, finite, heterogeneous populations, given the conditions of Theorem 6 are met. $c \sim N(1, 0.5)$ in the top row, $c \sim N(1, 0.8)$ in the bottom row. Light and dark curves represent different initial states, respectively.