# Stability of pairwise social dilemma games: Destructive agents, constructive agents, and their joint effects

Khadija Khatun,[1,2] Chen Shen,[3,*] Lei Shi,[4,†] and Jun Tanimoto[3,1,‡]

[1]*Interdisciplinary Graduate School of Engineering Sciences, Kyushu University, Fukuoka 816-8580, Japan*
[2]*Department of Applied Mathematics, University of Dhaka, Dhaka-1000, Bangladesh*
[3]*Faculty of Engineering Sciences, Kyushu University, Kasuga-koen, Kasuga-shi, Fukuoka 816-8580, Japan*
[4]*School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming 650221, China*

(Received 21 February 2024; revised 27 June 2024; accepted 9 August 2024; published 9 September 2024)

Destructive agents, who opt out of the game and indiscriminately harm others, paradoxically foster cooperation, representing an intriguing variant of the voluntary participation strategy. Yet, their impact on cooperation remains inadequately understood, particularly in the context of pairwise social dilemma games and in comparison to their counterparts, constructive agents, who opt out of the game but indiscriminately benefit others. Furthermore, little is known about the combined effects of both agent types on cooperation dynamics. Using replicator dynamics in infinite and well-mixed populations, we find that contrary to their role in facilitating cooperation in multiplayer games, destructive agents fail to encourage cooperation in pairwise social dilemmas. Instead, they replace defection in the prisoners' dilemma and stag-hunt games. Similarly, in the chicken game, they can destabilize or replace the mixed equilibrium of cooperation and defection, undermining cooperation in the harmony (trivial) game. Conversely, constructive agents, when their payoffs exceed their contributions to opponents, can exhibit effects similar to destructive agents. However, if their payoffs are lower, while they destabilize defection in prisoners' dilemma and stag-hunt games, they do not disrupt the cooperation equilibrium in harmony games and have a negligible impact on the coexistence of cooperation in chicken games. The combination of destructive and constructive agents does not facilitate cooperation, but instead generates complex evolutionary dynamics, including bistable, tristable, and quadstable states, with outcomes contingent on their relative payoffs and game types. These results, taken together, enhance our understanding of the impact of the voluntary participation mechanism on cooperation, contributing to a more comprehensive understanding of its influence.

DOI: 10.1103/PhysRevE.110.034307

## I. INTRODUCTION

The persistence of cooperative behavior poses a significant evolutionary puzzle. Cooperation often incurs costs for individuals to help others, while the temptation of free riding, i.e., benefiting from others' assistance without contributing, threatens to undermine cooperative efforts [1,2]. According to the principle of "survival of the fittest," free riding, which saves the cost of helping, should have more evolutionary advantages than cooperation, leading to the latter's eventual extinction [3]. Evolutionary game theory offers a robust mathematical framework to unravel this paradox [4,5]. In particular, a public goods game (PGG) is a mathematical metaphor for exploring the cooperation conundrum in multiplayer games [6,7]. In the PGG, cooperators invest in a common pool by incurring costs, whereas defectors contribute nothing. The cumulative payoff in the common pool is then multiplied by an enhancement factor and distributed to all participants, irrespective of their contribution. In scenarios where the game is one-shot and anonymous [8,9], meaning that players never interact with the same individual more than once, and reciprocity mechanisms such as reputation [10,11], costly signals [12,13], and repeated interactions [14] are absent, fostering cooperation becomes particularly challenging [15]. In such contexts, social mechanisms such as reward [16–18], punishment [8,19–21], social exclusion [22,23], prior commitment [24,25], and voluntary participation [26,27] become crucial for the emergence of cooperative behavior.

While social punishment (and reward) has fostered cooperation, its efficacy relies on identifying and tracking defectors. However, the stability of these mechanisms is threatened by second-order free riders, i.e., those who contribute but avoid the costs of punishing (or rewarding), and antisocial punishers (or rewarders), i.e., those who defect yet punish (or reward) other defectors, potentially undermining the effectiveness of these social mechanisms [28,29]. In contrast, voluntary participation emerges as a simple yet effective strategy that promotes cooperation without the complexities associated with identifying and tracking defectors [26,30]. Importantly, this social mechanism does not face the same evolutionary challenges as punishment and reward, making it a subject of extensive study.

[*] Contact author: steven_shen91@hotmail.com
[†] Contact author: shi_lei65@hotmail.com
[‡] Contact author: tanimoto@cm.kyushu-u.ac.jp

Voluntary participants, also known as loners who abstain from partaking in the benefits generated from public goods and instead receive a fixed positive payoff by opting out, can effectively establish cooperation. This is achieved through a cyclic dominance effect, where cooperation yields to defectors, who, in turn, give way to loners, and loners give way to cooperators. Extending beyond the original research, studies have explored the effects of loners in networked populations [31,32], the role of loners in punishment dilemmas [33], and various other cooperation-related issues [34]. Moreover, researchers have investigated different variants of the loner strategy, such as abstention strategies, where individuals neither pay nor receive anything while their opponents bear a participation cost [35]. Exiters, who receive a fixed payoff but contribute nothing to their opponents, also receive attention [36,37]. Studies investigate the freedom to choose between homogeneous symmetric or asymmetric public resources [38,39], hedgers who enact tit-for-tat play without cooperation in the first move [40], and other related aspects [41,42]. Although these variants differ from the loner strategy, they all demonstrate a cooperation-promotion effect.

Given that the voluntary participation mechanism is a bottom-up scheme for public goods provision [43], its influence on public goods might also be related to one's personality. The loner model reflects an individualistic trait, as loners focus solely on their own payoff without directly influencing public goods. An intriguing variant of the loner strategy is the destructive agent, who, like loners, abstain from participating in public goods but actively harm others without personal gain [44,45]. This represents a competitive trait, focusing on defeating opponents. These aspects lead us to consider the existence of constructive agents, who positively contribute to public goods and can be understood as reflecting prosocial traits. We are first curious about how destructive agents impact cooperation dynamics in pairwise social dilemma games, where distinct equilibrium points exist (e.g., dominance of cooperation in the harmony game, defection in the prisoner's dilemma, bistable equilibrium in the stag-hunt game, and mixed strategies equilibrium in the chicken game), compared to their positive effects on cooperation in public goods games, which only exhibit cooperation and defection equilibria [44,45]. Second, what would be the impact of constructive agents on cooperation dynamics, in contrast to destructive agents? Lastly, it is crucial to explore the joint effects of constructive and destructive agents on cooperation dynamics in social dilemma games, especially regarding how the presence of constructive agents may alter the influence of destructive agents on cooperation dynamics.

To explore these questions, we extend the framework of social dilemma games to incorporate both destructive and constructive agents. Initially, we analyze their effects on promoting cooperation within well-mixed populations separately, before investigating their combined impact. Our model incorporates key parameters such as dilemma strength ($D_g$, $D_r$), categorizing games into harmony, chicken, stag-hunt, and prisoner's dilemma, along with incentives for agents to exit the game $d$, and the respective damage $d_1$ and benefit $d_2$ caused by destructive and constructive agents. Utilizing replicator dynamical equations, we discover that destructive agents fail to encourage cooperation in pairwise social dilemmas, in contrast to their role in promoting cooperation in public goods games. Instead, they destabilize defection, ultimately replacing it in prisoner's dilemma and stag-hunt games while undermining cooperation in chicken and harmony games. Conversely, constructive agents sustain the coexistence of cooperation in the chicken game and minimally influence the cooperative equilibrium in the harmony game, particularly when their payoffs are less than their contributions to opponents. Otherwise, their impact tends to mimic that of destructive agents. When both constructive and destructive agents are active simultaneously, their combined influence often mirrors the effects observed when each agent type acts alone. For example, the coexistence of destructive and constructive agents can disrupt defection in the prisoner's dilemma and stag-hunt games, while also compromising cooperation in chicken and harmony games. Furthermore, in scenarios where constructive agents confer benefits exceeding their gains, these joint effects can lead to the emergence of complex dynamics, including bistable, tristable, or quadstable equilibria, contingent on game types and parameter conditions.

These results enhance our understanding of the impact of the voluntary participation mechanism on cooperation, contributing to a more comprehensive understanding of its influence.

## II. MODEL

Our method contains two necessary basic components: (a) payoff matrices and (b) population settings and game dynamics. A brief description of each section is given as follows.

### A. Payoff matrices

In this study, we assume a symmetric pairwise game, where the evolutionary dynamics of cooperation within dyadic interactions involve the strategic interplay of cooperation ($C$) and defection ($D$). In instances where both players opt for cooperation, they are endowed with the payoff denoted as $R$ (reward). Conversely, if both players choose defection, the resulting payoff is designated as $P$ (punishment). When one player cooperates while the other defects, two distinct payoffs emerge: $T$, representing the temptation to defect, signifying an advantageous outcome for the defector; and $S$, denoting the sucker's payoff, indicating a disadvantageous outcome for the cooperator. Based on the relative ordering of these payoffs, four types of social dilemma games can be identified: the prisoner's dilemma, characterized by $T > R > P > S$; the stag hunt, characterized by $R > T > P > S$; the chicken or snowdrift game, characterized by $T > R > S > P$; and the harmony game, characterized by $R > T > S > P$.

To observe cooperation dynamics, we have used the concept of universal scaling of dilemma strength [46], where $D_g = T - R$ and $D_r = P - S$ are used to quantify the game's dilemma strength, encapsulating aspects characteristic of both chicken-type dilemmas (originating from greed) and stag-hunt-type dilemmas (originating from fear). The nature of the equilibrium depends on the signs of $D_g$ and $D_r$: a prisoner's dilemma scenario, where both $D_g$ and $D_r$ are positive, leads to mutual defection as the equilibrium state. A positive $D_g$

TABLE I. Payoff matrix of social dilemma game for destructive agents.

| | $C$ | $D$ | $DA$ |
|---|---|---|---|
| $C$ | $1$ | $-D_r$ | $-d_1$ |
| $D$ | $1+D_g$ | $0$ | $-d_1$ |
| $DA$ | $d$ | $d$ | $d$ |

combined with a negative $D_r$, resembling the chicken game, results in a mixed equilibrium of cooperation and defection. The stag-hunt game, indicated by a negative $D_g$ and a positive $D_r$, presents a bistable equilibrium, where both mutual cooperation and mutual defection are stable strategies. Finally, in the harmony game scenario, where both $D_g$ and $D_r$ are negative, cooperation emerges as the dominant equilibrium strategy.

*Pairwise game with destructive agents (DA).* Incorporating destructive agents named Joker, which inflicts equal damage on both cooperators and defectors, without receiving any benefit, was initially introduced in a public goods game (PGG) [44]. In this study, we introduce destructive agents into the pairwise game as a third strategy with no payoff. Then, we relax the strong assumption (Joker does not receive any benefit) with a positive payoff from destructive agents. The benefit received by destructive agents playing with others is $d \in [0, 1)$ and the damage this imposes on its opponents is $d_1 \in [0, 1)$. The payoff matrix is given in Table I.

*Pairwise game with constructive agents (CA).* Constructive agents in pairwise games strive to equal benefits between cooperators and defectors and also receive some benefits in participation. The aid that normal players receive from playing with constructive agents is $d_2 \in [0, 1)$ and the benefit received by the constructive agent is the same as the destructive agent, i.e., $d \in [0, 1)$. The payoff matrix is given as Table II.

*Pairwise game in mixed of destructive and constructive agents.* To comprehensively assess the impact of both constructive and destructive agents, we synthesized the strategies outlined in Tables I and II to create a new payoff matrix. This matrix incorporates four strategies: cooperation ($C$), defection ($D$), constructive agents ($CA$), and destructive agents ($DA$). The detailed interactions and resultant payoffs are presented in Table III.

### B. Population setting and game dynamics

We consider a well-mixed and infinite population model, wherein individuals engage in random pairwise interactions with each other.

TABLE II. Payoff matrix of social dilemma game for constructive agents.

| | $C$ | $D$ | $CA$ |
|---|---|---|---|
| $C$ | $1$ | $-D_r$ | $d_2$ |
| $D$ | $1+D_g$ | $0$ | $d_2$ |
| $CA$ | $d$ | $d$ | $d$ |

TABLE III. Payoff matrix of social dilemma game for the combined effect of destructive and constructive agents.

| | $C$ | $D$ | $DA$ | $CA$ |
|---|---|---|---|---|
| $C$ | $1$ | $-D_r$ | $-d_1$ | $d_2$ |
| $D$ | $1+D_g$ | $0$ | $-d_1$ | $d_2$ |
| $DA$ | $d$ | $d$ | $d$ | $d$ |
| $CA$ | $d$ | $d$ | $d$ | $d$ |

*Destructive agent's game dynamics.* Let $x, y, z$ denote the fractions of cooperation, $C$, defection, $D$, and destructive agent, $DA$, in the population, where $0 \leqslant x, y, z \leqslant 1$, and $x + y + z = 1$. The expected payoff for each player is given as

$$\begin{aligned} \Pi_C &= x - D_r y - d_1 z, \\ \Pi_D &= (1 + D_g)x - d_1 z, \\ \Pi_{DA} &= d. \end{aligned} \tag{1}$$

The replicator equations are

$$\begin{aligned} \dot{x} &= x(\Pi_C - \overline{\Pi}_{DA}), \\ \dot{y} &= y(\Pi_D - \overline{\Pi}_{DA}), \\ \dot{z} &= z(\Pi_{DA} - \overline{\Pi}_{DA}), \end{aligned} \tag{2}$$

where $\overline{\Pi}_{DA} = x\Pi_C + y\Pi_D + z\Pi_{DA}$.

*Constructive agent's game dynamics.* Let $w$ denote the fractions of the constructive agent, $CA$, in the population; then, $0 \leqslant x, y, w \leqslant 1$, and $x + y + w = 1$. The expected payoff for each player and the replicator dynamics are given as Eq. (3) and (4), respectively:

$$\begin{aligned} \Pi_C &= x - D_r y + d_2 w, \\ \Pi_D &= (1 + D_g)x + d_2 w, \\ \Pi_{CA} &= d; \end{aligned} \tag{3}$$

$$\begin{aligned} \dot{x} &= x(\Pi_C - \overline{\Pi}_{CA}), \\ \dot{y} &= y(\Pi_D - \overline{\Pi}_{CA}), \\ \dot{w} &= w(\Pi_{CA} - \overline{\Pi}_{CA}), \end{aligned} \tag{4}$$

where $\overline{\Pi}_{CA} = x\Pi_C + y\Pi_D + w\Pi_{CA}$.

*Game dynamics of the joint effects of DA and CA.* When both destructive and constructive agents simultaneously interact with the cooperation and defection, then $0 \leqslant x, y, z, w \leqslant 1$ and $x + y + z + w = 1$. The expected payoff for each player is given as

$$\begin{aligned} \Pi_C &= x - yD_r - d_1 z + d_2 w, \\ \Pi_D &= (1 + D_g)x - d_1 z + d_2 w, \\ \Pi_{DA} &= d, \\ \Pi_{CA} &= d. \end{aligned} \tag{5}$$

The replicator equations are

$$\begin{aligned} \dot{x} &= x(\Pi_C - \overline{\Pi}), \\ \dot{y} &= y(\Pi_D - \overline{\Pi}), \end{aligned}$$

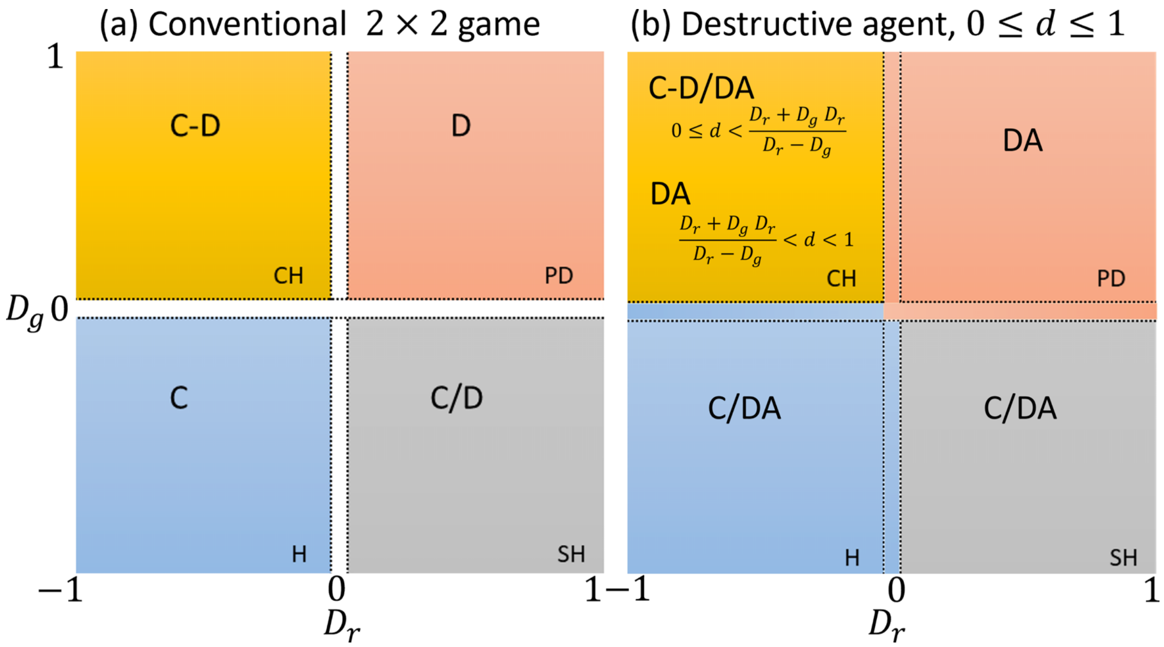
(a) Conventional 2 × 2 game
(b) Destructive agent, $0 \le d \le 1$
C-D
D
CH
PD
C
C/D
H
SH
C-D/DA
$0 \le d < \frac{D_r + D_g D_r}{D_r - D_g}$
DA
$\frac{D_r + D_g D_r}{D_r - D_g} < d < 1$
CH
DA
PD
C/DA
C/DA
H
SH
1
$D_g$ 0
−1
0
1−1
0
1
$D_r$
$D_r$

FIG. 1. Destructive agents destabilize defection for $D_r > 0$, cooperation and a mix of cooperation and defection when $D_r < 0$ (b). The defection of the prisoner's dilemma is destabilized by destruction, and stag-hunt's bistable equilibrium becomes bistable with destruction. On the other hand, chicken's mixed cooperation and defection are transformed into a bistable mix of cooperation and defection or monomorphic destruction (after a threshold value of $d$), and the cooperation of harmony turns into bistable cooperation and destruction. The diagrams can be divided into four regions (denoted by different colors) corresponding to prisoner's dilemma (PD), stag-hunt (SH), harmony (H), and chicken (CH) games, and the boundary ($D_r = 0$ and $D_g = 0$) separated by the black dotted lines. The tristability and the coexistence of cooperation, defection, and destruction are defined as C/D/DA and C-D-DA. Equilibria, stable on the boundary, are shown in the same color as the interior.

$$\dot{z} = z(\Pi_{DA} - \overline{\Pi}),$$
$$\dot{w} = w(\Pi_{CA} - \overline{\Pi}), \tag{6}$$

where $\overline{\Pi} = x\Pi_C + y\Pi_D + z\Pi_{DA} + w\Pi_{CA}$. Detailed explanations of the equilibria and their stability of all replicator dynamics have been given in the Appendix.

## III. RESULTS

### A. Destructive agents

The presence of destructive agents in a PGG, paradoxically, promotes cooperation and destabilizes defection by cyclic dominance, where cooperation leads to defection, which leads to destruction, ultimately paving the way for cooperation again [44]. In contrast, the introduction of destructive agents in the prisoner's dilemma game—special two-player PGG—fails to foster cooperation; instead, it destabilizes the equilibrium of the prisoner's dilemma game [refer to the upper right of Fig. 1(b)]. In this scenario, defection is replaced by destruction; trajectories start from an unstable node directing to destruction directly or invading cooperation by defection, and defection by destruction, as illustrated in the upper right of Fig. 2.

Beyond the prisoner's dilemma, our study extended to assess the influence of destructive agents within other pairwise social dilemma games, such as chicken, harmony, and stag-hunt. We analyzed their impact on game equilibria, focusing on mixed strategies of cooperation and defection, pure cooperation, and the bistable equilibrium between cooperation and defection. Our findings reveal that akin to observations in the prisoner's dilemma, destructive agents fail to promote cooperation; instead, they tend to destabilize existing equilibria [going back to Fig. 1(b) for detailed illustrations]. In the chicken game, the introduction of destructive agents transforms the mixed strategy equilibrium into a bistable system. This system is characterized by a possible coexistence of cooperation and defection, which is separated by a critical saddle point leading to destruction. The game dynamics evolve from two unstable equilibria towards these divergent outcomes, illustrated in Fig. 2, (upper left). The harmony game's monostable cooperation becomes bistable with destructive agents, with trajectories separated into either cooperation or destruction starting from two different unstable nodes (lower left). In the stag-hunt game, the bistable cooperation or defection equilibrium shifts to the bistable equilibrium of cooperation or destruction; trajectories stemming from an internal unstable node present two possible outcomes: either direct cooperation or destruction, which prevails over defection (lower right).

The initial assumption regarding destructive agents posits that they receive no additional payoff from opting out, which can be seen as somewhat restrictive. Given the rarity of individuals who would opt out of the game without any potential benefits, we have decided to relax this assumption. Now, agents can derive benefits from opting out of the game. Similar to nonbeneficial destructive agents, beneficial destructive agents do not facilitate cooperation. The destabilization of equilibria in all games is akin to the impact of nonbeneficial destructive agents, the only exception in the chicken

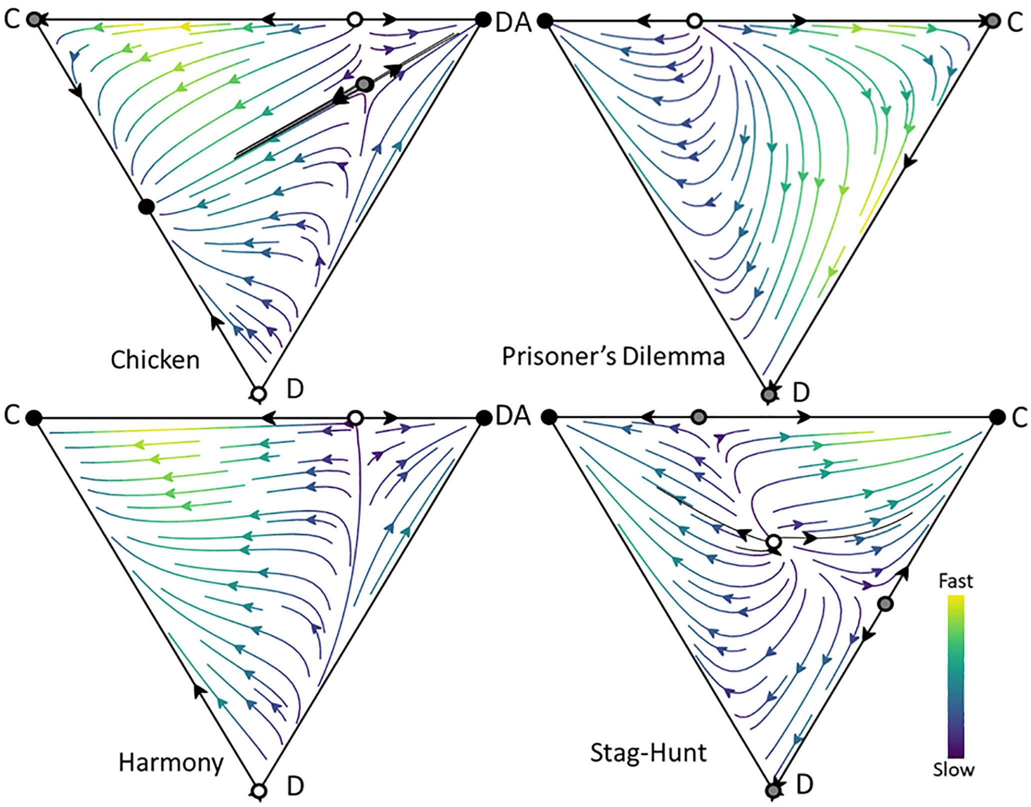

FIG. 2. The stable defection of the prisoner's dilemma is replaced by destruction; the chicken's mixed equilibrium of cooperation and defection is transformed into a bistable, either a mix of cooperation and defection or monomorphic destruction; cooperation of harmony turns into bistable cooperation and destruction; and stag-hunt bistable equilibrium's defection is replaced with destruction. The parameters are fixed at $d_1 = 0.4$, $d = 0.0$, ($D_g = D_r = 0.5$; PD), ($D_g = 0.5$, $D_r = -0.5$; CH), ($D_g = -0.5$, $D_r = -0.5$; H), and ($D_g = -0.5$, $D_r = 0.5$; SH). Solid black dots are stable nodes, whites are unstable nodes, and grays are saddle points. Images are generated by a modified version of the egttools PYTHON Package [57]. All subsequent plots are likewise generated using this same package.

game, where the mixed equilibrium is either similar to that of nonbeneficial destructive agents (when $0 \leqslant d < \frac{D_r+D_gD_r}{D_r-D_g}$) or monostable destruction ($\frac{D_r+D_gD_r}{D_r-D_g} < d < 1$; described in Appendix A 1).

At a glance, destructive agents cannot promote cooperation in pairwise social dilemmas. However, they can destabilize and potentially replace defection in the prisoner's dilemma and stag-hunt games; likewise, they can disrupt or supersede the mixed cooperation-defection equilibrium in the chicken game and undermine cooperation entirely in the harmony game. In contrast to destructive agents, which exploit or harm either cooperators or defectors, constructive agents emerge as a concept that benefits both parties equally and receives rewards for abstaining from participation. This introduces a different avenue of investigation into how constructive agents influence the dynamics of cooperation in pairwise social dilemma games, which we will explore further in subsequent analyses.

### B. Constructive agents

Similar to destructive agents, incorporating constructive agents in pairwise social dilemmas does not encourage cooperation. Rather, introducing these agents changes the stability of the equilibria in the dilemmas. Two distinct scenarios have been observed based on the relative payoffs received by constructive agents and the payoffs offered by constructive agents to others. When constructive agents experience greater payoff than the contributions they make to others, the destabilization and transformation of these agents mirror that of destructive agents, except that the outcome shifts from destruction to construction, as illustrated in Figs. 3(a) and 4 (a theoretical analysis is given in Appendix A 2).

Constructive agents, when receiving lower payoffs compared to the benefits they provide to others, disrupt defection equilibria in the prisoner's dilemma and stag-hunt games. However, their introduction has no significant impact on cooperation in the harmony game and only a negligible effect on the coexistent equilibria of cooperation and defection in the chicken game [see Fig. 3(b)]. In the prisoner's dilemma game, when trajectories originate at an unstable equilibrium of purely constructive agents and sequentially lead to cooperation and then defection, the result is a polymorphic stable mix of defection and construction that supplants the monostable defection equilibrium illustrated in Fig. 5 (upper right). Similarly, in the stag-hunt game, the bistable equilibria of cooperation and defection become bistable cooperation or

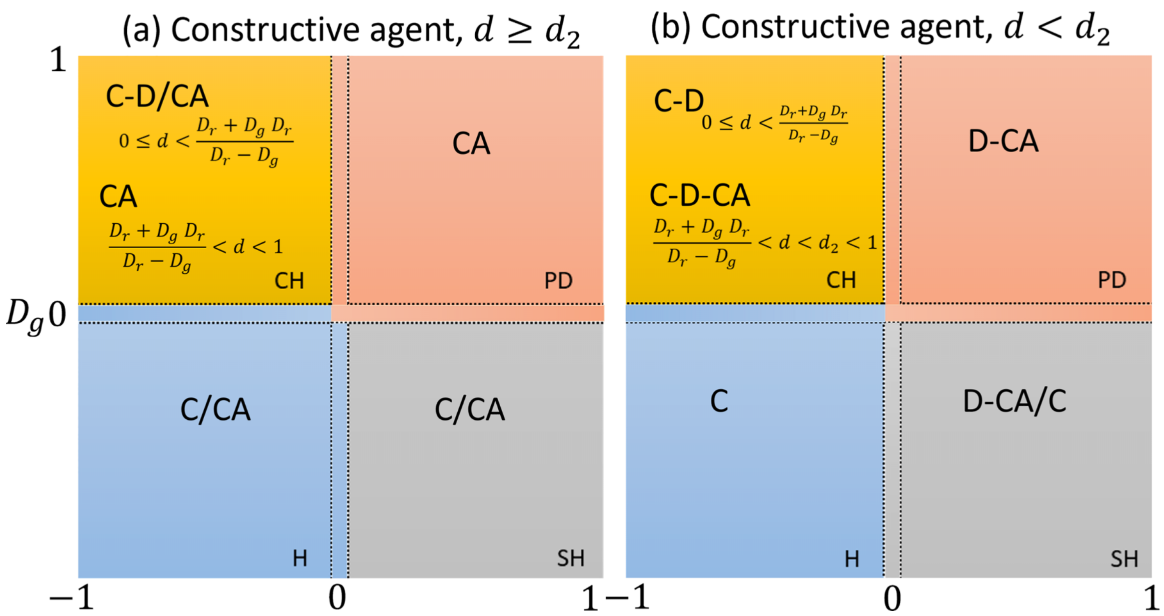
(a) Constructive agent, $d \geq d_2$
(b) Constructive agent, $d < d_2$
1
C-D/CA
$0 \leq d < \frac{D_r + D_g D_r}{D_r - D_g}$
CA
$\frac{D_r + D_g D_r}{D_r - D_g} < d < 1$
CH
CA
PD
$D_g$ 0
C/CA
H
C/CA
SH
−1
0
1
−1
0
1
C-D
$0 \leq d < \frac{D_r + D_g D_r}{D_r - D_g}$
C-D-CA
$\frac{D_r + D_g D_r}{D_r - D_g} < d < d_2 < 1$
CH
D-CA
PD
C
H
D-CA/C
SH

FIG. 3. When the constructive agents' payoff exceeds others (a), construction replaces defection if $D_r > 0$ and destabilizes cooperation and a mix of cooperation and defection if $D_r < 0$. The stable equilibrium of prisoner's dilemma and stag-hunt is construction and a bistable of cooperation and construction. In contrast, chicken's mixed equilibrium is bistable, either embracing a blend of cooperation and defection or construction or monostable construction (depending on $d$ values), and harmony's cooperation demonstrates bistability with construction. When the constructive agents' payoff is lower than others (b), defection is changed to polymorphic defection and construction if $D_r > 0$, but does not influence cooperation and a mix of cooperation and defection if $D_r < 0$. The defection of prisoner's dilemma and stag-hunt changes to a coexistence of defection and construction, and the stability of the equilibria remains unchanged in the harmony and chicken games.

a polymorphic mixture of defection and construction (lower right). The mixed equilibria of chicken's analogously may be unchanged (when $0 \leqslant d < \frac{D_r+D_gD_r}{D_r-D_g}$; see the analytical result in Appendix A 2) or shifted to polymorphic stable mixtures of cooperation, defection, and construction ($\frac{D_r+D_gD_r}{D_r-D_g} < d < 1$; upper left).

To sum up, constructive agents, when their payoffs surpass their contributions to opponents, may demonstrate effects akin to destructive agents. Conversely, when their payoffs are lower, although they destabilize defection in the prisoners' dilemma and stag-hunt games, they neither disturb cooperation in harmony games nor exert a significant influence on the coexistent equilibrium in chicken games. At this point, it is entirely natural to investigate the combined impact of both destructive and constructive agents.

### C. Mixed destructive agents and constructive agents

The introduction of both destructive and constructive agents in social dilemma games does not foster cooperation. Instead, it results in intricate evolutionary dynamics, where the end equilibrium is contingent on the relative payoff received by constructive agents and the payoffs offered by constructive agents to others. When the constructive agents' payoff exceeds the aids they have given to others, they displace defection fully in the prisoners' dilemma and stag-hunt games and can destabilize cooperation and coexistent cooperation and defection in the harmony and chicken games [refer to Fig. 6(a); see Appendix A 3 for theoretical analysis]. In the prisoner's dilemma, defection is substituted by a coexistence of destruction and construction; see Fig. 7 (upper right). In this scenario, in simplex (C, DA, CA), for instance, all trajectories either converge to cooperation or coexistence of destruction and cooperation; an introduction of mutant defection can invade cooperation [refer to the simplex (C, D, DA) in the same figure], but not the mixture which leads the mixture as final equilibrium. The bistable equilibrium of stag-hunt becomes bistable between the cooperation and coexistence of destruction and construction (lower right). All trajectories divided by a collection of unstable nodes [simplex (C, DA, CA), for example, in the same figure] converge either towards cooperation or the coexistence of destruction and construction; the introduction of mutant defection is unable to infiltrate the stability, consequently, bistability between cooperation and the mix of destruction and construction is sustained. Similarly, chicken's mixed equilibrium may become bistable, encompassing either a mixture of cooperation and defection or destruction and construction (when $0 \leqslant d < \frac{D_r+D_gD_r}{D_r-D_g}$; upper left), or monostable, encompassing a mixture of destruction and construction ($\frac{D_r+D_gD_r}{D_r-D_g} < d < 1$; see Appendix A 3), and harmony's cooperation exhibits bistability with a mix of destruction and construction (lower left).

However, when constructive agents receive lower payoffs than the benefits given to opponents, the equilibria in prisoner's dilemma and stag-hunt shift to a complex coexistence of defection, destruction, and construction, showing expanded multistability, while the equilibria in harmony and chicken remain unchanged as constructive agents have higher

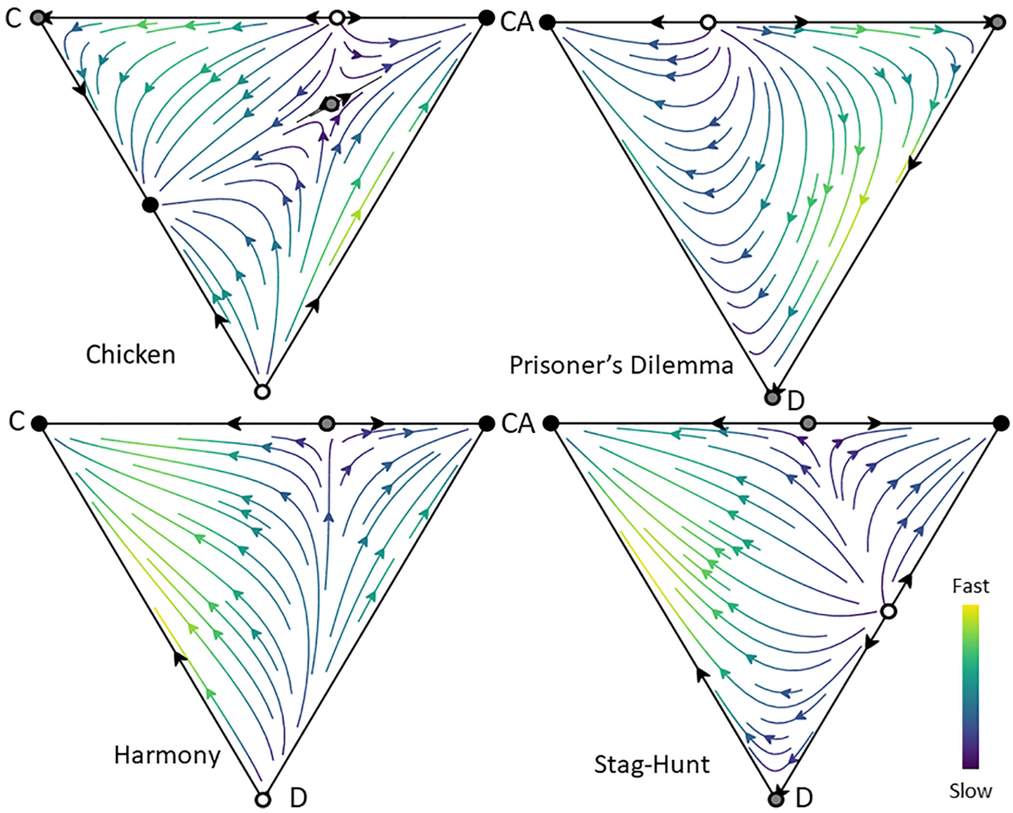

FIG. 4. When constructive agents achieve higher payoffs than others, both defection and cooperation are destabilized by it. In the prisoner's dilemma and stag-hunt, the stability of defection is replaced with construction; while chicken's mixed equilibrium becomes bistable, either embracing a blend of cooperation and defection or construction; and harmony's cooperation demonstrates bistability with construction. The parameters are fixed at $d = 0.4$ and $d_2 = 0.1$, ($D_g = D_r = 0.5$; PD), ($D_g = 0.5, D_r = -0.5$; CH), ($D_g = -0.5, D_r = -0.5$; H), and ($D_g = -0.5, D_r = 0.5$; SH). Stable nodes are marked with solid black dots, unstable nodes with white dots, and saddle points with gray dots.

payoffs, illustrated in Fig. 6(b) and theoretically analyzed in Appendix A 3. In the prisoner's dilemma, the monostable defection equilibrium is replaced by either the coexistence of defection-destruction-construction or the coexistence of destruction-cooperation or pure destruction, exhibited in Fig. 8 (upper right). In this context, trajectories in simplex (C, DA, CA) are divided by a branch of unstable nodes into cooperation or a mix of destruction and cooperation; an introduction of mutant defection can invade cooperation to a mix of defection, destruction, and construction [refer to the simplex (D, DA, CA)] or destruction only [in the simplex (C, D, DA)] in the same figure, but no influence on the mixture of destruction and cooperation, which leads a tristable state, with either coexistent of defection, destruction, and construction or a mix of destruction and construction or destruction only. Similarly, in the stag-hunt game, the bistable equilibria of cooperation and defection become tetrastable cooperation or the coexistence of defection-destruction-construction or the coexistence of destruction-cooperation or pure destruction (lower right). In this scenario, trajectories within the simplex (C, DA, CA) are partitioned by a branch of unstable nodes, creating a bistability between cooperation and a combination of destruction and construction. The introduction of mutant defection does not invade cooperation, but results in a bistable state, either a mixture of defection, destruction, and construction [observed in the simplex (D, DA, CA)] or destruction only [within the simplex (C, D, DA)] in the same figure. This mutant defection has no impact on the blend of destruction and construction, maintaining a quadstable state that encompasses cooperation, the coexistence of defection, destruction, and construction, or a combination of destruction and construction, or destruction alone.

## IV. DISCUSSION

In this paper, we have demonstrated that contrary to their role in facilitating cooperation within public goods games, the introduction of destructive agents into pairwise social dilemma games fails to encourage cooperation. Specifically, destructive agents destabilize the system. This leads to a shift from equilibria of single defection to destruction, even single cooperation, or mixed states to regions of bistability. In the prisoner's dilemma and stag-hunt game, we observe the replacement of defection to destruction; on the other hand, single cooperation in harmony and a mixed state in chicken can change to bistable with destruction.

Additionally, we introduced. a different agent type akin to destructive agents: constructive agents. These agents exit the game upon receiving a benefit, yet they also endow their opponents with additional benefits. Our findings suggest that when constructive agents secure higher payoffs than those they bestow on opponents, they can destabilize defection in the prisoner's dilemma and stag-hunt games and disrupt cooperation in the chicken and harmony games, mirroring the

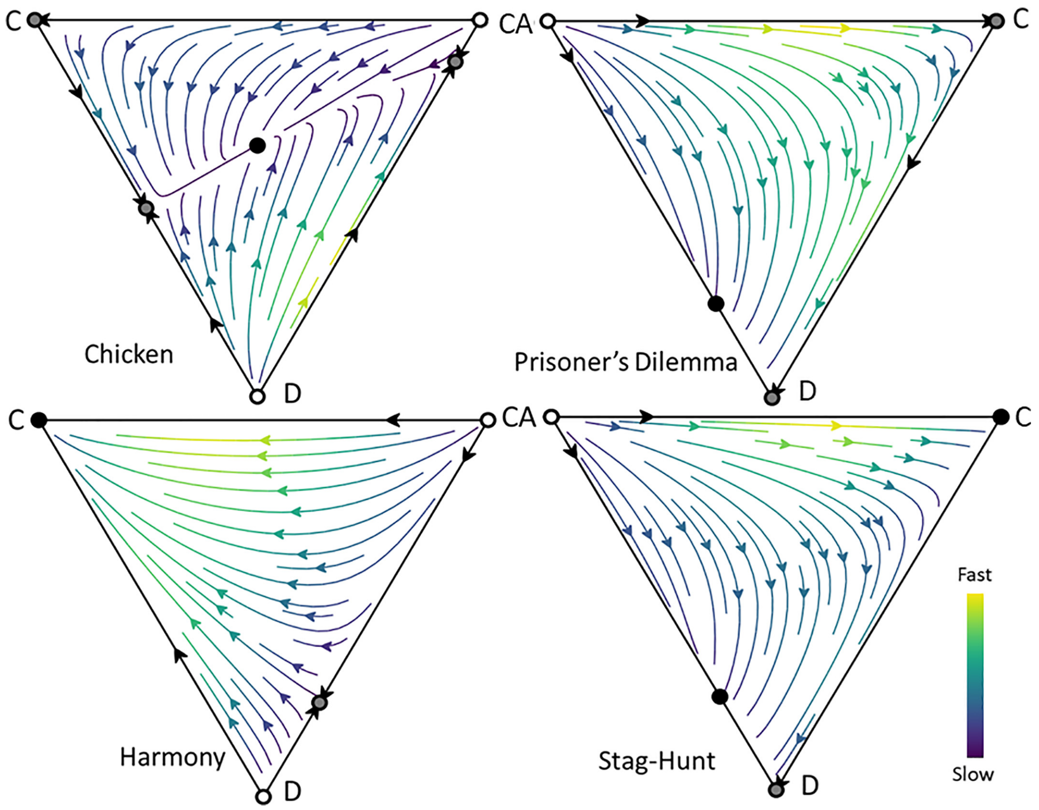

FIG. 5. Coexistence of construction with cooperation and defection in chicken game and disruption of defection states by a mixture of defection and construction in prisoner's dilemma and stag-hunt; no influence in harmony's cooperation. The parameters are fixed at $d = 0.1$ and $d_2 = 0.4$, ($D_g = D_r = 0.5$; PD), ($D_g = 0.5, D_r = -0.5$; CH), ($D_g = -0.5, D_r = -0.5$; H), and ($D_g = -0.5, D_r = 0.5$; SH). Stable nodes are marked with solid black dots, unstable nodes with white dots, and saddle points with gray dots.

destabilizing influence of destructive agents. However, if the payoff for constructive agents is less than what they provide to their opponents, they predominantly disrupt defection states. This leads to new equilibria where defection coexists with constructive actions in the prisoner's dilemma, and a bistable state between mixed defection and construction, and cooperation in the stag-hunt games, leaving the dynamics in the chicken and harmony games unaffected.

Moreover, combining destructive and constructive agents does not inherently promote cooperation, but introduces more complex dynamics, especially when the payoff for constructive agents is lower than what they bestow upon opponents. For instance, in the prisoner's dilemma, a tristable state emerges, characterized by mixed defection, destructive, and constructive agents; a mixed state of destructive and constructive agents; and a state dominated by destructive agents. In the stag-hunt game, a quadstable state arises, featuring mixed states of defection, destruction, and constructive agents; a mixed destructive and constructive agent state; a purely destructive state; and a state of pure cooperation. The harmony game exhibits bistability between pure cooperation and a mixed destructive and constructive agent state. In the chicken game, dynamics are parameter dependent, sometimes resulting in bistability involving a mixed cooperation and defection state, and a mixed destructive and constructive agents state, or leading to a singular mixed state of destructive and constructive agents under different conditions.

The concept of loner strategy, alongside destructive and constructive agents, parallels the notion of social value orientation [47]. In this framework, loners embody individualistic values, seeking personal payoff without impacting their opponents. Destructive agents align with competitive values, aiming to harm their opponents while securing non-negative benefits. Conversely, constructive agents represent prosocial values by benefiting their opponents, while also obtaining non-negative payoffs. While the influence of these strategies on cooperation has been extensively studied, the role of voluntary participation in fostering cooperation remains underexplored. These strategies, being specific, do not encapsulate the broader spectrum of potential behaviors. Beyond these, the social value orientation framework suggests additional motivations for innovative variants of voluntary strategies. These include masochism, where individuals accept negative payoffs by exiting the game without affecting others; martyrdom, which entails negative personal payoffs alongside generating positive outcomes for others; and sadomasochism, characterized by negative personal payoffs coupled with inflicting harm on opponents; among others. Therefore, developing a comprehensive theoretical model that integrates a general voluntary participation strategy, rooted in social value orientations, presents a compelling research direction. This approach aims to investigate how diverse social values impact the evolution of cooperation and assess their effectiveness in enhancing cooperative behaviors. Such an

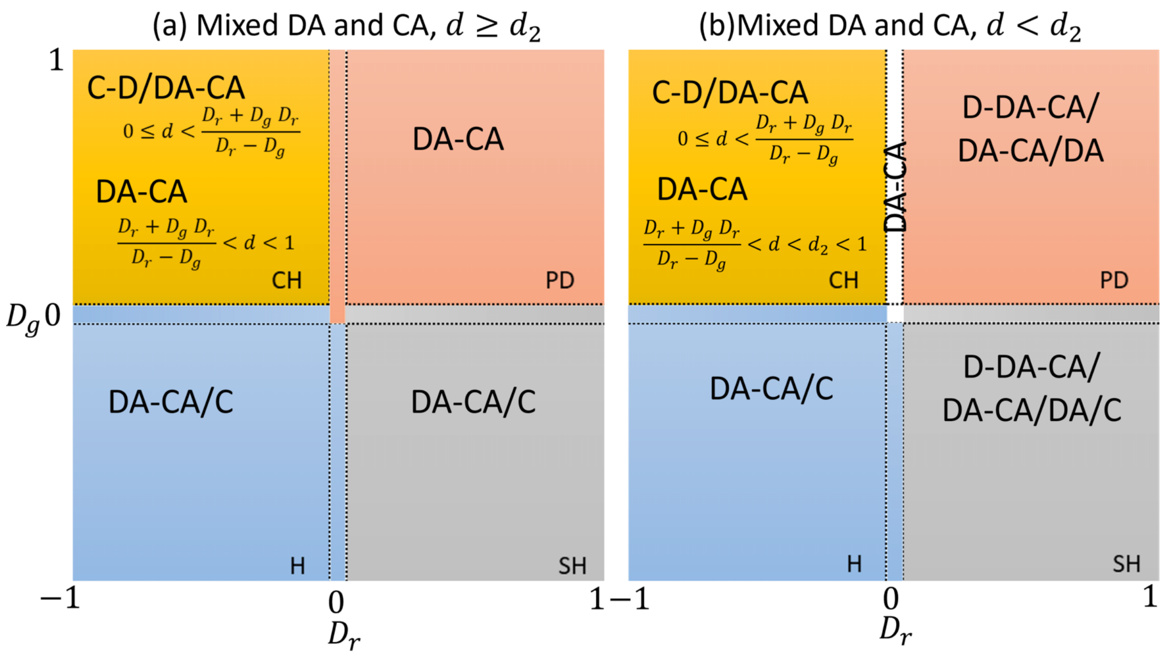

FIG. 6. (a) When the constructive agents' payoff exceeds others, a polymorphic mixture of destruction and construction replaces defection if $D_r > 0$ and disrupts cooperation and a mix of cooperation and defection if $D_r < 0$. The stable equilibrium of prisoner's dilemma and stag-hunt is a coexistence of destruction and construction and a bistable of cooperation and coexistence of destruction and construction. In contrast, chicken's mixed equilibrium becomes either bistable, i.e., either embracing a blend of cooperation and defection or coexistence of destruction and construction, or a monostable coexistence of destruction and construction, and harmony's cooperation demonstrates bistability with the coexistence of destruction and construction. (b) When the constructive agents' payoff is lower than others, defection is changed to either coexistence of defection-destruction-construction or coexistence of destruction-cooperation, or pure destruction if $D_r > 0$, but if $D_r < 0$ stability remains, then it is the same as (a).

endeavor is poised to deepen our understanding of how various voluntary participation strategies can address the enduring puzzle of cooperation.

The critical assumptions of this study—namely, one-shot, anonymous, and well-mixed scenarios—present a most challenging context for the evolution of cooperation. While we found that both constructive and destructive agents do not facilitate cooperation in the context of pairwise social dilemma games, the investigation of the impact of these agents warrants further exploration, as realistic situations often involve repeated interactions or some prior information. It is of significant interest to investigate the impact of these agents on cooperation dynamics in scenarios involving repeated interactions [48], networked populations [49,50], higher-order interactions [51], and other scenarios [52,53].

The programs for theoretical analysis and image generation are given in Ref. [54].

## ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (Grant No. 11931015). We also acknowledge support from 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 23H03499) to C.S.; the National Natural Science Foundation of China (Grants No. 11931015, No. 12271471, and No. 11671348), and the major Program of National Fund of Philosophy and Social Science of China (Grants No. 22&ZD158 and No. 22VRCO49) to L.S.; a Grant-in-Aid for Scientific Research from JSPS, Japan, KAKENHI (Grants No. JP 20H02314 and No. JP 23H03499) awarded to J.T.; and a Japanese Government (MEXT) scholarship, Japan (Grant No. 222143) awarded to K.K.

C.S. and J.T. conceived the research. K.K. and C.S. performed the analytical analysis. All co-authors discussed the results and wrote the manuscript.

The authors declare no conflict of interest.

## APPENDIX

### 1. Equilibria and stability of destructive agent

Four realistic equilibrium points exist in the presence of destructive agents obtained from the solution of the replicator dynamics given by Eq. (2): $E_{A1} = (1, 0, 0)$, $E_{A2} = (0, 1, 0)$, $E_{A3} = (0, 0, 1)$, and $E_{A4} = (\frac{-D_r}{D_g - D_r}, \frac{D_g}{D_g - D_r}, 0)$.

First we reduce the system of equations into a lower dimension, setting $z = 1 - x - y$ into Eq. (2); then the new set of equations will be

$$\dot{x} = x[(1 - x)(\Pi_C - \Pi_{DA}) - y(\Pi_D - \Pi_{DA})] = f_C(x, y),$$
$$\dot{y} = y[(1 - y)(\Pi_D - \Pi_{DA}) - x(\Pi_C - \Pi_{DA})] = f_D(x, y). \tag{A1}$$

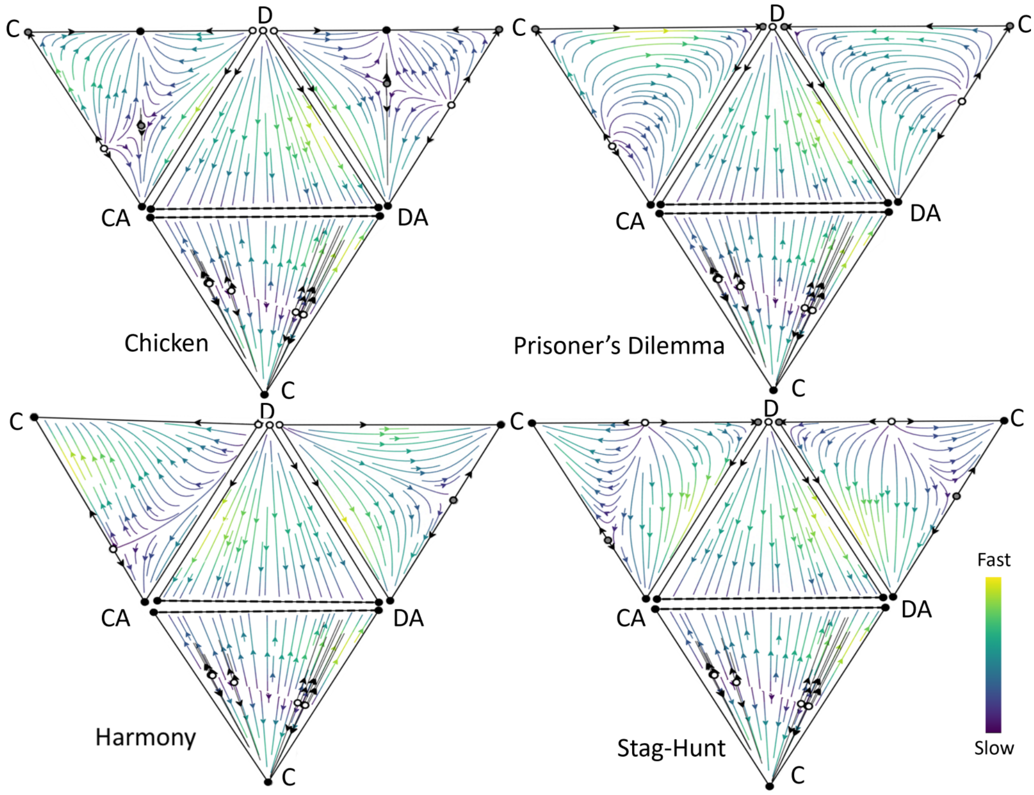

FIG. 7. The mixture of destruction and construction shifts defection in the prisoner's dilemma and stag-hunt and destabilizes cooperation and coexistent cooperation and defection in harmony and chicken game. Four three-simplex combined as a four-simplex; for instance, in prisoner's dilemma simplex (C, DA, CA) is bistable cooperation and mix of destruction and construction. A mutant defection can invade cooperation and leads to a monostable mixture of destruction and construction. The parameters are fixed at $d = 0.4$, $d_1 = 0.1$, $d_2 = 0.1$, ($D_g = D_r = 0.5$; PD), ($D_g = 0.5, D_r = -0.5$; CH), ($D_g = -0.5, D_r = -0.5$; H), and ($D_g = -0.5, D_r = 0.5$; SH). Stable nodes are marked with solid black dots, all points are stable in the thick black dashed line, unstable nodes with white dots, and saddle points with gray dots.

To examine the stability of these equilibrium points, we calculate the eigenvalues of the Jacobin matrix,

$$J_A = \begin{bmatrix} \frac{\partial f_C(x,y)}{\partial x} & \frac{\partial f_C(x,y)}{\partial y} \\ \frac{\partial f_D(x,y)}{\partial x} & \frac{\partial f_D(x,y,)}{\partial y} \end{bmatrix}, \tag{A2}$$

where

$$\begin{aligned}
\frac{\partial f_C(x,y)}{\partial x} &= -\{[-d + (1 + D_g)x - d_1(1 - x - y)]y\} + (1 - x)[-d + x - d_1(1 - x - y) - D_r y] \\
&\quad + x[d + (1 + d)(1 - x) - x + d_1(1 - x - y) - (1 + d_1 + D_g)y + D_r y], \\
\frac{\partial f_C(x,y)}{\partial y} &= x[d + (d_1 - D_r)(1 - x) - (1 + D_g)x + d_1(1 - x - y) - d_1 y], \\
\frac{\partial f_D(x,y)}{\partial x} &= y[d - x - (1 + d_1)x + (1 + d_1 + D_g)(1 - y) + d_1(1 - x - y) + D_r], \\
\frac{\partial f_D(x,y)}{\partial y} &= [-d + (1 + D_g)x - d_1(1 - x - y)](1 - y) + [d - (1 + D_g)x - (d - D_r)x \\
&\quad + d_1(1 - y) + d_1(1 - x - y)]y - x[-d + x - d_1(1 - x - y) - D_r y].
\end{aligned} \tag{A3}$$

For a dynamical system represented by its equilibrium points, stability [55] analysis involves examining the real parts of its eigenvalues. If all eigenvalues possess negative real parts, the equilibrium is deemed stable due to the system's tendency to return to this state over time. Conversely, if any eigenvalue has a positive real part, the equilibrium becomes unstable, indicating divergence from the steady state. When eigenvalues include negative real parts and those with

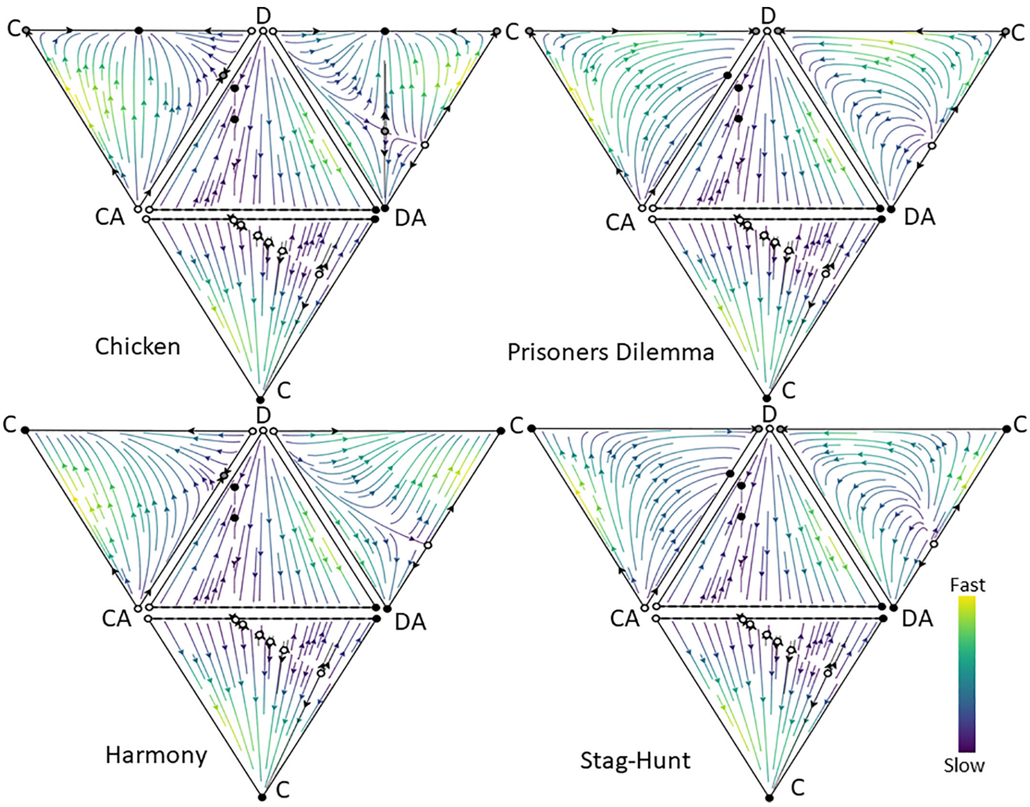

FIG. 8. In the prisoner's dilemma, the monostable defection equilibrium is replaced by either the coexistence of defection-destruction-construction or the coexistence of destruction-cooperation or pure destruction; and in the stag-hunt game, the bistable equilibria of cooperation and defection become tetrastable cooperation or the coexistence of defection-destruction-construction or the coexistence of destruction-cooperation or pure destruction. Destabilization of cooperation and coexistent cooperation and defection also take place in the harmony and chicken game as in the previous one. In the prisoners dilemma, simplex (D, DA, CA) is a tristable coexistence of defection, destruction, and construction, the coexistence of destruction and construction, and destruction. A mutant cooperation cannot change the stability as it is invaded by defection. The parameters are fixed at $d = 0.1$, $d_1 = 0.4$, $d_2 = 0.4$, ($D_g = D_r = 0.5$; PD), ($D_g = 0.5, D_r = -0.5$; CH), ($D_g = -0.5, D_r = -0.5$; H), and ($D_g = -0.5, D_r = 0.5$; SH). Stable nodes are marked with solid black dots, all points are stable in the thick black dashed line, unstable nodes with white dots, and saddle points with gray dots.

real parts equal to zero, necessitating a deeper analysis, applying the center manifold theorem [56] becomes crucial to understanding the system's behavior near that particular point.

### *Stability of the equilibria*

(1) $E_{A1}$: $\lambda_1 = D_g$ and $\lambda_2 = -1 + d$, so the real parts of the eigenvalues will be negative if $d < 1$ and $D_g < 0$. Hence, the equilibrium point $E_{A1}$ is stable if $D_g < 0$. However, at $D_g = 0$, we find a zero eigenvalue; to conclude the stability of this point, we need to use the center manifold theorem here. The Jacobin matrix at $E_{A1}$ is

$$J_{A1} = \begin{bmatrix} -1+d & -1+d \\ 0 & 0 \end{bmatrix}. \tag{A4}$$

An invertible matrix $U$ is constructed by arranging the eigenvectors of the matrix $J_{A1}$ as its column elements, which can diagonalize the matrix,

$$U = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}. \tag{A5}$$

Therefore,

$$U^{-1} J_{A1} U = \begin{bmatrix} -1+d & 0 \\ 0 & 0 \end{bmatrix}. \tag{A6}$$

The new coordinates are Eq. (A7), and Eq. (A1) has been transformed into Eq. (A8),

$$\begin{bmatrix} u \\ v \end{bmatrix} = U^{-1} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x+y \\ y \end{bmatrix}, \tag{A7}$$

$$\begin{aligned} \dot{u} &= (-1+u)[du - d_1(-1+u)u - (u-v)(u - D_r v)], \\ \dot{v} &= -v[d + d_1(-1+u)^2 - du + (u-v)(-1+u-D_r v)]. \end{aligned} \tag{A8}$$

Set $u = u_1 + 1$. Then, Eq. (A8) is converted to a diagonal form given by Eq. (A9),

$$\begin{aligned} \dot{u}_1 &= u_1\big\{ - \big[(1+d_1)u_1^2\big] + d(1+u_1) - (-1+v) \\ &\quad \times (-1 + D_r v) + u_1(-2 - d_1 + v + D_r v)\big\}, \\ \dot{v} &= -v\big[(1+d_1)u_1^2 + D_r(-1+v)v \\ &\quad - u_1(-1 + d + v + D_r v)\big], \end{aligned} \tag{A9}$$

which can be written as

$$\dot{X} = PX + F(X, Y),$$
$$\dot{Y} = QY + G(X, Y). \tag{A10}$$

Here, $X = v$, $Y = u_1$, and $P = 0$, $Q = -1 + d$; $F$ and $G$ are functions of $X$ and $Y$ and $F(0) = G(0) = 0$, $F'(0) = G'(0) = 0$; there exists a $\delta > 0$ and a function $h \in C^r[N_\delta(0)]$, $\forall r \geqslant 1$, so that $h(0) = h'(0) = 0$ defines the local center manifold $\{(X, Y) \in R^2 | u_1 = h(v)$ for $|v| < \delta\}$ and satisfies $h'(v)\{Pv + F[v, h(v)]\} = Qh(v) + G[v, h(v)]$.

Set $u_1 = O(v^2)$. Then we obtain

$$\dot{v} = D_r v^2 + O(v^3). \tag{A11}$$

If $D_r < 0$, the central manifold will be stable at the origin. So we can say that at $D_g \leqslant 0$, and $0 \leqslant d, d_1 \leqslant 1$, $E_{A1}$ will be stable when $D_r < 0$.

(2) $E_{A2}$: $\lambda_1 = d$ and $\lambda_2 = -D_r$, unstable for all $d > 0$ and $D_r < 0$. If $d = 0$, $E_{A2}$ has a zero eigenvalue with a negative eigenvalue for $D_r > 0$. But, for this condition, the central manifold is zero, so we cannot make any conclusion. Hence, using a numerical method, we see that defection is unstable when $D_g > 0$ (prisoner dilemma) and $D_g < 0$ (stag-hunt), illustrated in the right panels of Fig. 2. Therefore, $E_{A2}$ is unstable for all $d \geqslant 0$ and $-1 \leqslant D_g, D_r \leqslant 1$.

(3) $E_{A4}$: $\lambda_1 = \frac{D_r + d*D_g - d*D_r + D_g*D_r}{D_g - D_r}$ and $\lambda_2 = \frac{D_g*D_r}{D_g - D_r}$, will be stable if $D_r < 0$, $D_g > 0$ and $0 \leqslant d \leqslant \frac{-(D_r + D_g D_r)}{D_g - D_r}$.

### 2. Equilibria and stability of constructive agent

There are six realistic equilibrium points in the presence of constructive agents obtained from the solution of replicator dynamics given by Eq. (4): $E_{B1} = (1, 0, 0)$, $E_{B2} = (0, 1, 0)$, $E_{B3} = (0, 0, 1)$, $E_{B4} = (\frac{-D_r}{D_g - D_r}, \frac{D_g}{D_g - D_r}, 0)$, $E_{B5} = (0, \frac{d_2 - d}{d_2}, \frac{d}{d_2})_{d_2 > d}$, and $E_{B6} = (\frac{D_r(d - d_2)}{d_2 D_g + D_r - d_2 D_r + D_g D_r}, \frac{-D_g(d - d_2)}{d_2 D_g + D_r - d_2 D_r + D_g D_r}, \frac{d D_g + D_r - d D_r + D_g D_r}{d_2 D_g + D_r - d_2 D_r + D_g D_r})_{d_2 > d}$.

Similarly, we reduce the system of equations into a lower dimension, setting $w = 1 - x - y$ into Eq. (4). Then the new set of equations will be

$$\dot{x} = x[(1 - x)(\Pi_C - \Pi_A) - y(\Pi_D - \Pi_{CA})] = g_C(x, y),$$
$$\dot{y} = y[(1 - y)(\Pi_D - \Pi_A) - x(\Pi_C - \Pi_{CA})] = g_D(x, y). \tag{A12}$$

To examine the stability of these equilibrium points, we calculate the eigenvalues of the Jacobin matrix:

$$J_B = \begin{bmatrix} \frac{\partial g_C(x,y)}{\partial x} & \frac{\partial g_C(x,y)}{\partial y} \\ \frac{\partial g_D(x,y)}{\partial x} & \frac{\partial g_D(x,y,)}{\partial y} \end{bmatrix}, \tag{A13}$$

where

$$\begin{aligned} \frac{\partial g_C(x, y)}{\partial x} &= -\{[-d + (1 + D_g)x + d_2(1 - x - y)]y\} + (1 - x)[-d + x + d_2(1 - x - y) - D_r y] \\ &\quad + x[d + (1 - d_2)(1 - x) - x - d_2(1 - x - y) - (1 - d_2 + D_g)y + D_r y], \\ \frac{\partial g_C(x, y)}{\partial y} &= x[d + (-d_2 - D_r)(1 - x) - (1 + D_g)x - d_2(1 - x - y) + d_2 y], \\ \frac{\partial g_D(x, y)}{\partial x} &= y[d - x - (1 - d_2)x + (1 - d_2 + D_g)(1 - y) - d_2(1 - x - y) + D_r y], \\ \frac{\partial g_D(x, y)}{\partial y} &= [-d + (1 + D_g)x + d_2(1 - x - y)](1 - y) + [d - (1 + D_g)x - (-d_2 - D_r)x \\ &\quad - d_2(1 - y) - d_2(1 - x - y)]y - x[-d + x + d_2(1 - x - y) - D_r y]. \end{aligned} \tag{A14}$$

#### *Stability of the equilibria*

(1) $E_{B1}$: $\lambda_1 = D_g$ and $\lambda_2 = -1 + d$, so the real parts of the eigenvalues will be negative if $d < 1$ and $D_g < 0$. Hence, the equilibrium point $E_{A1}$ is stable if $D_g < 0$. However, at $D_g = 0$, we find a zero eigenvalue; to conclude the stability of this point, we need to use the center manifold theorem. We can find the transformed system in Eq. (A15) and the center manifold given by Eq. (A16) in the previous way,

$$\begin{aligned} \dot{u} &= u[(-1 + d_2)u^2 + d(1 + u) - (-1 + v)(-1 + D_r v) \\ &\quad + u(-2 + d_2 + v + D_r v)], \\ \dot{v} &= v[(-1 + d_2)u^2 - D_r(-1 + v)v \\ &\quad + u(-1 + d + v + D_r v)], \end{aligned} \tag{A15}$$

$$\dot{v} = D_r v^2 + O(v^3). \tag{A16}$$

The coefficient of $v^2$ will be negative if $D_r < 0$, and the center manifold is stable at the origin. Hence the point $E_{B1}$ is stable when $D_g \leqslant 0$ and $D_r < 0$.

(2) $E_{B2}$: $\lambda_1 = d$ and $\lambda_2 = -D_r$, unstable for all $d > 0$. If $d = 0$, then $E_{B2} = (0, 1, 0)$ has a zero eigenvalue with a negative eigenvalue for $D_r > 0$. In a similar process, we find the following transformed system given by Eq. (A17) and the center manifold given by Eq. (A18):

$$\begin{aligned} \dot{u} &= u[D_g u(1 + u + v) - D_r(1 + u)(1 + u + v) \\ &\quad + v(u + d_2 v)], \\ \dot{v} &= v[(D_g - D_r)u^2 + (1 + D_g - D_r)u(1 + v) \\ &\quad + d_2 v(1 + v)], \end{aligned} \tag{A17}$$

$$\dot{v} = d_2 v^2 + O(v^3), \tag{A18}$$

which is unstable at the origin as the coefficient of $v^2$ is positive for $0 \leqslant d_2 < 1$, so the equilibrium point $E_{B2}$ is unstable when $d \geqslant 0$ and $D_r > 0$.

(3) $E_{B3}$: $\lambda_1 = -d + d_2$ and $\lambda_2 = -d + d_2$, is stable for all $-1 \leqslant D_g, D_r \leqslant 1$ if $d_2 < d$ and unstable otherwise.

(4) $E_{B4}$: $\lambda_1 = \frac{D_r + d*D_g - d*D_r + D_g*D_r}{D_g - D_r}$ and $\lambda_2 = \frac{D_g*D_r}{D_g - D_r}$, will be stable if $D_r < 0$, $D_g > 0$ and $d \leqslant \frac{-D_r(1+D_g)}{D_g - D_r}$.

(5) $E_{B5}$: $\lambda_1 = \frac{-d(d_2-d)}{d_2}$ and $\lambda_2 = \frac{-D_r(d_2-d)}{d_2}$, will be stable if $-1 \leqslant D_g \leqslant 1$, $0 < D_r \leqslant 1$ and $0 \leqslant d < d_2$.

(6) $E_{B6}$: $\lambda_1 = \frac{(-d+d_2)D_gD_r(d_2D_g + D_r - d_2D_r + D_gD_r)}{(-d_2D_g - D_r + d_2D_r - D_gD_r)^2}$ and $\lambda_2 = -\frac{(-d+d_2)(dD_g + D_r - dD_r + D_gD_r)(d_2D_g + D_r - d_2D_r + D_gD_r)}{(-d_2D_g - D_r + d_2D_r - D_gD_r)^2}$, will be stable if $-1 < D_r < 0$, $0 < D_g \leqslant 1$, and $\frac{D_r + D_gD_r}{-D_g + D_r} < d < d_2 < 1$.

### 3. Equilibria and stability of the joint of destructive and constructive agents

In combination with destructive agents and constructive agents, there are seven realistic equilibrium points obtained from the solution of the replicator dynamics given by Eq. (6): $E_{C1} = (0, 0, a, 1-a)_{a\in[0,1]}$, $E_{C2} = (a, 0, \frac{-d+d_2+a-d_2a}{d_1+d_2}, \frac{d+d_1-a-d_1a}{d_1+d_2})_{a\in(0,1)}$, $E_{C3} = (0, 1, 0, 0)$, $E_{C4} = (1, 0, 0, 0)$, $E_{C5} = (\frac{-D_r}{D_g-D_r}, \frac{D_g}{D_g-D_r}, 0, 0)$, $E_{C6} = (0, a, \frac{-d+d_2-d_2a}{d_1+d_2}, \frac{d+d_1-d_1a}{d_1+d_2})_{a\in(0,1)}$, and $E_{C7} = (a, \frac{-D_ga}{D_r}, \frac{-dD_r+d_2D_r+d_2D_ga+D_ra-d_2D_ra+D_gD_ra}{(d_1+d_2)D_r}, \frac{dD_r+d_1D_r+d_1D_ga-D_ra-d_1D_ra-D_gD_ra}{(d_1+d_2)D_r})_{a\in(0,1)}$.

Set $w = 1 - x - y - z$ into Eq. (6). Then the system will be

$$\begin{aligned}
\dot{x} &= x[(1-x)(\Pi_C - \Pi_{CA}) - y(\Pi_D - \Pi_{CA}) - z(\Pi_{DA} - \Pi_{CA})] = h_C(x, y, z),\\
\dot{y} &= y[(1-y)(\Pi_D - \Pi_{CA}) - x(\Pi_C - \Pi_{CA})] = h_D(x, y, z),\\
\dot{z} &= z[-y(\Pi_D - \Pi_{CA}) - x(\Pi_C - \Pi_{CA})] = h_J(x, y, z).
\end{aligned} \tag{A19}$$

The Jacobin matrix is

$$J_C = \begin{bmatrix} \frac{\partial h_C(x,y,z)}{\partial x} & \frac{\partial h_C(x,y,z)}{\partial y} & \frac{\partial h_C(x,y,z)}{\partial z} \\ \frac{\partial h_D(x,y,z)}{\partial x} & \frac{\partial h_D(x,y,z)}{\partial y} & \frac{\partial h_D(x,y,z)}{\partial z} \\ \frac{\partial h_J(x,y,z)}{\partial x} & \frac{\partial h_J(x,y,z)}{\partial y} & \frac{\partial h_J(x,y,z)}{\partial z} \end{bmatrix}, \tag{A20}$$

where

$$\begin{aligned}
\frac{\partial h_C(x,y,z)}{\partial x} &= -y[-d + (1+D_g)x + d_2(1-x-y-z) - d_1z] + (1-x)[-d + x - D_ry + d_2(1-x-y-z) - d_1z]\\
&\quad + x[d + (1-d_2)(1-x) - x - (1-d_2+D_g)y + D_ry - d_2(1-x-y-z) + d_1z],\\
\frac{\partial h_C(x,y,z)}{\partial y} &= x[d + (-d_2 - D_r)(1-x) - (1+D_g)x + d_2y - d_2(1-x-y-z) + d_1z],\\
\frac{\partial h_C(x,y,z)}{\partial z} &= x[(-d_1 - d_2)(1-x) - (-d_1 - d_2)y],\\
\frac{\partial h_D(x,y,z)}{\partial x} &= y[d - x - (1-d_2)x + (1-d_2+D_g)(1-y) + D_ry - d_2(1-x-y-z) + d_1z],\\
\frac{\partial h_D(x,y,z)}{\partial y} &= (1-y)[-d + (1+D_g)x + d_2(1-x-y-z) - d_1z] - x[-d + x - D_ry + d_2(1-x-y-z) - d_1z)],\\
&\quad + y[d - (1+D_g)x - (-d_2 - D_r)x - d - 2(1-y) - d_2(1-x-y-z) + d_1z],\\
\frac{\partial h_D(x,y,z)}{\partial z} &= [-(-d_1 - d_2)x + (-d_1 - d_2)(1-y)]y,\\
\frac{\partial h_J(x,y,z)}{\partial x} &= z[d - x - (1-d_2)x - (1-d_2+D_g)y + D_ry - d_2(1-x-y-z) + d_1z],\\
\frac{\partial h_J(x,y,z)}{\partial y} &= z[d - (1+D_g)x - (-d_2 - D_r)x + d_2y - d_2(1-x-y-z) + d_1z],\\
\frac{\partial h_J(x,y,z)}{\partial z} &= \{-[(-d_1 - d_2)x] - (-d_1 - d_2)y\}z - y[-d + (1+D_g)x + d_2(1-x-y-z) - d_1z] - x[-d + x - D_ry\\
&\quad + d_2(1-x-y-z) - d_1z].
\end{aligned} \tag{A21}$$

***Stability of the equilibria***

(1) At $E_{C1}$: $\lambda_{1,2} = -d + d_2 - a(d_1 + d_2)$ and $\lambda_3 = 0$ are the eigenvalues; the real parts of $\lambda_{1,2} < 0$ for $0 \leqslant d_1 < 1$, if $0 \leqslant d_2 \leqslant d < 1$ and $a > 0$ or if $0 \leqslant d < d_2$ and $a > \frac{-d+d_2}{d_1+d_2}$. Since there is a zero eigenvalue, to conclude we have to use the center manifold theorem here.

The Jacobin matrix at $E_{C1} = (0, 0, a, 1 - a)$ is

$$J_{C1} = \begin{bmatrix} -d - ad_1 + (1-a)d_2, & 0 & 0 \\ 0 & -d - ad_1 + (1-a)d_2 & 0 \\ a[d + ad_1 - (1-a)d_2] & a[d + ad_1 - (1-a)d_2] & 0 \end{bmatrix}. \tag{A22}$$

An invertible matrix $U$ is constructed by arranging the eigenvectors of the matrix $J_{C1}$ as its column elements, which can diagonalize the matrix,

$$U = \begin{bmatrix} -\frac{1}{a} & -1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}. \tag{A23}$$

Therefore,

$$U^{-1} J_{C1} U = \begin{bmatrix} -d + d_2 - a(d_1 + d_2) & 0 & 0 \\ 0 & -d + d_2 - a(d_1 + d_2) & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{A24}$$

The new coordinates are Eq. (A25), and Eq. (A19) has been transformed into Eq. (A26),

$$\begin{bmatrix} u \\ v \\ w_1 \end{bmatrix} = U^{-1} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -a(x+y) \\ y \\ a(x+y)+z \end{bmatrix}, \tag{A25}$$

$$\begin{aligned} \dot{u} &= \frac{1}{a^2}(a+u)\{(-1+d_2)u^2 + a^2(D_g - D_r)v^2 - au[d + d_1 u + v - D_g v + D_r v + d_1 w_1 + d_2(-1 + u + w_1)]\}, \\ \dot{v} &= \frac{1}{a^2} v\{(-1+d_2)u^2 - au[1 + d + D_g + d_1 u + v - D_g v + D_r v + d_1 w_1 + d_2(-2 + u + w_1)] \\ &\quad - a^2[d + d_1 u + v + D_g v - D_g v^2 + D_r v^2 + d_1 w_1 + d_2(-1 + u + w_1)]\}, \\ \dot{w}_1 &= -\frac{1}{a^2}(a - w_1)\{(-1+d_2)u^2 + a^2(D_g - D_r)v^2 - au[d + d_1 u + v - D_g v + D_r v + d_1 w_1 + d_2(-1 + u + w_1)]\}. \end{aligned} \tag{A26}$$

Set $w_1 = w + a$. Then the system given by (A26) in $(u, v, w)$ will be

$$\begin{aligned} \dot{u} &= -\frac{1}{a^2}(a+u)\{-[(-1+d_2)u^2] + a^2[d_1 u + d_2 u + (-D_g + D_r)v^2] \\ &\quad + au[d + d_1 u + v - D_g v + D_r v + d_1 w + d_2(-1 + u + w)]\}, \\ \dot{v} &= \frac{1}{a^2} v\{(-1+d_2)u^2 - au[1 + d + D_g + d_1 u + v - D_g v + D_r v + d_1(a + w) + d_2(-2 + a + u + w)] \\ &\quad - a^2[d + d_1 u + v + D_g v - D_g v^2 + D_r v^2 + d_1(a + w) + d_2(-1 + a + u + w)]\}, \\ \dot{w} &= -\frac{1}{a^2} w\{-[(-1+d_2)u^2] + a^2[d_1 u + d_2 u + (-D_g + D_r)v^2] + au[d + d_1 u + v - D_g v + D_r v + d_1 w + d_2(-1 + u + w)]\}. \end{aligned} \tag{A27}$$

Equation (A27) can be written as

$$\begin{aligned} \dot{X} &= PX + F(X, Y), \\ \dot{Y} &= QY + G(X, Y). \end{aligned} \tag{A28}$$

Here, $X = w$, $Y = [\begin{smallmatrix} u \\ v \end{smallmatrix}]$, and $P = 0$, $Q = [\begin{smallmatrix} -d + d_2 - a(d_1 + d_2) & 0 \\ 0 & -d + d_2 - a(d_1 + d_2) \end{smallmatrix}]$; $F$ and $G$ are functions of $X$ and $Y$ and $F(0) = G(0) = 0, F'(0) = G'(0) = 0$; there exists a $\delta > 0$ and a function $H \in C^r[N_\delta(0)], \forall r \geqslant 1$, so that $H(0) = H'(0) = 0$ defines the local center manifold $\{[X, H(X)] \in R^3 | Y = H(w) \text{ for } |w| < \delta\}$ and satisfies $H'(w)\{Pw + F[w, H(w)]\} = QH(w) + G[w, H(w)]$.

Set $Y = O(w^2)$. We find the following center manifold:

$$\dot{w} = -\tfrac{1}{a}[a(d_1 + d_2) + (d - d_2)]w^3 + O(w^4). \tag{A29}$$

The coefficient of $w^3$ will be negative for either $d > d_2$ or $d < d_2$ for all $0 \leqslant d, d_1, d_2 < 1$, so the center manifold is stable at the origin. Hence, the equilibrium point $E_{C1}$ is stable for all $0 \leqslant d, d_1, d_2 < 1$.

(2) At $E_{C2}$: eigenvalues are $\lambda_1 = 0$, $\lambda_2 = a(1 - d)$, and $\lambda_3 = aD_g$. Here, $\lambda > 0$ for all $0 \leqslant d < 1$, so the equilibrium point is unstable.

(3) $E_{C3}$: eigenvalues are $\lambda_{1,2} = d$ and $\lambda_3 = -D_r$, so $E_{C3}$ is unstable as eigenvalues are positive for $d > 0$.

(4) $E_{C4}$: $\lambda_{1,2} = -1 + d$ and $\lambda_3 = D_g$, the real part of the eigenvalues will be negative if $d < 1$ and $D_g < 0$; if $D_g = 0$, then one eigenvalue will be zero. Using the center manifold theorem, we obtain the following transformed system given by Eq. (A30) and the center manifold is given by Eq. (A31):

$$\begin{aligned}
\dot{u} &= (1+u)[dv + d_1(1+u)v - v^2 + d_2 v(u+v) + vw + D_r vw - D_r w^2],\\
\dot{v} &= (-1+v)[dv + d_2 uv + d_1(1+u)v - v^2 + d_2 v^2 + vw + D_r vw - D_r w^2],\\
\dot{w} &= -w[d + d_1(1+u) - v - dv - d_1(1+u)v + v^2 - d_2(-1+v)(u+v) + w - vw - D_r vw + D_r w^2],
\end{aligned} \tag{A30}$$

$$\dot{w} = -(d + d_1)w + O(w^2). \tag{A31}$$

The center manifold is stable at the origin, which implies $E_{C4}$ is stable when $D_g \leqslant 0$ and $-1 \leqslant D_r \leqslant 1$ for all $0 \leqslant d, d_1, d_2 < 1$.

(5) $E_{C5}$: $\lambda_{1,2} = \frac{D_r + d*D_g - d*D_r + D_g*D_r}{D_g - D_r}$ and $\lambda_3 = \frac{D_g*D_r}{D_g - D_r}$, the real parts of the eigenvalues will be negative if $D_r < 0$, $D_g > 0$ and $d \leqslant \frac{-D_r(1+D_g)}{D_g - D_r}$, so $E_{C5}$ will be stable.

(6) $E_{C6}$: $\lambda_1 = -d + d_2$, $\lambda_2 = -ad$, and $\lambda_3 = -aD_r$ are the eigenvalues; the real parts of $\lambda_2 < 0$ and $\lambda_3 < 0$ if $D_r > 0$. To conclude the stability in a rather analytic way, we rely on the numerical procedure to avoid complexity (see Fig. 8). It is stable if $0 < d < d_2$, $D_r > 0$, and $-1 \leqslant D_g \leqslant 1$.

(7) $E_{C7}$: Coexistent of all strategies, we also rely on the numerical process to conclude this point's stability. It is unstable for all possible values of the parameters.

[1] R. Axelrod and W. D. Hamilton, The evolution of cooperation, Science **211**, 1390 (1981).

[2] E. Fehr and S. Gächter, Cooperation and punishment in public goods experiments, Am. Econ. Rev. **90**, 980 (2000).

[3] C. Darwin, *On the Origin of Species* (Harvard University Press, Cambridge, MA, 1964).

[4] J. W. Weibull, *Evolutionary Game Theory* (MIT Press, Cambridge, MA, 1997).

[5] J. M. Smith, Evolution and the theory of games, in *Did Darwin Get it Right? Essays on Games, Sex and Evolution* (Springer, New York, 1982), pp. 202–215.

[6] U. Fischbacher, S. Gächter, and E. Fehr, Are people conditionally cooperative? Evidence from a public goods experiment, Econ. Lett. **71**, 397 (2001).

[7] M. Archetti and I. Scheuring, Game theory of public goods in one-shot social dilemmas without assortment, J. Theor. Biol. **299**, 9 (2012).

[8] E. Fehr and S. Gächter, Altruistic punishment in humans, Nature (London) **415**, 137 (2002).

[9] J. H. Fowler and N. A. Christakis, Cooperative behavior cascades in human social networks, Proc. Natl. Acad. Sci. USA **107**, 5334 (2010).

[10] H. Ohtsuki and Y. Iwasa, How should we define goodness?-Reputation dynamics in indirect reciprocity, J. Theor. Biol. **231**, 107 (2004).

[11] H. Ohtsuki and Y. Iwasa, The leading eight: Social norms that can maintain cooperation by indirect reciprocity, J. Theor. Biol. **239**, 435 (2006).

[12] H. Gintis, E. A. Smith, and S. Bowles, Costly signaling and cooperation, J. Theor. Biol. **213**, 103 (2001).

[13] J. J. Jordan, M. Hoffman, P. Bloom, and D. G. Rand, Third-party punishment as a costly signal of trustworthiness, Nature (London) **530**, 473 (2016).

[14] R. Axelrod, Effective choice in the prisoner's dilemma, J. Conflict Resolut. **24**, 3 (1980).

[15] D. G. Rand and M. A. Nowak, Human cooperation, Trends Cognit. Sci. **17**, 413 (2013).

[16] J. Andreoni, W. Harbaugh, and L. Vesterlund, The carrot or the stick: Rewards, punishments, and cooperation, Am. Econ. Rev. **93**, 893 (2003).

[17] C. Hilbe and K. Sigmund, Incentives and opportunism: From the carrot to the stick, Proc. R. Soc. B: Biolog. Sci. **277**, 2427 (2010).

[18] Z. Wang, M. Jusup, L. Shi, J.-H. Lee, Y. Iwasa, and S. Boccaletti, Exploiting a cognitive bias promotes cooperation in social dilemma experiments, Nat. Commun. **9**, 2954 (2018).

[19] A. Dreber, D. G. Rand, D. Fudenberg, and M. A. Nowak, Winners don't punish, Nature (London) **452**, 348 (2008).

[20] X. Li, M. Jusup, Z. Wang, H. Li, L. Shi, B. Podobnik, H. E. Stanley, S. Havlin, and S. Boccaletti, Punishment diminishes the benefits of network reciprocity in social dilemma experiments, Proc. Natl. Acad. Sci. USA **115**, 30 (2018).

[21] Z. Wang, M. Jusup, R.-W. Wang, L. Shi, Y. Iwasa, Y. Moreno, and J. Kurths, Onymity promotes cooperation in social dilemma experiments, Sci. Adv. **3**, e1601444 (2017).

[22] T. Sasaki and S. Uchida, The evolution of cooperation by social exclusion, Proc. R. Soc. B: Biolog. Sci. **280**, 20122498 (2013).

[23] S. Li, C. Du, X. Li, C. Shen, and L. Shi, Antisocial peer exclusion does not eliminate the effectiveness of prosocial peer exclusion in structured populations, J. Theor. Biol. **576**, 111665 (2024).

[24] T. A. Han, L. M. Pereira, and T. Lenaerts, Evolution of commitment and level of participation in public goods games, Auton. Agents Multi-Agent Syst. **31**, 561 (2017).

[25] J. Duffy and N. Feltovich, Do actions speak louder than words? An experimental comparison of observation and cheap talk, Games Econ. Behav. **39**, 1 (2002).

[26] C. Hauert, S. De Monte, J. Hofbauer, and K. Sigmund, Volunteering as red queen mechanism for cooperation in public goods games, Science **296**, 1129 (2002).

[27] G. Szabó, C. Hauert, Phase transitions and volunteering in spatial public goods games, Phys. Rev. Lett. **89**, 118101 (2002).

[28] D. G. Rand, J. J. Armao, IV, M. Nakamaru, and H. Ohtsuki, Anti-social punishment can prevent the co-evolution of punishment and cooperation, J. Theor. Biol. **265**, 624 (2010).

[29] D. G. Rand and M. A. Nowak, The evolution of antisocial punishment in optional public goods games, Nat. Commun. **2**, 434 (2011).

[30] C. Hauert, S. De Monte, J. Hofbauer, and K. Sigmund, Replicator dynamics for optional public good games, J. Theor. Biol. **218**, 187 (2002).

[31] C. Hauert and G. Szabó, Game theory and physics, Am. J. Phys. **73**, 405 (2005).

[32] D. Jia, C. Shen, X. Dai, X. Wang, J. Xing, P. Tao, Y. Shi, and Z. Wang, Freedom of choice disrupts cyclic dominance but maintains cooperation in voluntary prisoner's dilemma game, Knowl.-Based Syst. **299**, 111962 (2024).

[33] C. Hauert, A. Traulsen, H. Brandt, M. A. Nowak, and K. Sigmund, Via freedom to coercion: The emergence of costly punishment, Science **316**, 1905 (2007).

[34] R. F. Inglis, J. M. Biernaskie, A. Gardner, and R. Kümmerli, Presence of a loner strain maintains cooperation and diversity in well-mixed bacterial communities, Proc. R. Soc. B: Biolog. Sci. **283**, 20152682 (2016).

[35] H. Pérez-Martínez, C. Gracia-Lazaro, F. Dercole, and Y. Moreno, Cooperation in costly-access environments, New J. Phys. **24**, 083005 (2022).

[36] C. Shen, M. Jusup, L. Shi, Z. Wang, M. Perc, and P. Holme, Exit rights open complex pathways to cooperation, J. R. Soc. Interface **18**, 20200777 (2021).

[37] C. Shen, Z. Song, L. Shi, J. Tanimoto, and Z. Wang, Exit options sustain altruistic punishment and decrease the second-order free-riders, but it is not a panacea, arXiv:2301.04849.

[38] M. Salahshour, Evolution of cooperation in costly institutions exhibits red queen and black queen dynamics in heterogeneous public goods, Communic. Biol. **4**, 1340 (2021).

[39] M. Salahshour, Freedom to choose between public resources promotes cooperation, PLoS Comput. Biol. **17**, e1008703 (2021).

[40] H. Guo, Z. Song, S. Geček, X. Li, M. Jusup, M. Perc, Y. Moreno, S. Boccaletti, and Z. Wang, A novel route to cyclic dominance in voluntary social dilemmas, J. R. Soc. Interface **17**, 20190789 (2020).

[41] S.-Y. Wang, Y.-P. Liu, M.-L. Li, C. Li, and R.-W. Wang, Superrational aspiration induced strategy updating helps resolve the tragedy of the commons in a cooperation system with exit rights, Biosystems **208**, 104496 (2021).

[42] H. Guo, Z. Wang, Z. Song, Y. Yuan, X. Deng, and X. Li, Effect of state transition triggered by reinforcement learning in evolutionary prisoner's dilemma game, Neurocomputing **511**, 187 (2022).

[43] L. Shi, I. Romić, Y. Ma, Z. Wang, B. Podobnik, H. E. Stanley, P. Holme, and M. Jusup, Freedom of choice adds value to public goods, Proc. Natl. Acad. Sci. USA **117**, 17516 (2020).

[44] A. Arenas, J. Camacho, J. A. Cuesta, and R. J. Requejo, The joker effect: Cooperation driven by destructive agents, J. Theor. Biol. **279**, 113 (2011).

[45] R. J. Requejo, J. Camacho, J. A. Cuesta, and A. Arenas, Stability and robustness analysis of cooperation cycles driven by destructive agents in finite populations, Phys. Rev. E **86**, 026105 (2012).

[46] Z. Wang, S. Kokubo, M. Jusup, and J. Tanimoto, Universal scaling for the dilemma strength in evolutionary games, Phys. Life Rev. **14**, 1 (2015).

[47] R. O. Murphy, K. A. Ackermann, and M. J. Handgraaf, Measuring social value orientation, Judg. Decis. Making **6**, 771 (2011).

[48] C. S. L. Rossetti and C. Hilbe, Direct reciprocity among humans, Ethology **130**, e13407 (2024).

[49] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki, Statistical physics of human cooperation, Phys. Rep. **687**, 1 (2017).

[50] Z. Wang, L. Wang, A. Szolnoki, and M. Perc, Evolutionary games on multilayer networks: A colloquium, Eur. Phys. J. B **88**, 124 (2015).

[51] H. Guo, D. Jia, I. Sendiña-Nadal, M. Zhang, Z. Wang, X. Li, K. Alfaro-Bittner, Y. Moreno, and S. Boccaletti, Evolutionary games on simplicial complexes, Chaos, Solitons Fractals **150**, 111103 (2021).

[52] C. Xia, J. Wang, M. Perc, and Z. Wang, Reputation and reciprocity, Phys. Life Rev. **46**, 8 (2023).

[53] H. Guo, Z. Wang, J. Xing, P. Tao, and Y. Shi, Cooperation and coordination in heterogeneous populations with interaction diversity, in *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)* (Auckland, New Zealand,2024), pp. 752–760.

[54] https://osf.io/4p3ch.

[55] J. J. Anagnost and C. A. Desoer, An elementary proof of the Routh-Hurwitz stability criterion, Circuits Syst. Signal Proc. **10**, 101 (1991).

[56] J. Carr, *Applications of Centre Manifold Theory* (Springer Science & Business Media, New York, 2012), Vol. 35.

[57] E. Fernández Domingos, F. C. Santos, and T. Lenaerts, Egttools: Evolutionary game dynamics in PYTHON, iScience **26**, 106419 (2023).