INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Song Z, Shen C, Capraro V, Han TA. 2026 Non-participant externalities reshape the evolution of altruistic punishment. *J. R. Soc. Interface* **23**: 20250820.
https://doi.org/10.1098/rsif.2025.0820

Received: 7 August 2025
Accepted: 19 December 2025

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
evolution

Keywords:
evolutionary game theory, public goods game, voluntary participation, cooperation, population dynamics, altruistic punishment

Author for correspondence:
Zhao Song
e-mails: songzhao.songz@gmail.com; Z.Song@tees.ac.uk

Non-participant externalities reshape the evolution of altruistic punishment

Zhao Song[1], Chen Shen[2], Valerio Capraro[3] and The Anh Han[1]

[1]Teesside University, Middlesbrough, England, UK
[2]Kyushu University, Fukuoka, Japan
[3]Department of Psychology, University of Milan-Bicocca, Milan, Lombardy, Italy

ZS, 0000-0002-0897-6724; CS, 0000-0002-2725-9763; VC, 0000-0002-0579-0166; TAH, 0000-0002-3095-7714

Understanding the evolutionary origins of altruistic punishment, a critical mechanism sustaining cooperation, remains a central challenge in behavioural science. Voluntary participation is considered a powerful approach that enables its emergence, but its explanatory power typically rests on the common assumption that non-participants have no impact on the public good. Yet, given the decentralized nature of voluntary participation, opting out does not necessarily preclude individuals from influencing the public good. Here, we revisit the role of voluntary participation by allowing non-participants to exert either positive or negative impacts on the public good. Using evolutionary analysis in a well-mixed finite population, we find that positive externalities from non-participants lower the synergy threshold required for altruistic punishment to dominate. In contrast, negative externalities raise this threshold, making altruistic punishment harder to sustain. Notably, when non-participants have positive impacts, altruistic punishment thrives only if non-participation is incentivized, whereas under negative impacts, it can persist even when non-participation is discouraged. Our findings reveal that efforts to promote altruistic punishment must account for the active role of non-participants, whose influence can make or break collective outcomes.

1. Introduction

The evolution of altruistic punishment, a key mechanism for sustaining cooperation, presents a long-standing theoretical puzzle [1–3]. Experimental evidence confirms that individuals will incur personal costs to punish non-cooperators, implementing sanctions through both decentralized peer punishment, where individuals punish defectors directly, and centralized pool punishment, where contributions are pooled to fund institutional sanctioning [4–7]. However, peer punishment is particularly vulnerable to 'second-order free-riders', cooperators who benefit from sanctions without paying the costs, a vulnerability that is largely mitigated by the institutionalized cost-sharing of pool punishment [8–11]. In addition, peer punishment is exposed to the risky retaliation from those they sanction [12–14]. To solve this paradox for decentralized systems, researchers have proposed mechanisms including group selection [15,16] and prior commitment [17,18].

Voluntary participation offers a powerful bottom-up approach to this problem [19,20]. To understand its importance, one must first consider the dynamics without it. In populations of only cooperators, defectors and punishers, a fragile rock-paper-scissors cycle can emerge: punishers suppress defectors, but once defectors become rare, non-punishing cooperators outperform punishers by avoiding punishment costs. This allows defectors to reinvade, completing a fragile cycle requiring careful fine-tuning of selection strength and payoffs; otherwise, it collapses into full defection, permanently

THE ROYAL SOCIETY PUBLISHING

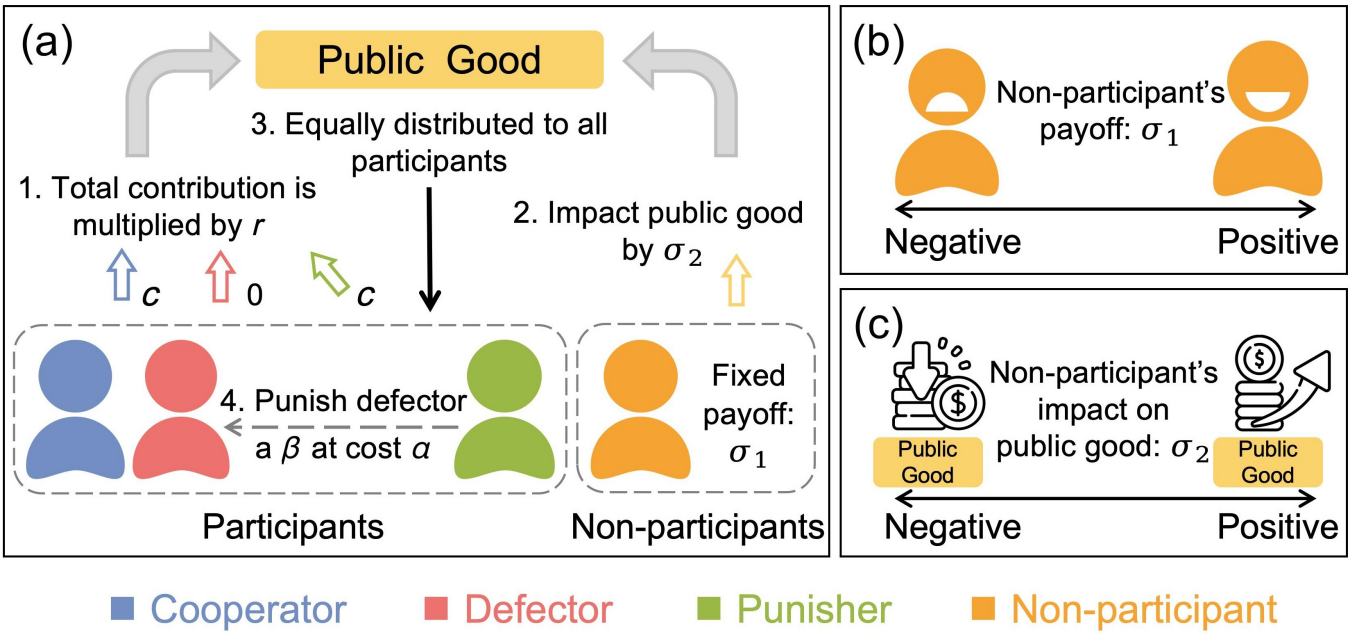

Figure 1. PGG with voluntary participation. (a) Each player has four options: cooperation, defection, punishment and non-participation. Cooperators and punishers contribute an endowment c to the public pool, while defectors contribute 0. Non-participants opt out of the game with a payoff σ_1. The total contribution from all group members will be multiplied by the synergy factor r as the public good. Each non-participant has an impact σ_2 on the public good. Then the public good is equally distributed among all participants, regardless of their initial contribution. (b) Non-participant's payoff σ_1 can be negative, positive or zero. (c) Non-participant's impact on the public good σ_2 can be negative, positive or zero.

eliminating punishment. Voluntary participation provides an 'escape hatch' by introducing a non-participation strategy, where players can opt out for a small, fixed payoff [21,22]. When defectors dominate and payoffs from the game are low, this option becomes more advantageous than mutual defection, allowing non-participants to invade and replace defectors. This crucial step enables cooperation to re-emerge, establishing a more robust cycle that prevents a total collapse to defection and allows punishment to emerge through the neutral drift between punishers and non-punishing cooperators [23]. However, this influential explanation rests on the critically narrow assumption that non-participants are passive actors, a simplification that overlooks their potential for more complex strategic behaviour.

In real-world social dilemmas, the decentralized nature of non-participation implies that an individual's freedom to opt out, with complex payoffs, can still have direct consequences for the public good. This active role stands in sharp contrast to the assumption of passive non-participation. For instance, in public healthcare, an individual opting for private services may receive a net positive payoff (better care) as well as a negative one (higher fees) [24,25], while their departure simultaneously reduces overcrowding, a positive influence on the public system. Conversely, in a knowledge-sharing group, the departure of an influential member harms the group's collective expertise while offering the departing individual a potential gain (e.g. increase in prestige) or loss (e.g. loss of collaborative opportunities) [26,27]. This real-world complexity motivates a critical re-examination of non-participation, raising key questions: will this mechanism, when more broadly defined, remain effective in enabling the evolution of altruistic punishment? And under what conditions will it help or hinder it?

To answer these questions, we generalize non-participation in the public goods game (PGG) by introducing a new parameter: the direct impact of non-participation on the public good (σ_2), reflecting a beneficial or harmful influence on the public good available to the remaining participants (see figure 1). Additionally, we extend the non-participation payoff (σ_1) beyond the non-negative range, representing the incentive or penalty for opting out. Alongside the non-participation, our model incorporates cooperation, defection and altruistic punishment. In the game, cooperators and punishers contribute to the public pool, while defectors do not. The collective contribution in the public pool will be enhanced and then modified by the impact of all non-participants before being distributed equally among all participants.

Through evolutionary analysis in well-mixed finite populations, we find that the capacity of voluntary participation to support altruistic punishment depends critically on whether non-participants contribute positively or negatively to the public good. When opting out yields a positive payoff and non-participants exert a beneficial influence on the public good, altruistic punishment can dominate, and the range of synergy factors where it thrives expands compared with the traditional assumption that non-participants have no impact on the public good. In contrast, when non-participants undermine the public good, altruistic punishment becomes harder to sustain and the viable synergy range narrows relative to the no-impact baseline. Notably, under positive impacts on the public good, the dominance of altruistic punishment requires positive incentives for non-participation, whereas under negative impacts, it can persist even when opting out is discouraged. These findings underscore the importance of accounting for the active role of non-participants when designing exit-based mechanisms to address the challenge of sustaining altruistic punishment.

2. Model and methods

2.1. Game set-up

In this study, we extend the assumption of non-participation within the framework of the PGG, utilizing one critical parameter: the direct impact σ_2, which represents the influence that non-participants have on the public good. Within the PGG involving G players, each player has four distinct strategic options: cooperation (C), defection (D), punishment (P) and non-participation (N), as illustrated in figure 1a. Cooperators and punishers contribute an endowment c into the public pool, while defectors contribute nothing, and non-participants opt out of the game, receiving a payoff σ_1. First, all contributions are multiplied by the synergy factor r as the public good. Each non-participant will impact the public good by σ_2. Then, the resulting public good is equally distributed among all players who have participated (i.e. excluding non-participants). Additionally, punishers incur a cost α to impose a loss $-\beta$ on defectors. Let x, y, z and w denote the number of players adopting C, D, P and N, respectively, where $x+y+z+w=G$. Therefore, the payoffs for the four strategies are given by

$$\begin{aligned}\pi_C &= \frac{(x+z)rc + w\sigma_2}{x+y+z} - c,\\ \pi_D &= \frac{(x+z)rc + w\sigma_2}{x+y+z} - z\beta,\\ \pi_P &= \frac{(x+z)rc + w\sigma_2}{x+y+z} - c - y\alpha,\\ \pi_N &= \sigma_1.\end{aligned} \tag{2.1}$$

Note that if only a single player chooses to participate, $x+y+z=1$, the PGG fails to be established, and the solitary participant also receives the non-participation payoff σ_1, as the other w N-players.

As depicted in figure 1b,c, the non-participant's payoff, $-1 \le \sigma_1 \le 1$, represents the net outcome of opting out, where a positive value incentivizes non-participation and a negative value encourages participation; the non-participant's impact, $-1 \le \sigma_2 \le 1$, reflects their direct influence on the public good, where a positive value benefits the participants and a negative value induces loss. This framework moves beyond classic models that typically assume a fixed, positive payoff and zero impact for non-participants. For a clear comparison with previous models, e.g. in [23], we set $G=5$, $c=1$, $\alpha=0.3$ and $\beta=1$, unless otherwise specified.

2.2. Well-mixed finite population

We analyse the evolutionary dynamics within a finite, well-mixed population of M players. Assuming there are m_i players adopting strategy i and $M-m_i$ players adopting j, then the probability that a group of G players is randomly sampled from the population to play the PGG is

$$H(n_i, G, m_i, M) = \frac{\binom{m_i}{n_i}\binom{M-m_i}{G-n_i}}{\binom{M}{G}}, \tag{2.2}$$

where n_i players adopting strategy i and $G-n_i$ players adopting strategy j are selected to form the group. The expected payoff for a focal player is calculated by averaging over all possible compositions of the $G-1$ co-players they might interact with in a sampled game group (see appendix A). For example, when the population consists of m cooperators and $M-m$ defectors, the average payoff for a focal cooperator and defector is

$$\begin{aligned}P_{CD} &= \sum_{x=0}^{G-1} H(x, G-1, m-1, M-1)\left(\frac{(x+1)rc}{G} - c\right) = \frac{rc}{G}\left(\frac{G-1}{M-1}(m-1)+1\right) - c,\\ P_{DC} &= \sum_{x=0}^{G-1} H(x, G-1, m, M-1)\left(\frac{xrc}{G}\right) = \frac{rc}{G}\frac{G-1}{M-1}m.\end{aligned} \tag{2.3}$$

The evolution of strategies is modelled using a Moran process with a pairwise social learning mechanism. In each time step, one player i is randomly chosen for strategy revision, and another player j is chosen as a potential role model. The probability that player i adopts player j's strategy is determined by their fitness difference according to the Fermi function [28,29],

$$p_{i\to j}(m_i) = \frac{1}{1+\exp[-s(P_{ji}-P_{ij})]}, \tag{2.4}$$

where P_{ij} and P_{ji} are the expected payoffs (fitness) of players i and j, respectively. The parameter s represents the intensity of selection, which is also referred to as the inverse temperature. It controls how strongly imitation depends on fitness differences. For $s=0$, imitation is random (neutral drift), while for $s\to\infty$, players always imitate strategies with higher fitness.

To determine the long-term success of each strategy, we first calculate the fixation probability, ρ_{ij}, which is the probability that a single mutant of strategy i will eventually take over a resident population of $M-1$ players of strategy j. The transition probabilities $T_{ij}^{\pm}$ for the number of i players to increase or decrease by one are given by

$$T_{ij}^{\pm}(m_i) = \frac{M-m_i}{M}\frac{m_i}{M}\frac{1}{1+\exp[\pm s(P_{ji}-P_{ij})]}. \tag{2.5}$$

The fixation probability ρ_{ij} is then calculated as [29]

$$\rho_{ij} = \frac{1}{1 + \sum_{k=1}^{M-1} \prod_{m_i=1}^{k} \frac{T_{ij}^{-}(m_i)}{T_{ij}^{+}(m_i)}}. \tag{2.6}$$

Assuming a small mutation limit, where any mutant either fixates or goes extinct before another mutation occurs [30,31], the fixation probabilities ρ_{ij} define the transition matrix, A, of a Markov process between the four homogeneous population states [32,33]. The elements of this matrix are given by $A_{ij,i\neq j} = \rho_{ij}/(q-1)$ and $A_{ii} = 1 - \sum_{j=1,j\neq i}^{q} A_{ij}$, where q is the number of strategies. The long-term outcome of this evolutionary process is captured by the stationary distribution of this matrix. Mathematically, this distribution corresponds to the normalized eigenvector of the transposed transition matrix with an eigenvalue of 1. In our analysis, this stationary distribution represents the fraction of time the population is expected to spend in each monomorphic state in the long run. We therefore use it as our primary measure of a strategy's long-term evolutionary success.

2.3. Risk dominance

To analyse the short-term invasion dynamics between any two strategies, we use the concept of risk dominance. Strategy *i is* risk dominant against strategy *j* when it satisfies [34,35]

$$\sum_{k=1}^{G} \pi_i(k) \geq \sum_{k=0}^{G-1} \pi_j(k+1), \tag{2.7}$$

where $\pi_i(k)$ and $\pi_j(k+1)$ are the payoffs of strategy i and strategy j, respectively, when their group consists of k players of strategy i and $G-k$ players of strategy j. Detailed derivations for the conditions under which each strategy is risk-dominant can be found in appendix B.

3. Results

3.1. Non-participation has a conditional impact on the evolution of altruistic punishment

We first present our general finding that compared with the case without the opting-out option, the effectiveness of non-participation in promoting the emergence of punishment is highly conditional. This finding is robust, as evidenced by our examination of the stationary distribution of punishment averaged over 10 000 randomly sampled parameter sets, as shown in figure 2 (see also figure 7 in appendix C).

The non-participant's payoff, σ_1, plays the most decisive role [17,36]. When this payoff is negative (the top row in figure 2 and the first column in figure 7), punishment is strongly suppressed, with its frequency remaining near zero in the vast majority of parameter settings, regardless of the non-participant's impact. In addition, the frequencies of cooperation and non-participation also remain near zero, while defection becomes the overwhelmingly dominant strategy, with its frequency approaching 1 in the vast majority of parameter settings, regardless of the non-participant's impact. A negative σ_1 removes the incentives for non-participation, namely, the viability of voluntary participation. This causes the system to collapse to a state of full defection where no public good is produced, thereby rendering any externality inconsequential. Conversely, a positive payoff (the bottom row in figure 2 and the second column in figure 7) is a necessary condition for punishment to evolve. Within this positive payoff regime, however, the non-participant's direct impact plays a significant and intricate role, creating polarized outcomes where either punishment or, in some cases, non-participation can flourish to dominate the population. While the frequency of cooperation is slightly enhanced in this regime, its prevalence remains low, typically below 0.2. In essence, a positive payoff for non-participants is essential for punishment to emerge, but the nature of the non-participant's impact then determines the balance between its risk of extinction and its ultimate success.

To better understand this general finding, below we provide a detailed analysis of how the positive or negative impact of non-participation can significantly influence evolutionary dynamics, contrasting with the no-impact setting typically considered in previous studies [11,23].

3.2. Impact of non-participants alters the synergy threshold for altruistic punishment to dominate

First, compared with the case without impact, the direct impact of non-participants significantly alters the synergy factor required for punishment to dominate the population (i.e. greater than 50% in the stationary distribution), with the positive impact lowering the threshold and the negative impact raising it. Compared with the scenario with no impact ($\sigma_2 = 0$, dashed line in figure 3a–c), where punishment prevails when $r \geq 2$, a positive impact considerably extends this range for the dominance of punishment, lowering the required synergy factor to approximately $r \approx 1.4$ when $\sigma_2 = 1$. Conversely, a negative impact restricts these conditions, requiring a larger synergy factor to $r > 3$ when $\sigma_2 = -1$. A detailed analysis of risk dominance is shown in appendix B. Additionally, punishment is the priority dominant cooperative behaviour with higher stationary distribution, rather than cooperation, comparing figure 3b,c.

The underlying evolutionary dynamics reveal that the stability of a cyclic dominance among C, D and N is the key mechanism for the dominance of punishment. For example, when $r = 3$, in the case without impact ($\sigma_2 = 0$, figure 3d), punishment is strongly

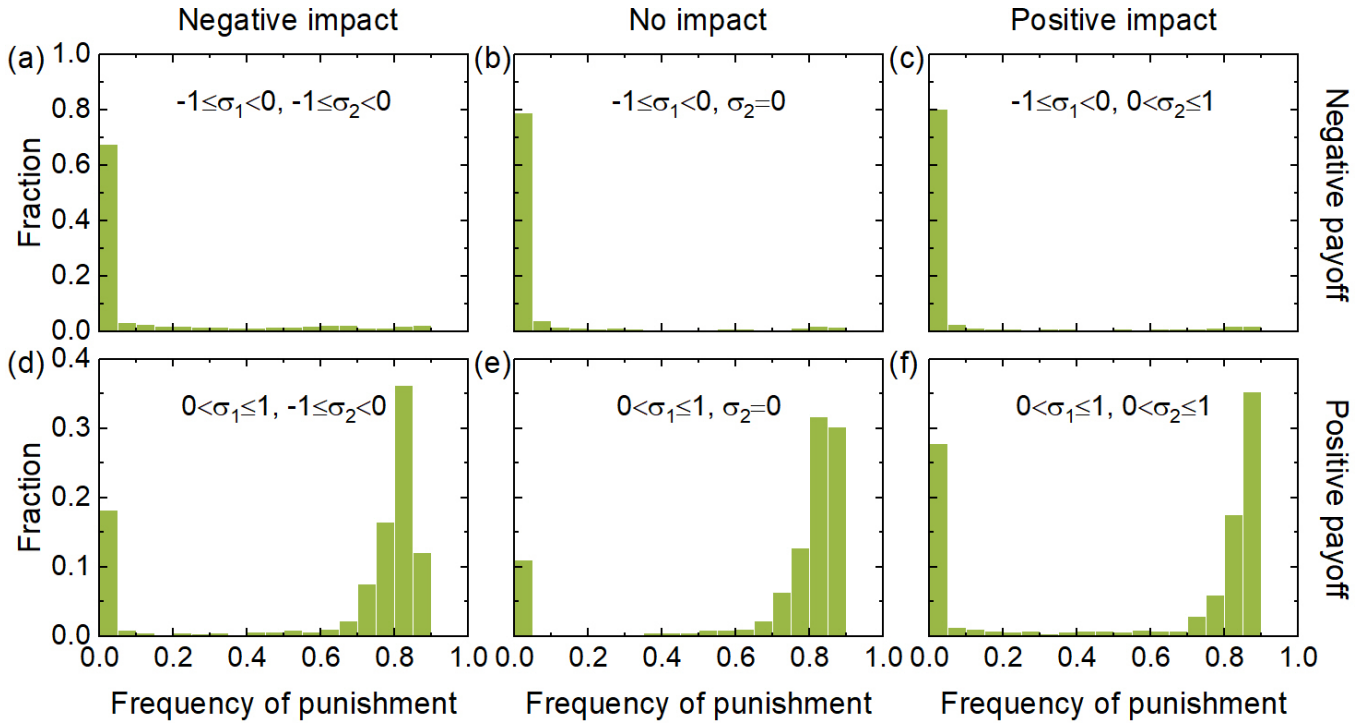

Figure 2. Compared with the case without voluntary participation, the effectiveness of non-participation for promoting the emergence of punishment is limited. Shown are the results of 10 000 numerical calculations for $s = 1$. Parameters are randomly sampled from uniform distributions on the intervals $\alpha \in [0, 1]$, $\beta \in [\alpha, 5]$, $r \in [1, 5]$ and (a) $\sigma_1 \in [-1, 0)$, $\sigma_2 \in [-1, 0)$, (b) $\sigma_1 \in [-1, 0]$, $\sigma_2 = 0$, (c) $\sigma_1 \in [-1, 0)$, $\sigma_2 \in (0, 1]$,(d) $\sigma_1 \in (0, 1]$, $\sigma_2 \in [-1, 0)$, (e) $\sigma_1 \in (0, 1]$, $\sigma_2 = 0$, (f) $\sigma_1 \in (0, 1]$, $\sigma_2 \in (0, 1]$. Detailed cumulative fractions are shown in figure 8.

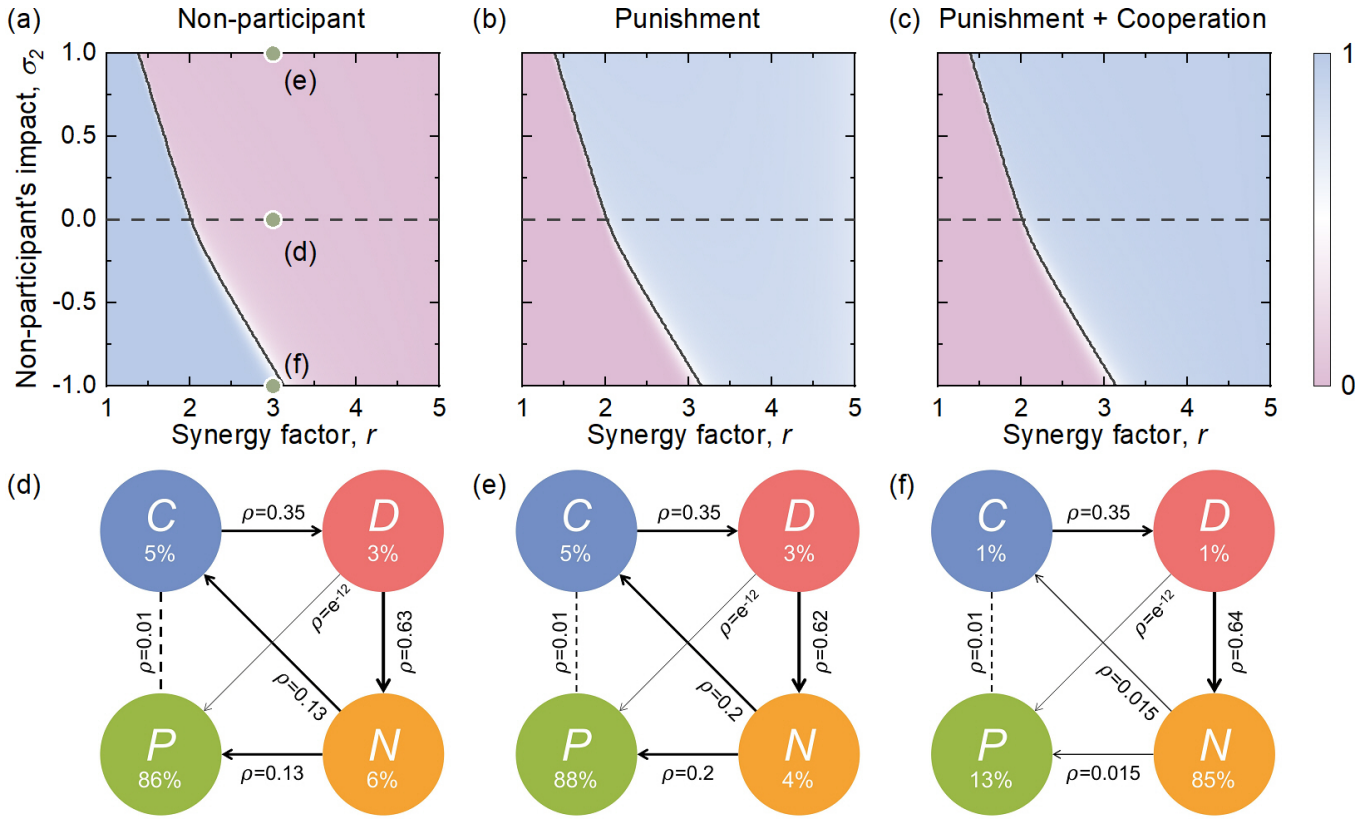

Figure 3. Compared with the case of a negligible non-participation impact, a positive impact lowers the threshold of the synergy factor (r) for punishment survival, while a negative impact increases it. (a)–(c) Show the frequency of non-participation, punishment and the amount of cooperation and punishment as a function of the synergy factor and the non-participant's impact, respectively. The dashed line represents the case of non-participation having no impact on the public good. The solid line marks the threshold beyond which the indicated strategy dominates, i.e. with a frequency of at least 50%. (d)–(f) Show the stationary distribution and the transition probabilities for the selected games indicated by the green circles in (a). Black arrows show the stronger transitions within a pair of strategies, dashed arrows show neutral transitions and ρ denotes the transition probability. Parameters are set as $s = 1$, $\sigma_1 = 1$; $r = 3$ for panels (d–f) with $\sigma_2 = 0$ in (d), $\sigma_2 = 1$ in (e) and $\sigma_2 = -1$ in (f).

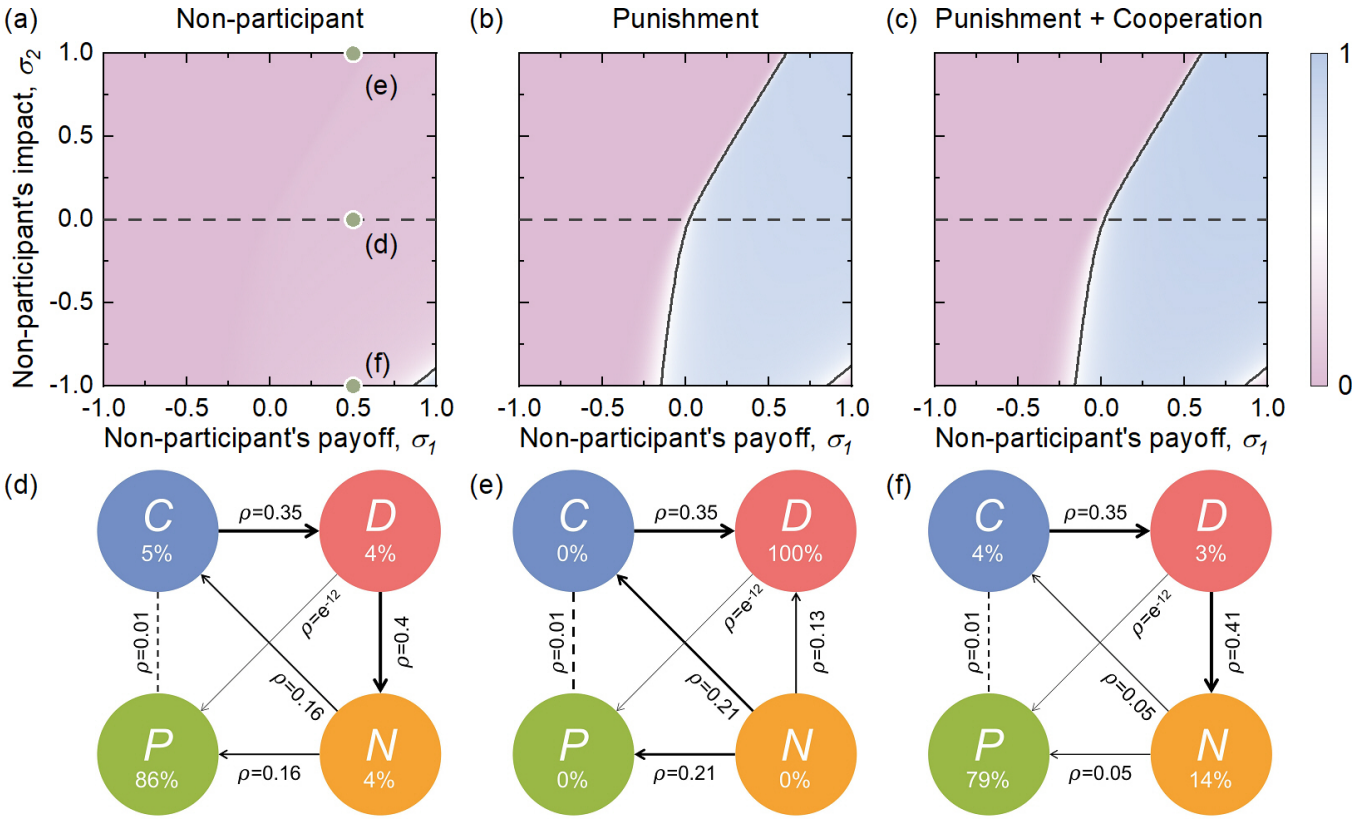

Figure 4. A positive impact from non-participants requires a higher payoff incentive for punishment to dominate, while a negative impact allows punishment to persist even when non-participation is disincentivized. (a)–(c) Show the frequency of non-participation, punishment and the amount of cooperation and punishment as a function of the non-participant's payoff and the non-participant's impact, respectively. The dashed line represents the non-participation without the impact on the public good. The solid line marks the threshold beyond which the strategy dominates, meaning its stationary distribution probability is more than 50%. (d)–(f) Show the stationary distribution and the transition probability for the selected points in (a), respectively. The black arrow shows the stronger transition between the two strategies, the dashed arrow shows the neutral transition, and ρ denotes the transition probability. Parameters are set as $s = 1$, $r = 3$; $\sigma_1 = 0.5$ for panels (d–f) with $\sigma_2 = 0$ in (d), $\sigma_2 = 1$ in (e) and $\sigma_2 = -1$ in (f).

favoured, comprising 86% of the stationary distribution. When a positive impact is introduced ($\sigma_2 = 1$, figure 3e), the dominance of punishment is slightly enhanced to 88%, while non-participation is reduced. In contrast, a negative impact completely reverses this outcome ($\sigma_2 = -1$, figure 3f). Non-participation becomes a highly stable state, making up 85% of the stationary distribution. Although the transition from non-participation to punishment is weak ($\rho = 0.015$), the reverse transition towards non-participation is even weaker. This slightly maintains the cyclic dynamics among C, D and N, leading to a state of non-participation with a minimal level of punishment surviving (13%). According to our risk dominance analysis, these three cases fall within a region where both cooperation and punishment are risk-dominant over non-participation (as shown in figure 6a). Therefore, varying σ_2 in this regime primarily alters the transition probabilities rather than reversing the (stronger) transition direction between these strategies.

3.3. Impact of non-participants influences their required incentives to achieve the dominance of altruistic punishment

Second, compared with the no-impact setting, to achieve the dominance of punishment, a positive impact requires higher incentives for non-participants, while a negative impact does not, even when non-participation is disincentivized. In the scenarios with no impact ($\sigma_2 = 0$, dashed line in figure 4), punishment dominates over a wide range, prevailing as long as the non-participant's payoff σ_1 is larger than 0. However, a positive impact severely restricts these conditions by raising the σ_1 threshold, meaning punishment can only dominate if the incentive to opt out is stronger. For instance, at a high positive impact ($\sigma_2 = 1$), punishment cannot dominate unless the non-participant's payoff is larger than 0.6. Conversely, a negative impact creates a different trade-off. It allows punishment to be maintained at a slightly broader range of σ_1 values compared with the baseline, but it also causes punishment to fail if the payoff σ_1 becomes milder negative (approx. $\sigma_1 < -0.3$).

This is because a positive impact can destroy the cyclic dynamics that induce punishment. With a moderate payoff ($\sigma_1 = 0.5$) and no impact ($\sigma_2 = 0$, figure 4d), punishment thrives (86%) because a strong cycle reliably suppresses defection. However, when a strong positive impact is introduced ($\sigma_2 = 1$, figure 4e), punishment collapses completely and defection takes over (100%). The positive impact fundamentally breaks the enforcement cycle: the ability of non-participants to invade and replace defectors is severely disrupted, where the transition from defection towards non-participation is reversed. Without this crucial mechanism, the population succumbs to defection. This reversal is directly explained by risk dominance analysis, as this case crosses the boundary into the region where defection is risk-dominant over non-participation (figure 6b). In addition, when a negative impact is considered

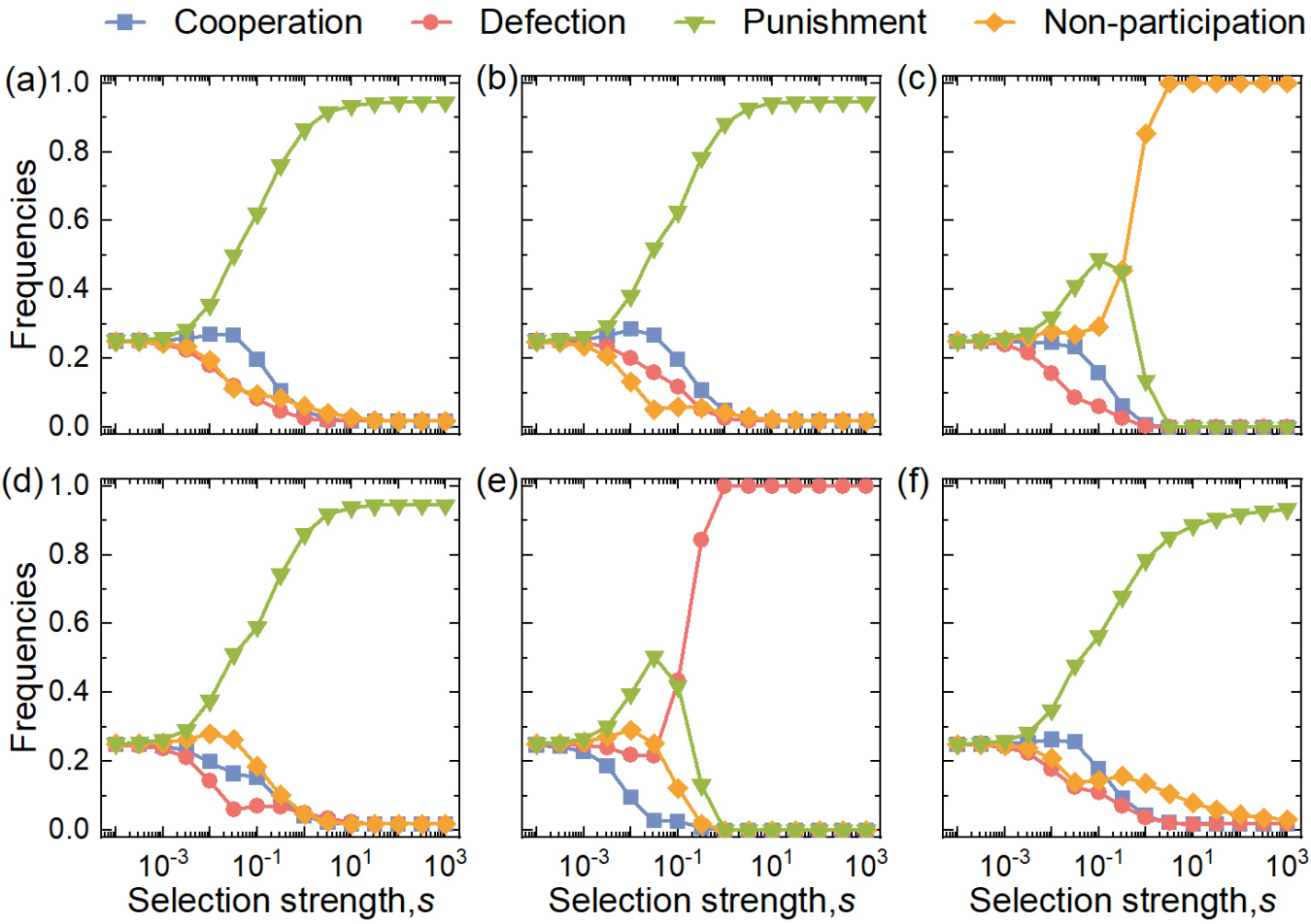

Figure 5. The robustness across strong selection scenarios. Stationary distribution as a function of selection strength s. Parameters are set as $r = 3$, (a) $\sigma_1 = 1$ and $\sigma_2 = 0$, (b) $\sigma_1 = 1$ and $\sigma_2 = 1$, (c) $\sigma_1 = 1$ and $\sigma_2 = -1$, (d) $\sigma_1 = 0.5$ and $\sigma_2 = 0$, (e) $\sigma_1 = 0.5$ and $\sigma_2 = 1$, and (f) $\sigma_1 = 0.5$ and $\sigma_2 = -1$.

($\sigma_2 = -1$, figure 4f), though the transition from non-participation towards cooperation and punishment is weakened (ρ drops from 0.16 to 0.05), the cyclic dynamics exist, and therefore, punishment prevails.

These results indicate that, though the positive impact on public good benefits participants, it potentially destroys the cyclic dynamics where non-participation loses the advantages when competing with defection, therefore, the extinction of punishment. Furthermore, the negative impact decreases public good, while the advantage of non-participation towards defection is the key to sustaining the cyclic dominance. Our findings further underscore the limitation of voluntary participation in sustaining punishment, considering the impact of non-participants on public good.

3.4. Robustness across strong selection intensity

So far, the analysis focused on an intermediate selection intensity ($s = 1$). We now examine the evolutionary outcomes across a broad spectrum of selection intensity.

Our findings are robust across the strong selection scenarios. In detail, when the non-participant payoff is high ($\sigma_1 = 1$, the top row in figure 5), a negative impact ($\sigma_2 = -1$, figure 5c) leads to the dominance of non-participation under strong selection, while neutral or positive impacts lead to the dominance of punishment (figure 5a,b). Similarly, for a moderate payoff ($\sigma_1 = 0.5$, the bottom row in figure 5), a positive impact ($\sigma_2 = 1$, figure 5e) uniquely leads to the dominance of defection under strong selection, while neutral and negative impacts sustain the dominance of punishment, where the latter decreases the frequency of punishment slightly (figure 5d,f). Further results across selection intensities for different synergy factors r are shown in figures 9 and 10. In all these cases, weak selection leads to the coexistence of four strategies (approx. $s < 1$). These findings align with our previous conclusion and further highlight the careful implementation of voluntary participation as a promoter of punishment.

4. Discussion

In this work, we have re-examined the role of voluntary participation in sustaining altruistic punishment in the one-shot PGG. To this end, we have extended the consequences of non-participation by a crucial parameter: its direct impact, which can be beneficial or harmful to the public good. Additionally, we have expanded the payoff for non-participants to a broader range, which acts as either an incentive or a penalty for opting out. Our results reveal that the effectiveness of non-participation is highly conditional.

Specifically, a negative payoff consistently leads to the dominance of defection. Within the regime of a positive payoff, the externality then dictates the precise conditions for the success of punishment. On the one hand, compared with the case without impact, the positive impact lowers the threshold of synergy factors for punishment to thrive, while the negative impact raises it. On the other hand, to achieve the dominance of punishment, the positive impact demands a higher incentive for non-participants, while the negative impact does not, even when non-participation is disincentivized. These findings emphasize the limitations of viewing voluntary participation as a universal promoter of punishment.

Our analysis also reveals an intriguing non-monotonic effect where the frequency of punishment often peaks at intermediate selection strengths. Notably, this peak is absent in the setting without externalities, suggesting the richer dynamics introduced by the non-participants' impact (figures 5, 9 and 10). We hypothesize that this is an optimal balance between selection and stochasticity: selection is strong enough for punishment to be effective, yet not so strong as to deterministically eliminate the complex, externality-driven dynamics. This benefit from intermediate stochasticity recalls the critical role of mutation, which can sustain cooperation by preventing the irreversible fixation of defection [37]. Exploring how these different sources of beneficial randomness, be it from selection intensity or mutation, interact with social mechanisms like externalities is a promising direction for future research.

Our findings build upon previous studies by showing that the effect of voluntary participation in promoting punishment is highly dependent on the assumption of passive non-participants [11,23]. We demonstrate that the cyclic dynamics inducing punishment are fragile and can be disrupted by the impact of non-participation, whether positive or negative, on the public good. For example, our results indicate that a positive impact from non-participants, while seemingly beneficial, can break the cycle dynamics by reversing the transition from defection to non-participation, leading to the prevalence of defection. In addition, a negative impact can weaken the ability of punishers and cooperators to suppress non-participants, resulting in a decrease in punishment. This result aligns with a broader re-evaluation of voluntary participation in sustaining cooperation, highlighting its limitation in solving social dilemmas given the non-negligible impact on the public good [38]. In other words, our findings underscore the necessity of carefully considering voluntary participation as a universal solution for social dilemmas from the evolutionary theoretical perspective.

The practical implications of these findings can be exemplified in real-world scenarios like public health systems. In this scenario, opting for private services can be seen as creating a positive externality on the public system by reducing waiting lists and overcrowding for the remaining participants. Our model predicts a policy paradox in such cases: to encourage the evolution of social norms, there must be a genuine and attractive incentive for individuals to opt out, such as the availability of high-quality private care. However, if this incentive becomes too strong, it risks a mass exodus that could destabilize the public good. Conversely, our results show that if there are significant penalties or high costs associated with leaving the public system, this would suppress the very sanctioning behaviours needed to maintain its standards. Therefore, managing such public goods requires a delicate balance, structuring incentives so the option to leave is viable enough to foster prosocial norms, without being so attractive that it undermines the system itself. However, our analysis relies on a symmetric framework where all non-participants are homogeneous and exert the same externality. A crucial next step is to explore these dynamics in asymmetric or heterogeneous settings, where different groups may have varying abilities or costs associated with opting out, reflecting greater real-world complexity.

Finally, we acknowledge the key simplifications in our model—the consideration of a single, homogeneous type of non-participant within a well-mixed population. By demonstrating that even this simplified, singular type of non-participant can fundamentally alter and disrupt cooperative outcomes, our findings show the critical need for further investigation into more complex scenarios. First, it would be valuable to investigate hybrid populations containing both adaptive human players and simple, committed bots. While 'loner bots' (who always opt out) in the optional prisoner's dilemma have been shown to have no impact on cooperation in well-mixed populations [39], it remains an open question whether our non-participant bots would facilitate altruistic punishment in this context. Second, the role of population structure should be explored. Future work could move from the current well-mixed setting to pairwise networks, where loner bots are known to facilitate cooperation [39] and cooperative behaviours can cascade through human social interactions [40]. An even more advanced step would be to consider higher-order networks, which more faithfully represent group interactions and can uniquely promote prosocial behaviours in ways that simple networks cannot [41]. Third, our framework could be extended to pool punishment models to investigate whether non-participant externalities can relax the assumption of overpunishing needed to sustain cooperation [42]. Finally, our framework could be applied to other games involving moral behaviours to explore how the stability of norms depends on the externalities created by non-participants [43,44]. Such work would further refine our understanding of how voluntary participation can be effectively managed to support, rather than undermine, cooperation.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. This article has no additional data.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. Z.S.: investigation, methodology, resources, software, visualization, writing—original draft; C.S.: conceptualization, investigation, writing—review and editing; V.C.: project administration, validation, writing—review and editing; T.A.H.: conceptualization, supervision, writing—review and editing. All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. We acknowledge the support provided by EPSRC (grant EP/Y00857X/1) to Z.S. and T.A.H., and JSPS KAKENHI (Grant no. JP 23H03499) to C.S.

Appendix A.

For the population, we denote by m_x, m_y, m_z and m_w the numbers of cooperators, defectors, punishers and non-participants, respectively, where $m_x + m_y + m_z + m_w = M$. Among the selected players, if only one player participates in the game, the player acts as a non-participant, which happens with probability $\binom{m_w}{G-1}/\binom{M-1}{G-1}$. Otherwise, the expected payoffs for pairwise encounters are:

$$\begin{aligned}
P_{CD} &= \frac{rc}{G}\left(\frac{G-1}{M-1}(m_x-1)+c\right)-c,\\
P_{DC} &= \frac{rc}{G}\frac{G-1}{M-1}m_x,\\
P_{CP} &= P_{PC} = rc-1,\\
P_{CN} &= P_{PN} = \left(1-\frac{\binom{m_w}{G-1}}{\binom{M-1}{G-1}}\right)\left(\frac{G\sigma_2}{\binom{M-1}{G-1}}\sum_{n_i=1}^{G-1}\frac{\binom{M-m_w-1}{n_i}\binom{m_w}{G-1-n_i}}{n_i+1}-\sigma_2+rc-c\right)+\frac{\binom{m_w}{G-1}}{\binom{M-1}{G-1}}\sigma_1,\\
P_{NC} &= P_{ND} = P_{NP} = \sigma_1,\\
P_{DN} &= \left(1-\frac{\binom{m_w}{G-1}}{\binom{M-1}{G-1}}\right)\left(\frac{G\sigma_2}{\binom{M-1}{G-1}}\sum_{n_i=1}^{G-1}\frac{\binom{M-m_w-1}{n_i}\binom{m_w}{G-1-n_i}}{n_i+1}-\sigma_2\right)+\frac{\binom{m_w}{G-1}}{\binom{M-1}{G-1}}\sigma_1,\\
P_{DP} &= \frac{(G-1)(rc-G\beta)}{G(M-1)}m_z,\\
P_{PD} &= \frac{(G-1)(rc+G\alpha)}{G(M-1)}(m_z-1)+\frac{rc}{G}-c-\alpha(G-1).
\end{aligned} \tag{A 1}$$

Appendix B.

Here, we derive the risk dominance conditions for the non-trivial pairwise comparisons among the four strategies. Of the six possible pairs, two have predetermined outcomes based on the game's structure: defection strictly dominates cooperation, and punishment and cooperation are neutral. We therefore focus on the conditions for the four remaining key comparisons below (figure 6).

First, we derive the condition for punishment to be risk-dominant against non-participation and defection, respectively.

— Punishment is risk-dominant against defection when

$$\sum_{k=1}^{G}\left[\frac{rck}{G}-c-(G-k)\alpha\right]\geq\sum_{k=0}^{G-1}\left(\frac{rck}{G}-k\beta\right), \tag{B 1}$$

which is simplified as

$$\beta-\alpha\geq\frac{2c(G-r)}{G(G-1)}. \tag{B 2}$$

— Punishment is risk-dominant against non-participation when

$$\sigma_1+\sum_{k=2}^{G}\left(\frac{rck+(G-k)\sigma_2}{k}-c\right)\geq\sum_{k=0}^{G-1}\sigma_1, \tag{B 3}$$

which is simplified as

$$rc-c+\frac{H_G G-2G+1}{G-1}\sigma_2\geq\sigma_1, \tag{B 4}$$

where $H_G=\sum_{k=1}^{G}\frac{1}{k}$.

Then, we derive the condition for non-participation to be risk-dominant against cooperation and defection, respectively.

— Non-participation is risk-dominant against cooperation when

$$\sum_{k=1}^{G}\sigma_1\geq\sigma_1+\sum_{k=0}^{G-2}\left(\frac{rc(G-k)+k\sigma_2}{G-k}-c\right), \tag{B 5}$$

which is simplified as

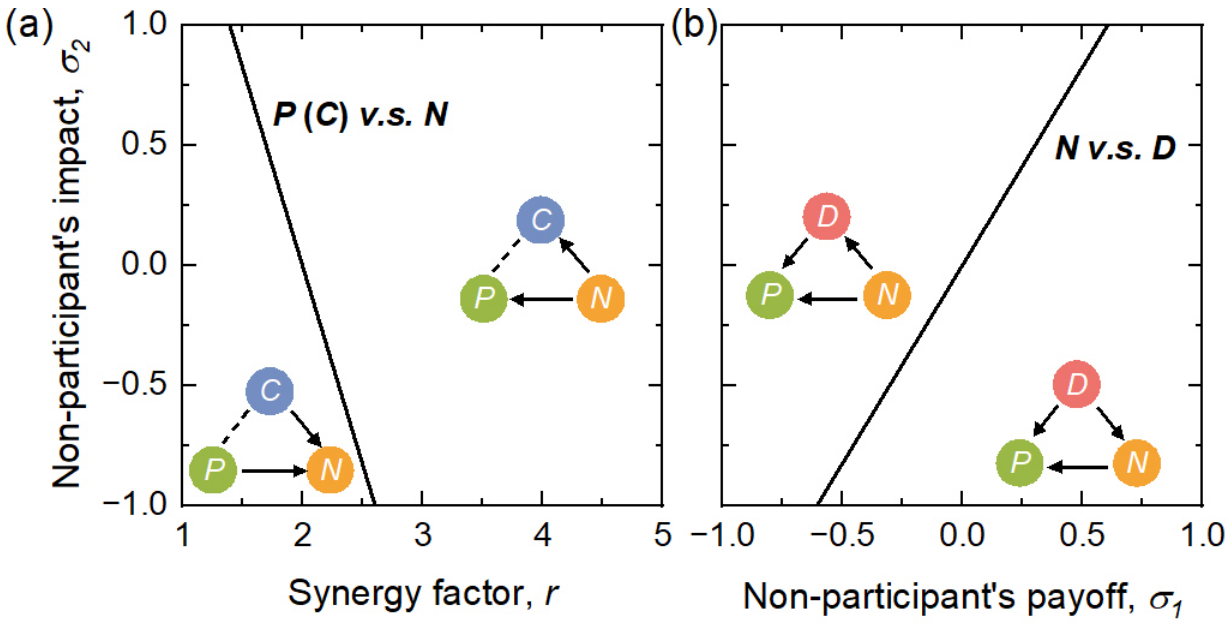

Figure 6. Risk dominance analysis explains the direction of evolutionary transitions. The phase diagrams show the analytically derived regions where each strategy is risk-dominant over its competitors. (a) Boundaries in the (σ_2, r) space for a fixed $\sigma_1 = 1$, corresponding to figure 3. For these parameters, both punishment and non-participation are always risk-dominant over defection, so only the dynamic boundary between the punishment (cooperation) and non-participation is shown. (b) Boundaries in the (σ_2, σ_1) space for a fixed $r = 3$, corresponding to figure 4. In this case, punishment is always risk-dominant over both defection and non-participation, and cooperation is also risk-dominant over non-participation, so only the boundary between defection and non-participation is shown.

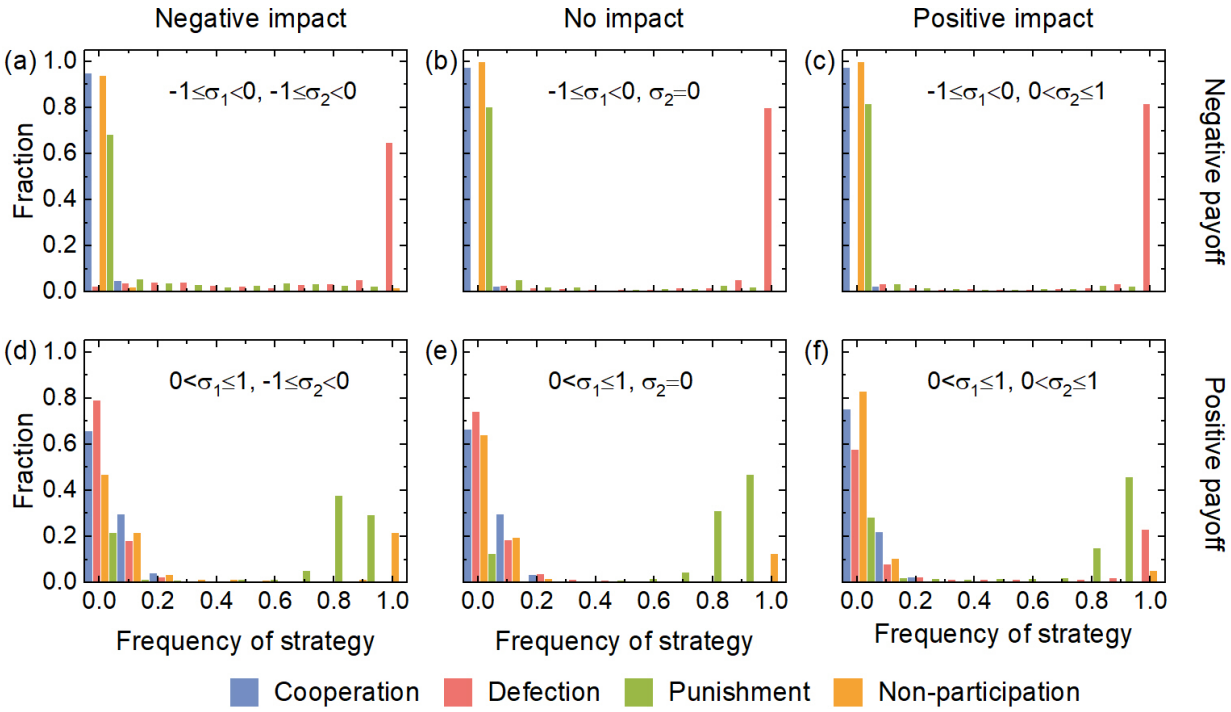

Figure 7. The effectiveness of non-participation for promoting the emergence of punishment is limited. Shown are the results of 10 000 numerical calculations for $s = 1$. Parameters are randomly sampled from uniform distributions on the intervals $a \in [0, 1]$, $\beta \in [\alpha, 5]$, $r \in [1, 5]$, and (a) $\sigma_1 \in [-1, 0)$, $\sigma_2 \in [-1, 0)$, (b) $\sigma_1 \in [-1, 0]$, $\sigma_2 = 0$, (c) $\sigma_1 \in [-1, 0)$, $\sigma_2 \in (0, 1]$, (d) $\sigma_1 \in (0, 1]$, $\sigma_2 \in [-1, 0)$, (e) $\sigma_1 \in (0, 1]$, $\sigma_2 = 0$, (f) $\sigma_1 \in (0, 1]$, $\sigma_2 \in (0, 1]$.

$$\sigma_1 \geq \frac{GH_G - 2G + 1}{G - 1}(rc + \sigma_2) - c. \tag{B 6}$$

— Non-participation is risk-dominant against defection when

$$\sum_{k=1}^{G} \sigma_1 \geq \sigma_1 + \sum_{k=0}^{G-2} \frac{k\sigma_2}{G - k}, \tag{B 7}$$

which is simplified as

$$\sigma_1 \geq \frac{GH_G - 2G + 1}{G - 1}\sigma_2. \tag{B 8}$$

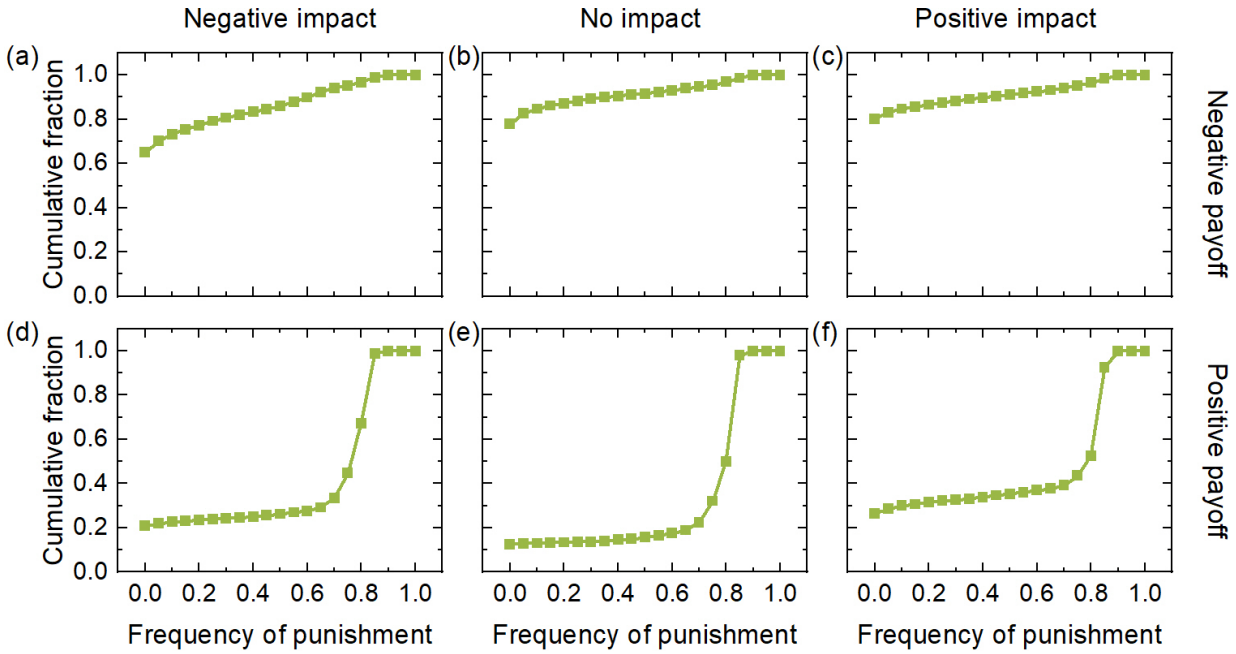

Figure 8. Cumulative fractions of the frequency of punishment. Shown are the results of 10 000 numerical calculations for $s = 1$. Parameters are randomly sampled from uniform distributions on the intervals $\alpha \in [0, 1]$, $\beta \in [, 5]$, $r \in [1, 5]$, and (a) $\sigma_1 \in [-1, 0)$, $\sigma_2 \in [-1, 0)$, (b) $\sigma_1 \in [-1, 0]$, $\sigma_2 = 0$, (c) $\sigma_1 \in [-1, 0)$, $\sigma_2 \in (0, 1]$, (d) $\sigma_1 \in (0, 1]$, $\sigma_2 \in [-1, 0)$, (e) $\sigma_1 \in (0, 1]$, $\sigma_2 = 0$, (f) $\sigma_1 \in (0, 1]$, $\sigma_2 \in (0, 1]$.

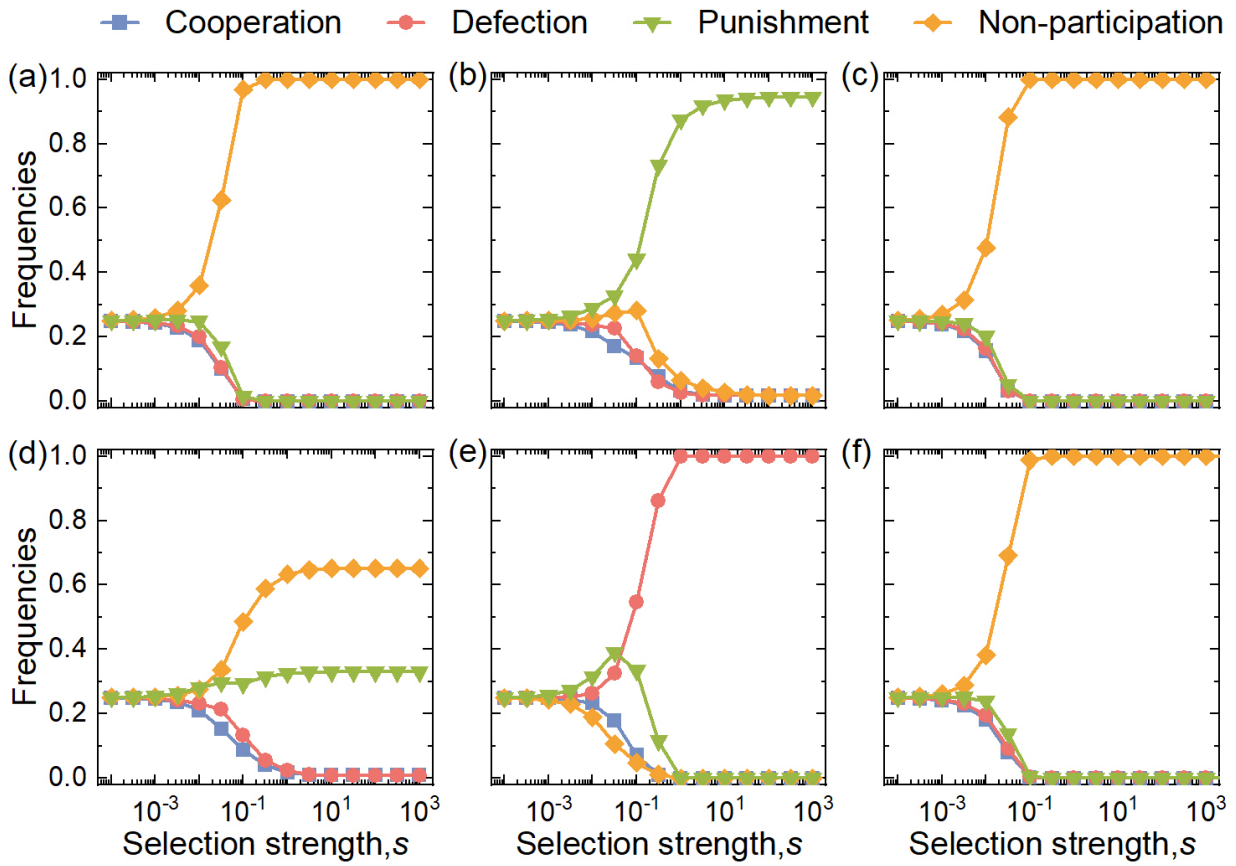

Figure 9. The robustness across strong selection scenarios. Shown are the stationary distributions as a function of selection strength s. Parameters are set as $r = 1.5$, (a) $\sigma_1 = 1$ and $\sigma_2 = 0$, (b) $\sigma_1 = 1$ and $\sigma_2 = 1$, (c) $\sigma_1 = 1$ and $\sigma_2 = -1$, (d) $\sigma_1 = 0.5$ and $\sigma_2 = 0$, (e) $\sigma_1 = 0.5$ and $\sigma_2 = 1$, and (f) $\sigma_1 = 0.5$ and $\sigma_2 = -1$.

Appendix C.

(See figures 7–10)

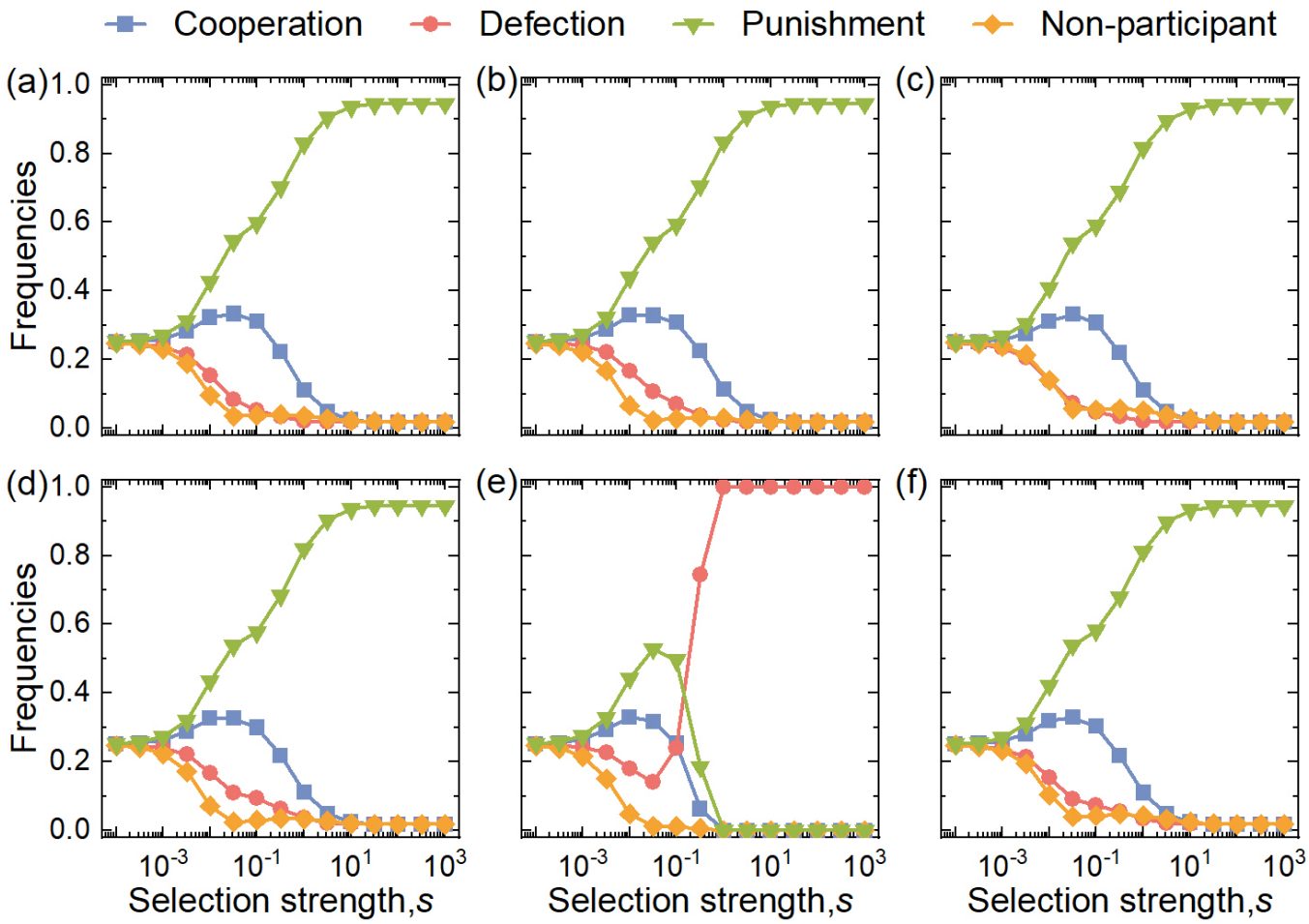

Figure 10. The robustness across strong selection scenarios. Shown are the stationary distributions as a function of selection strength s. Parameters are set as $r = 4.5$, (a) $\sigma_1 = 1$ and $\sigma_2 = 0$, (b) $\sigma_1 = 1$ and $\sigma_2 = 1$, (c) $\sigma_1 = 1$ and $\sigma_2 = -1$, (d) $\sigma_1 = 0.5$ and $\sigma_2 = 0$, (e) $\sigma_1 = 0.5$ and $\sigma_2 = 1$, and (f) $\sigma_1 = 0.5$ and $\sigma_2 = -1$.

References

1. Sigmund K, Hauert C, Nowak MA. 2001 Reward and punishment. *Proc. Natl Acad. Sci. USA* **98**, 10757–10762. (doi:10.1073/pnas.161155698)
2. Boyd R, Gintis H, Bowles S, Richerson PJ. 2003 The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535. (doi:10.1073/pnas.0630443100)
3. Raihani NJ, Thornton A, Bshary R. 2012 Punishment and cooperation in nature. *Trends Ecol. Evol.* **27**, 288–295. (doi:10.1016/j.tree.2011.12.004)
4. Fehr E, Gächter S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)
5. Boehm C, Barclay HB, Dentan RK, Dupre MC, Hill JD, Kent S, Knauft BM, Otterbein KF, Rayner S. 1993 Egalitarian behavior and reverse dominance hierarchy [and comments and reply]. *Curr. Anthropol.* **34**, 227–254. (doi:10.1086/204166)
6. Yamagishi T. 1986 The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51**, 110–116. (doi:10.1037//0022-3514.51.1.110)
7. Gürerk O, Irlenbusch B, Rockenbach B. 2006 The competitive advantage of sanctioning institutions. *Science* **312**, 108–111. (doi:10.1126/science.1123633)
8. Kaul I, Conceição P, Goulven KL, Mendoza RU. 2003 *Providing global public goods: managing globalization*. New York, NY, USA: Oxford University Press.
9. Fehr E, Gächter S. 2000 Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994. (doi:10.1257/aer.90.4.980)
10. Egas M, Riedl A. 2008 The economics of altruistic punishment and the maintenance of cooperation. *Proc. R. Soc. B* **275**, 871–878. (doi:10.1098/rspb.2007.1558)
11. Sigmund K, De Silva H, Traulsen A, Hauert C. 2010 Social learning promotes institutions for governing the commons. *Nature* **466**, 861–863. (doi:10.1038/nature09203)
12. Denant-Boemont L, Masclet D, Noussair CN. 2007 Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Econ. Theory* **33**, 145–167. (doi:10.1007/s00199-007-0212-0)
13. Janssen MA, Bushman C. 2008 Evolution of cooperation and altruistic punishment when retaliation is possible. *J. Theor. Biol.* **254**, 541–545. (doi:10.1016/j.jtbi.2008.06.017)
14. Fehr E, Rockenbach B. 2003 Detrimental effects of sanctions on human altruism. *Nature* **422**, 137–140. (doi:10.1038/nature01474)
15. Powers ST, Lehmann L. 2013 The co-evolution of social institutions, demography, and large-scale human cooperation. *Ecol. Lett.* **16**, 1356–1364. (doi:10.1111/ele.12178)
16. Boyd R, Richerson PJ. 1992 Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195. (doi:10.1016/0162-3095(92)90032-Y)
17. Han TA. 2016 Emergence of social punishment and cooperation through prior commitments. In *Proc. of the 30th AAAI Conf. on Artificial Intelligence*, vol. **30**, pp. 2494–2500, Palo Alto, CA, USA: AAAI Press. (doi:10.1609/aaai.v30i1.10120)
18. Han TA, Pereira LM, Lenaerts T. 2017 Evolution of commitment and level of participation in public goods games. *Auton. Agent. Multi Agent Syst.* **31**, 561–583. (doi:10.1007/s10458-016-9338-4)
19. Fowler JH. 2005 Altruistic punishment and the origin of cooperation. *Proc. Natl Acad. Sci. USA* **102**, 7047–7049. (doi:10.1073/pnas.0500938102)
20. Brandt H, Hauert C, Sigmund K. 2006 Punishing and abstaining for public goods. *Proc. Natl Acad. Sci. USA* **103**, 495–497. (doi:10.1073/pnas.0507229103)
21. Hauert C, De Monte S, Hofbauer J, Sigmund K. 2002 Replicator dynamics for optional public good games. *J. Theor. Biol.* **218**, 187–194. (doi:10.1006/jtbi.2002.3067)

22. Requejo RJ, Camacho J, Cuesta JA, Arenas A. 2012 Stability and robustness analysis of cooperation cycles driven by destructive agents in finite populations. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **86**, 2026105. (doi:10.1103/PhysRevE.86.026105)
23. Hauert C, Traulsen A, Brandt H, Nowak MA, Sigmund K. 2007 Via freedom to coercion: the emergence of costly punishment. *Science* **316**, 1905–1907. (doi:10.1126/science.1141588)
24. Meleddu M, Pulina M, Scuderi R. 2020 Public and private healthcare services: what drives the choice? *Socio. Econ. Plan. Sci*. **70**, 100739. (doi:10.1016/j.seps.2019.100739)
25. Hoel M, Saether EM. 2003 Public health care with waiting time: the role of supplementary private health care. *J. Health Econ*. **22**, 599–616. (doi:10.1016/S0167-6296(03)00007-9)
26. Tan D, Rider CI. 2017 Let them go? How losing employees to competitors can enhance firm status. *Strateg. Manag. J.* **38**, 1848–1874. (doi:10.1002/smj.2630)
27. Das BL. 2013 Employee retention: a review of literature. *J. Bus. Manag*. **14**, 8–16. (doi:10.9790/487X-1420816)
28. Traulsen A, Pacheco JM, Nowak MA. 2007 Pairwise comparison and selection temperature in evolutionary game dynamics. *J. Theor. Biol*. **246**, 522–529. (doi:10.1016/j.jtbi.2007.01.002)
29. Traulsen A, Nowak MA, Pacheco JM. 2006 Stochastic dynamics of invasion and fixation. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys*. **74**, 1011909. (doi:10.1103/PhysRevE.74.011909)
30. Nowak MA, Sasaki A, Taylor C, Fudenberg D. 2004 Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650. (doi:10.1038/nature02414)
31. Imhof LA, Fudenberg D, Nowak MA. 2005 Evolutionary cycles of cooperation and defection. *Proc. Natl Acad. Sci. USA* **102**, 10797–10800. (doi:10.1073/pnas.0502589102)
32. Fudenberg D, Imhof LA. 2006 Imitation processes with small mutations. *J. Econ. Theory* **131**, 251–262. (doi:10.2139/ssrn.619203)
33. Karlin S. 2014 *A first course in stochastic processes*. New York, NY, USA: Academic Press.
34. Sigmund K. 2010 *The calculus of selfishness*. Princeton, NJ, USA: Princeton University Press. (doi:10.1515/9781400832255)
35. Gokhale CS. 2010 Evolutionary games in the multiverse. *Proc. Natl Acad. Sci. USA* **107**, 5500–5504. (doi:10.1073/pnas.0912214107)
36. Rand DG, Nowak MA. 2011 The evolution of antisocial punishment in optional public goods games. *Nat. Commun*. **2**, 434. (doi:10.1038/ncomms1442)
37. Tkadlec J. 2023 Mutation enhances cooperation in direct reciprocity. *Proc. Natl Acad. Sci. USA* **120**, 2221080120. (doi:10.1073/pnas.2221080120)
38. Khatun K, Shen C, Tanimoto J. 2025 Optional participation only provides a narrow scope for sustaining cooperation. *Phys. Rev. E* **111**, 024306. (doi:10.1103/PhysRevE.111.024306)
39. Sharma G, Guo H, Shen C, Tanimoto J. 2023 Small bots, big impact: solving the conundrum of cooperation in optional Prisoner's Dilemma game through simple strategies. *J. R. Soc. Interface* **20**, 20230301. (doi:10.1098/rsif.2023.0301)
40. Fowler JH. 2010 Cooperative behavior cascades in human social networks. *Proc. Natl Acad. Sci. USA* **107**, 5334–5338. (doi:10.1073/pnas.0913149107)
41. Majhi S, Perc M, Ghosh D. 2022 Dynamics on higher-order networks: a review. *J. R. Soc. Interface* **19**, 20220043. (doi:10.1098/rsif.2022.0043)
42. Dercole F, De Carli M, Della Rossa F, Papadopoulos AV. 2013 Overpunishing is not necessary to fix cooperation in voluntary public goods games. *J. Theor. Biol*. **326**, 70–81. (doi:10.1016/j.jtbi.2012.11.034)
43. Capraro V, Perc M. 2021 Mathematical foundations of moral preferences. *J. R. Soc. Interface* **18**, 20200880. (doi:10.1098/rsif.2020.0880)
44. Kumar A, Capraro V, Perc M. 2020 The evolution of trust and trustworthiness. *J. R. Soc. Interface* **17**, 20200491. (doi:10.1098/rsif.2020.0491)