Third-Party Intervention of Cooperation in Multilayer Networks

Hao Guo, Zhao Song, Matjaž Perc, Member, IEEE, Xuelong Li, Fellow, IEEE, and Zhen Wang, Senior Member, IEEE

Abstract—The conflicts in human societies have often been studied through evolutionary games. In social dilemmas, for example, individuals fair best if they defect, but the society is best off if everybody cooperates. Cooperation therefore often requires a mechanism or third parties to evolve and remain viable. To study how third parties affect the evolution of cooperation, we develop a novel game theoretic framework composed of two layers. One layer contains cooperators and defectors, while the other, the third-party layer, contains interveners. Interveners can be peacemakers, troublemakers, or a hybrid of these two. Focusing on two-player two-strategy games, we show that intervention, as an exogenous factor, can stimulate (or inhibit) cooperation by weakening (or strengthening) the dilemma strength of the game the disputant plays. Moreover, the outcome in the disputant layer that is triggered by intervention, in turn, stimulates its own evolution. We analyze the co-evolution of intervention and cooperation and find that even a minority of interveners can promote higher cooperation. By conducting stability analyses, we derive the conditions for the emergence of cooperation and intervention. Our research unveils the potential of third parties to control the evolution of cooperation.

Index Terms—Cooperative systems, decision making, dynamics, game theory, networks.

Manuscript received 15 March 2023; accepted 4 May 2023. Date of publication 3 July 2023; date of current version 17 October 2023. This work was supported in part by the National Science Fund for Distinguished Young Scholars under Grant 62025602; in part by the National Natural Science Foundation of China under Grant U22B2036, Grant 11931015, and Grant 11961138010; in part by New Cornerstone Science Foundation through the XPLORER PRIZE; in part by the Fok Ying-Tong Education Foundation, China, under Grant 171105; and in part by the Key Technology Research and Development Program of Science and Technology-Scientific and Technological Innovation Team of Shaanxi Province under Grant 2020TD-013. This article was recommended by Associate Editor L. C. Rego. (Corresponding author: Zhen Wang.)

Hao Guo and Zhao Song are with the School of Mechanical Engineering and the School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, Xi’an 710072, Shaanxi, China (e-mail: ghffreezing@outlook.com; songzhao@mail.nwpu.edu.cn).

Matjaž Perc is with the Faculty of Natural Sciences and Mathematics, University of Maribor, 2000 Maribor, Slovenia, also with the Department of Medical Research, China Medical University Hospital, China Medical University, Taichung 404332, Taiwan, also with Complexity Science Hub Vienna, 1080 Vienna, Austria, also with Alma Mater Europaea, 2000 Maribor, Slovenia, and also with the Department of Physics, Kyung Hee University, Seoul 02447, Republic of Korea (e-mail: matjaz.perc@gmail.com).

Xuelong Li is with the School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, Xi’an 710072, Shaanxi, China (e-mail: xuelong_li@nwpu.edu.cn).

Zhen Wang is with the School of Mechanical Engineering, the School of Artificial Intelligence, Optics and Electronics, and the School of Cybersecurity, Northwestern Polytechnical University, Xi’an 710072, Shaanxi, China (e-mail: zhenwang0@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSMC.2023.3278048.

Digital Object Identifier 10.1109/TSMC.2023.3278048
life. In the context of government policy, for example, interventions can reduce the negative impact on the environment while encouraging green production [32]. Furthermore, many experimental results suggest that third-party punishment is an important factor in explaining high levels of human cooperation [33], [34], [35]. Theoretical analysis also reveals a cost-effective external intervention for promoting fairness and cooperation in the prisoner’s dilemma game (PDG) and ultimatum game [36], [37]. The optimal incentive that minimizes intervention cost while maximizing the benefit has also been explored in the context of public cooperation [38]. These works consider a single population model and study how to provide intervention in a cost-efficient way. However, they have a little discussion about the emergence of intervention and ignore that third parties are essentially a group and may be risky [39]. A failed intervention may have to bear the consequence of loss, which is a selfish reason that one gives up being an intervener and becomes a silencer. Therefore, it is natural to ask: How to develop a system to study the interplay of intervention and cooperation? How does intervention, as an external factor, control the evolution of cooperation? What is the reason for the emergence of intervention and cooperation?

In addition to examining the unilateral impact of interventions on conflicts [38], [40], it is also necessary to consider how intervention outcomes affect the behavior of third parties. It is believed that parents (or supervisors) can influence how their children perceive and respond to conflicts. By adopting various approaches, parents may either weaken or strengthen their children’s attitudes toward conflict, or a combination of both. Another typical example is when conflicts arise between employees, the employer often acts as a mediator, as the benefits of intervention generally outweigh those of nonintervention. Simultaneously, a company may gain more if it exacerbates conflicts between other companies. This scenario also occurs between countries. Therefore, the underlying rewards gained from intervention are critical to motivating individuals or entities to get involved in the conflict.

We here address these questions by proposing a framework that couples third parties with disputant players to understand how outcome-based interveners affect the evolutionary dynamics in disputant players. Moreover, this framework involves two layers, one is the disputant layer, and the other is the third-party layer. Specifically, players in the disputant layer participate in TPTS games that an exogenous environment can control, i.e., the strategy set of the disputant layer. Meanwhile, the third-party layer contains interveners whose payoff is closely related to the evolutionary outcome in the disputant layer, i.e., the distribution of cooperation and defection. We answer our key research questions by analyzing the interplay between cooperation and intervention using replicator equations for infinitely large populations and Monte Carlo simulations (MCSs) for finitely large square lattices. We provide the condition where cooperation and intervention dominate the respective layer (see Theorem 1). Furthermore, complete cooperation in the disputant layer is not necessary for the dominance of intervention (see Theorems 3 and 6). The emergence of cooperation is influenced by the dilemma strength of the basic game if there is no intervention (see Theorems 2 and 5), or the strength of intervention if intervention exists (see Theorem 6). It is noteworthy that in some cases, a minority of interventions can actually encourage a majority of disputants to engage in cooperation. The simulation results obtained from finitely large square lattices provide more evidence to support our conclusions.

Our main contributions are summarized as follows.

1) We develop a novel evolutionary game theoretical framework to model the coupling effects between strategic conflicts and third-party intervention. This framework overcomes the limitation that only unidirectional relationships are considered in previous studies and allows for analyzing the dynamics of a coupled system.

2) The model enables us to explore the interplay between the intervener’s type and income-preference pattern (IPP). We propose three types of interveners based on their effects on the dilemma strength, including peacemakers, troublemakers, and a hybrid of the two. Particularly, we demonstrate that peacemakers are effective at promoting cooperation. Furthermore, IPPs of intervention are crucial in shaping the coexistence of cooperation and intervention.

3) We show that intervention, by itself, can regulate individual decision making by monitoring strategic conflicts between disputants. The outcome in the disputant layer, which is instigated by intervention, in turn, stimulates its own evolution. This provides a new viewpoint for understanding the source of cooperation and intervention.

4) By analyzing the co-evolution dynamics of cooperation and intervention, we find various equilibria and derive their stability conditions. These include monostable states, such as co-extinction, co-dominance, and coexistence of cooperation and intervention, as well as bistable states under different IPPs. We then expand this system into networks with local interactions and develop an evolutionary game transition algorithm. Our research unveils the potential of third parties to control the evolution of cooperation.

The remainder of this article is organized as follows. In Section II, we give the notations and preliminaries. Section III formulates the system coupling problem of third party and human conflict. In Section IV, we give the model description and theory results of infinitely large well-mixed populations. In Section V, we provide the agent-based model and simulation results of square lattices. Finally, we conclude this article in Section VI.

II. NOTATIONS AND PRELIMINARIES

The notation of this article is summarized as follows. $$S = \{C, D\}$$ and $$A = \{I, Q\}$$ represent the strategy set of players in the disputant layer ($D$) and the third-party layer ($T$), respectively. Denote the payoff matrix of a TPTS game as

$$M = \begin{pmatrix} R & S \n T & P \end{pmatrix}$$

(1)

where mutual cooperation acquires a reward $R$, while mutual defection receives a punishment $P$. A cooperator obtains a sucker’s payoff $S$ if interacting with a defector who obtains
temptation $T$ simultaneously. In detail, the game is a PDG if the parameters satisfy $T > R > P > S$; snowdrift game (SDG) if the parameters satisfy $T > R > S > P$; stag hunt game (SHG) if the parameters satisfy $R > T > P > S$; harmony game (HG) if the parameters satisfy $R > T, S > P$. To measure the strength of social dilemma, Wang et al. [41] rescaled these four parameters as two indicators, named risk-averting and gamble-intending dilemma, defined by $D_r = (P - S/R - P)$ and $D_g = (T - R/R - P)$, respectively. $\pi_\star$ represents the payoff of strategy $\star$, and $P_i$ is the payoff of player $i$. $x$ and $\phi$ represent the fraction of cooperation in the disputant population and intervention in third-party population, respectively. Denote $\dot{x}$ and $\dot{\phi}$ as $x$’s and $\phi$’s derivative with respect to time, respectively. $J$ represents Jacobian in the stability analysis. $W_{S_i\rightarrow S_j}$ is the probability that player $i$ imitates the strategy of $j$.

Denote $G = \{V, E\}$ as a network, where $V = \{1, 2, \ldots, N\}$ represents the node set, $E \subseteq V \times V$ is link set. Let $a_{ij} \in \mathbb{R}$ be the element of adjacent matrix, if the $i$th player has a connection with $j$th player $a_{ij} = 1$; otherwise, $a_{ij} = 0$. Here, we consider an undirected and connected network, thus the degree of each node $k_i = \sum_{j=1}^{N} a_{ij}$. If $k_i = k_j$ $\forall i, j \in V$, $G$ is a homogeneous network. We call $G$ as complete graph if $k_i = N - 1$ $\forall i \in V$. A complete graph with the same weight is also known as a well-mixed population in EGT. In particular, $N \rightarrow \infty$ means an infinitely large well-mixed population.

### III. Problem Formulation

Since many conflict scenarios involve competition between cooperation and defection, we employ TPTS games [41]. In detail, players in the disputant layer have the same opportunity to choose cooperation (C) or defection (D) from set $S$. Meanwhile, as an exogenous factor, players in the third-party layer can choose either intervention (I) or silence (Q) from set $A$. Intervention to mediate conflicts between disputant players is rewarded according to the outcome of the disputants. Therefore, the system we study can be modeled by multilayer networks composed of disputant layer $D = G^D$ and third-party layer $T = G^T$. A sketch is given in Fig. 1, where we take square lattices as an example. Since nodes between two layers are one-to-one, the node sets are identical, and $V^D = V^T$. The edges between nodes in each particular layer of this system can be the same or different. Subsequently, the coupled effect can be depicted by an additional edge between two layers. This multilayer network is similar to an interconnected network where nodes have intraconnections within their own network, and interconnections with the other network [42]. The difference is that nodes between two layers are one-to-one. Specifically, an intervener controls the conflict by intervening in the game that its corresponding disputant play. Evolutionary outcomes related to this disputant then affect the payoff obtained by the intervener.

The basic game ($G_1$) involved in the disputant layer is given by the payoff matrix $M_1$. For simplicity yet without loss of generality, we here set $R = 1$ and $P = 0$ throughout this article [see Fig. 2(a)]

$$M_1 = \begin{pmatrix} 1 & S_1 \\ T_1 & 0 \end{pmatrix}.$$  \hspace{1cm} (2)

As the term dilemma strength is closely associated with the equilibrium of the game [41], we here utilize it to measure the conflicts between disputant players. Two types of
dilemma strength are employed, including \( D_s \) and \( D_e \). The first term \( D_s \) measures a risk-averting dilemma, and the second term \( D_e \) measures a gamble-intending dilemma. The dilemma strength of \( G_1 \) can be controlled by \( S_1 \) and \( T_1 \), namely, \( D_s = -S_1 \) and \( D_e = T_1 - 1 \). Generally, the higher the dilemma strength, the lower the cooperation rate. In our proposed model, players’ gains in the disputant layer depend not only on their own and neighbors’ strategies but also on third parties who act as exogenous environments and can trigger game transitions. Specifically, if supervised by an interner, the player in the disputant layer will participate in another game \( G_2 \) whose payoff matrix is determined as follows:

\[
M_2 = \begin{pmatrix}
1 & S_2 \\
T_2 & 0
\end{pmatrix}
\]

where \( S_2 = S_1 + \delta_1 \) and \( T_2 = T_1 + \delta_2 \). \( -1 \leq \delta_1, \delta_2 \leq 1 \) measure the strength of intervention. Subsequently, the dilemma strength of \( G_2 \) can be given by \( T_1, S_1, \delta_1, \) and \( \delta_2 \), namely, \( D_s = -S_2 = -S_1 - \delta_1 \) and \( D_e = T_2 - T_1 = T_1 + \delta_2 - 1 \). The relationship between dilemma strengths is \( D_s = D_s + \delta_1 \), \( D_e = D_e + \delta_2 \). It is easy to see that the dilemma strength of \( G_2 \) is strengthened (weakened) if \( \delta_1 > 0 \) or \( \delta_2 > 0 \) (\( \delta_1 > 0 \) or \( \delta_2 < 0 \)). A hybrid effect emerges if \( \delta_1 \) and \( \delta_2 \) have the same sign. According to the variation in dilemma strength [see Fig. 2(b)], here we define the type of the third-party interner as a peacemaker (PM) if \( \delta_1 > 0 \) or \( \delta_2 < 0 \) (the dilemma strength of \( G_2 \) is weakened), as a troublemaker (TM) if \( \delta_1 < 0 \) or \( \delta_2 > 0 \) (the dilemma strength of \( G_2 \) is strengthened), and as a mixer (MIX) if \( \delta_1 > 0 \) and \( \delta_2 > 0 \), or \( \delta_1 < 0 \) and \( \delta_2 < 0 \) are satisfied. When it comes to third parties, the payoff of interveners is determined by two key factors: 1) the evolutionary outcome in the disputant layer, which can be reflected by strategy pairs between the corresponding player and its connected neighbors and 2) the IPP, which refers to the payoff received from distinct pairs. In contrast, the payoff of the silence strategy is fixed and does not depend on external conflicts.

IV. INFINITELY LARGE WELL-MIXED POPULATION

A. Coupled Replicator Equation

We first consider infinitely large well-mixed populations where each player has the same probability of interacting with other players in the same population. Due to coupling with third parties, interactions between players in the disputant layer are influenced by the frequency of intervention. Therefore, the expected payoffs of cooperation and defection are given as follows:

\[
\pi_C = \phi(xR + (1 - x)S_2) + (1 - \phi)(xR + (1 - x)S_1)
\]

\[
\pi_D = \phi(xT_2 + (1 - x)P) + (1 - \phi)(xT_1 + (1 - x)P)
\]

where \( x \) and \( 1 - x \) mean the frequency of cooperation and defection in the disputant layer. \( \phi \) and \( 1 - \phi \) represent the fraction of intervention and silence in the third-party layer. The first term of the right-hand side represents the payoff received from \( G_2 \) (under intervention), while the second term means the payoff received from \( G_1 \) (without intervention). As stated above, we denote the payoff of the silence strategy by \( \beta \). While the payoff of intervention is determined by the distribution of pairs between cooperators (CC-pair), pairs between cooperators and defectors (CD-pair), and pairs between defectors (DD-pair). Thus, the expected payoffs of intervention and silence are given as follows:

\[
\pi_I = A_1x^2 + 2A_2x(1 - x) + A_3(1 - x)^2
\]
\[ \pi_Q = \beta \tag{5} \]

where \( A_1, A_2, \) and \( A_3 \) are the gains that interverner obtains from \( CC-, CD- \) and \( DD- \) pairs, respectively. Here, we define the IPP as \( CC \)- pair dominance if \( \max(A_1, A_2, A_3) = A_1 \), \( CD- \) pair dominance if \( \max(A_1, A_2, A_3) = A_2 \), and \( DD- \) pair dominance if \( \max(A_1, A_2, A_3) = A_3 \).

Replicator equation [29], [43], [44], [45] is a powerful tool to describe the evolutionary dynamics of collective behavior. Here, we illustrate the dynamics of this system by the fraction \( \frac{\pi_i}{\pi_Q} \) to describe the evolutionary dynamics of collective behavior.

\[ \dot{x} = x(\pi_C - \tilde{\pi}_1) := f(x, \phi) \]
\[ \dot{\phi} = \phi(\pi_1 - \tilde{\pi}_2) := g(x, \phi) \tag{6} \]

where the dot means the derivative with respect to time, \( \tilde{\pi}_1 \) and \( \tilde{\pi}_2 \) represent the expected payoff of the disputant and third-party layers, respectively. Subsequently, the expected payoff can be calculated by

\[ \tilde{\pi}_1 = x\pi_C + (1 - x)\pi_D \]
\[ \tilde{\pi}_2 = \phi\pi_1 + (1 - \phi)\pi_Q. \tag{7} \]

By considering a mean-field (MF) description, ignoring the spatial topology and stochasticity in evolutionary dynamics, the trajectories of \( x \) and \( \phi \) are determined by the expected payoff of cooperation and intervention, respectively. Note that there is no motivation to choose silence if \( \min(A_1, A_2, A_3) > \beta \), because intervention is a gain-only option and \( g(x, \phi) \geq 0 \) is always true despite of \( x \). On the other hand, there is no motivation to choose intervention if \( \max(A_1, A_2, A_3) < \beta \), because silence is a gain-only option. Improved by [39], we consider a gain-and-loss scenario here to reflect the underlying risk that comes with intervention, namely, \( \max(\pi_1) > \pi_Q \) and \( \min(\pi_1) < \pi_Q \). In the remainder of this section, we first give general results and then discuss three special cases by fixing \( A_1, A_2, \) and \( A_3 \).

**B. Equilibrium and Stability Analysis**

By solving the coupled replicator equation given by (6), we can derive several fixed (or equilibrium) points.

1) \( x = 0 \) and \( \phi = 0 \), i.e., equilibrium \( F_1 = (0, 0) \) which means co-extinction of \( C \) and \( I \).

2) \( x = 1 \) and \( \phi = 0 \), i.e., equilibrium \( F_2 = (1, 0) \) which means a polarized state with complete \( C \) and extinction of \( I \).

3) \( x = 0 \) and \( \phi = 1 \), i.e., equilibrium \( F_3 = (0, 1) \) which means a polarized state with complete \( I \) extinction of \( C \).

4) \( x = 1 \) and \( \phi = 1 \), i.e., equilibrium \( F_4 = (1, 1) \) which means co-dominance of \( C \) and \( I \).

5) \( x = (S_1/[S_1 + T_1 - 1]) \) and \( \phi = 0 \), i.e., equilibrium \( F_5 = (S_1/[S_1 + T_1 - 1], 0) \) which means the existence of \( C \) in the absence of \( I \).

Note that this equilibrium point exists if and only if \( S_1 + T_1 \neq 1 \) and \( 0 < x < 1 \).

6) \( x = ([S_1 + \delta_1]/[S_1 + T_1 + \delta_1 + \delta_2 - 1]) \) and \( \phi = 1 \), i.e., equilibrium \( F_6 = ([S_1 + \delta_1]/[S_1 + T_1 + \delta_1 + \delta_2 - 1], 1) \) which means the existence of \( C \) in the presence of \( I \).

Note that if and only if \( S_1 + T_1 + \delta_1 + \delta_2 \neq 1 \) and \( 0 < x < 1 \), this equilibrium point exists. In addition, there are two interior equilibrium points that depend on the value of \( A_1, A_2, \) and \( A_3 \).

7) \( x^* = ([A_3 - A_2 \pm \sqrt{A_2^2 + \beta(A_1 + A_3 - 2A_2) - A_1A_3}]/[A_1 - 2A_2 + A_3]) \) and \( \phi^* = ([S_1 - (S_1 + T_1 - 1)x^*/(\delta_1 + \delta_2)x^* - \delta_1]), \) i.e., equilibrium \( F_7 \) and \( F_8 \) which mean the co-existence of \( C, D, I, \) and \( Q \).

Note that if and only if \( A_1 - 2A_2 + A_3 \neq 0 \), \( (\delta_1 + \delta_2)x^* - \delta_1 \neq 0 \), \( 0 < x^* < 1 \), and \( 0 < \phi^* < 1 \), these equilibrium points exist. Solution of \( \pi_C - \pi_D = 0 \) and \( \pi_1 - \pi_Q = 0 \) yields the interior equilibrium points.

To determine the stability of each fixed point, we use Lyapunov’s indirect method. By doing so, Jacobian is given as follows:

\[ J = \begin{bmatrix} \frac{\partial f(x, \phi)}{\partial x} & \frac{\partial f(x, \phi)}{\partial \phi} \\ \frac{\partial g(x, \phi)}{\partial x} & \frac{\partial g(x, \phi)}{\partial \phi} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{8} \]

where

\[ a_{11} = ([3\delta_1 + 3\delta_2]x + 3S_1 + 3T_1 - 3x^2 + [(-4\delta_1 - 2\delta_2)x - 4S_1 - 2T_1 + 2]x + \delta_1x + S_1 \\
 a_{12} = x(x - 1)((\delta_1 + \delta_2)x - \delta_1) \\
 a_{21} = -2(\phi - 1)((A_1 - 2A_2 + A_3)x + A_2 - A_3) \\
 a_{22} = -2(\phi - 0.5) \]

\[ \lambda^2 - Tr\lambda + \Delta = 0 \tag{10} \]

where

\[ Tr = a_{11} + a_{22} \]
\[ \Delta = |J| = a_{11}a_{22} - a_{12}a_{21}. \tag{11} \]

Solving the characteristic roots of the characteristic equation yields \( \lambda = ([Tr \pm \sqrt{Tr^2 - 4\Delta}]/2) \). The fixed point is asymptotically stable provided that the real part of all the characteristic roots is less than 0, i.e., \( (Re(\lambda_k) < 0) \). This is equivalent to the condition that the trace of matrix \( J \) is less than 0 and the determinant is greater than 0, that is, \( Tr < 0, \Delta > 0 \). It is worth noting that stability in this part refers to locally asymptotic stability [46]. Then, we showcase the following theorems.

**Theorem 1:** The equilibrium point \((1, 1)\) is the stable state if \( \delta_2 < 1 - T_1 \) and \( A_1 > \beta \).

**Proof:** The trace and determinant of equilibrium point \((1, 1)\) are \( T_2 - 1 - A_1 + \beta \) and \(- (T_2 - 1)(A_1 - \beta) \). When \( \delta_2 < 1 - T_1 \) and \( A_1 > \beta \), the trace and determinant satisfy \( Tr < 0 \) and \( \Delta > 0 \). Thus, equilibrium point \((1, 1)\) is stable.

Based on the parameters established within the proven range, this theorem shows that co-dominance of \( C \) and \( I \) can be achieved. Particularly, this condition has no requirement for the dilemma strength of basic game \( G_1 \) and the payoff from \( CD- \) and \( DD- \) pair. If the strength of intervention is powerful enough \((\delta_2 < 1 - T_1)\) and the payoff
from CC-pair is larger than that from choosing silence, cooperation and intervention can dominate their own layer. As intervention gains from CC-pair, the more cooperation in the disputant layer, the better for the evolution of intervention. This means that mitigating conflict in the disputant layer is beneficial for the evolution of cooperation and the profit of intervention.

**Theorem 2:** The equilibrium point $(1, 0)$ is the stable state if $T_1 < 1$ and $A_1 < \beta$.

**Proof:** The trace and determinant of equilibrium point $(1 , 0)$ are $T_1 - 1 + A_1 - \beta$ and $(T_1 - 1)(A_1 - \beta)$. When $T_1 < 1$ and $A_1 < \beta$, the trace and determinant satisfy $\text{Tr} < 0$ and $\Delta > 0$. Thus, the equilibrium point $(1, 0)$ is stable.

We clarify that in the absence of third-party intervention, cooperation dominates the disputant layer only when $T_1 < 1$. The stability of this point is influenced effectively by the basic game $G_1$. Therefore, if playing PDG and SDG, disputants can never reach a complete cooperation state. On the other hand, in the case of cooperation dominating the disputant layer, intervention vanishes only when $A_1 < \beta$.

**Corollary 1:** Cooperation can dominate in the disputant layer if $1 - T_1 > \min(\delta_2, 0)$.

Incorporating Theorems 1 and 2, we can conclude that the domination of cooperation is fully determined by the value of $T_1 - 1$, $\delta_2$, $A_1$, and $\beta$. If $1 - T_1 > \min(\delta_2, 0)$, cooperation can always dominate either $A_1 > \beta$ or $A_1 < \beta$. This indicates that besides the dilemma strength of $G_1$, intervention strength also plays an important role in deciding the domination of cooperation.

**Theorem 3:** The equilibrium point $(0, 1)$ is the stable state if $\delta_1 < -S_1$ and $A_3 > \beta$.

**Proof:** The trace and determinant of equilibrium state $(0, 1)$ are $S_1 + \delta_1 - A_3 + \beta$ and $-(S_1 + \delta_1)(A_3 - \beta)$. When $\delta_1 < -S_1$ and $A_3 > \beta$, the trace and determinant satisfy $\text{Tr} < 0$ and $\Delta > 0$. Thus, the equilibrium point $(0, 1)$ is stable.

This theorem reveals that in the presence of intervention, cooperation vanishes when $\delta_1 < -S_1$. It means that the stability of this equilibrium point is closely related to intervention strength. On the other hand, the complete intervention state relies on the payoff from DD-pair, i.e., the condition $A_3 > \beta$.

**Theorem 4:** The equilibrium point $(0, 0)$ is the stable state if $S_1 < 0$ and $A_3 < \beta$.

**Proof:** The trace and determinant of equilibrium point $(0, 0)$ are $S_1 + A_3 - \beta$ and $S_1(A_3 - \beta)$. When $S_1 < 0$ and $A_3 < \beta$, the trace and determinant satisfy $\text{Tr} < 0$ and $\Delta > 0$. Thus, the equilibrium point $(0, 0)$ is stable.

We emphasize that if and only if $S_1 < 0$ and $A_3 < \beta$, the co-extinction of $C$ and $I$ occurs. In contrast, this point will not be stable if the basic game that disputants participate in is SDG and HG. Taking Theorem 3 into consideration, the extinction condition of cooperation is determined by $S_1$, $\delta_1$, $A_3$, and $\beta$.

**Corollary 2:** Given cooperation dominates (or is extinct) in the disputant layer, the dominance of intervention in the third-party layer relies on $A_1$ (or $A_3$) and $\beta$.

The previous discussion shows that when cooperation dominates the disputant layer, intervention can dominate the third-party layer if it benefits more from CC-pair than silence, i.e., $A_1 > \beta$. However, when defection dominates the disputant layer, intervention can dominate the third-party layer if it benefits more from DD-pair than silence, i.e., $A_3 > \beta$. In addition to these equilibrium points, we also find two boundary solutions, namely, $F_5$ and $F_6$.

**Theorem 5:** The equilibrium point $([S_1/T_1 + 1], 0)$ is the stable state if $S_1 > 0$, $T_1 > 1$ and $U_1 = (A_3 - \beta)(T_1 - 1)^2 + 2S_1(A_2 - \beta)(T_1 - 1) + S_1^2(A_1 - \beta) < 0$.

**Proof:** The trace and determinant of equilibrium point $([S_1/T_1 + 1], 0)$ are $((1 - S_1(T_1 - 1))(S_1 + T_1 - 1))/(S_1 + T_1 - 1)^2$ and $-(((1 - T_1)S_1U_1)/(S_1 + T_1 - 1)^3)$. When $S_1 > 0$, $T_1 > 1$, and $U_1 < 0$, the trace and determinant satisfy $\text{Tr} < 0$ and $\Delta > 0$. Thus, the equilibrium point $([S_1/T_1 + 1], 0)$ is stable.

We showcase that in the absence of intervention, cooperation is possible, but it must build upon $S_1 > 0$, $T_1 > 1$, and $U_1 > 0$. It means only when disputants play SDG, cooperation may exist in this system.

**Theorem 6:** The equilibrium point $([S_1 + \delta_1]/[S_1 + T_1 + \delta_1 + \delta_2 - 1], 1)$ is the stable state if $\delta_1 > -S_1$, $\delta_2 > -T_1$ and $U_2 = (A_3 - \beta)(T_2 - 1)^2 + 2S_2(A_2 - \beta)(T_2 - 1) + S_2^2(A_1 - \beta) > 0$.

**Proof:** The trace and determinant of equilibrium point $([S_1 + \delta_1]/[S_1 + T_1 + \delta_1 + \delta_2 - 1], 1)$ are $-((U_2 + S_2(T_2 - 1)(S_2 + T_2 - 1))/((S_2 + T_2 - 1)^2))$ and $(S_2)(U_2(T_2 - 1))/(S_2 + T_2 - 1)^3)$. When $\delta_1 > -S_1$, $\delta_2 > -T_1$, and $U_2 > 0$, the trace and determinant satisfy $\text{Tr} < 0$ and $\Delta > 0$. Thus, the equilibrium point $([S_1 + \delta_1]/[S_1 + T_1 + \delta_1 + \delta_2 - 1], 1)$ is stable.

This theorem reveals that under the dominance of intervention, regardless of the basic game $G_1$, if $\delta_1 > -S_1$, $\delta_2 > -T_1$, $U_2 > 0$ are satisfied, cooperation and defection coexist.

**Theorem 7:** The interior equilibrium point $(x^*, \phi^*)$ has an associated Jacobian

$$J^* = \begin{bmatrix} x(1-x)\frac{\partial h_1(x, \phi)}{\partial x} & x(1-x)\frac{\partial h_1(x, \phi)}{\partial \phi} \\ \phi(1-x)\frac{\partial h_2(x)}{\partial x} & 0 \end{bmatrix}$$

where $h_1(x, \phi) = [(\delta_1 - \delta_2)\phi - S_1 - T_1 + 1]x + \delta_1\phi + S_1$ and $h_2(x) = (A_1 - 2A_2 + A_3)x^2 + 2(A_2 - A_3)x + A_3 - \beta$. As $0 < x^* < 1$ and $0 < \phi^* < 1$, it is easy to derive that the trace $\text{Tr} < 0$ and $\Delta > 0$. Thus, if $V_1 > 0$ and $U_3 < 0$.

Thus, far, we have clarified the stability condition of each equilibrium point. In order to study the gain-and-loss scenario, we assume that the silence strategy relies on a fixed payoff $\beta = 2$. Subsequently, denote $A_1 = 4$, $A_2 = 1$, $A_3 = 0$ as CC-pair dominance intervention, where interveners receive the largest payoff $A_1$ from CC-pair. As shown in Fig. 3, compared with the silence strategy, the advantage of intervention changes with the frequency of cooperation (the outcome of the disputant layer). Similarly, denote CD-pair dominance intervention as $A_1 = 0$, $A_2 = 8$, $A_3 = 0$, and DD-pair dominance intervention as $A_1 = 0$, $A_2 = 1$, $A_3 = 4$. Since the valid parameter space of our model is $-1 \leq S_1 \leq 1$, $0 \leq T_1 \leq T_2 \leq 2$, and $0 \leq x, \phi \leq 1$, we only discuss results that satisfy this space. Without a specific statement, we obtain the following results by fixing $T_1 = 1.1$ and $S_1 = -0.1$. 

Authorized licensed use limited to: University of Maribor. Downloaded on October 18, 2023 at 08:02:51 UTC from IEEE Xplore. Restrictions apply.
Fig. 3. Payoff difference of intervention and silence as a function of cooperation rate. For the CC-pair dominance pattern (black line), the payoff difference depends mainly on the fraction of cooperation. The larger the cooperation rate, the higher the payoff difference. DD-pair dominance pattern produces the opposite results (blue line). For the CD-pair dominance, the optimal payoff difference is obtained at $x = 0.5$.

1) CC-Pair Dominance Pattern: In the case of $A_1 = 4$, $A_2 = 1$, and $A_3 = 0$, (5) can be rewritten as

$$\pi_I = 4x^2 + 2x(1 - x)$$

$$\pi_Q = \beta.$$  \hspace{1cm} (13)

There is only one interior equilibrium $F_7 = (\sqrt{\delta_1} - 1/2), [S_1 - (\delta_1 + T_1 - 1)x_1]/([\delta_1 + \delta_2]x_1 - \delta_1)$, where $x_1 = (\sqrt{\delta_1} - 1/2)$. Its stability condition can be obtained according to Theorem 7. Hereafter, in Fig. 4(a), we showcase a phase diagram as a function of $(\delta_1, \delta_2)$ pair. Under these values, it is easy to deduce that equilibrium points $F_2 = (1, 0)$ and $F_3 = (0, 1)$ are unstable. While equilibrium point $F_1 = (0, 0)$ is always stable. For $\delta_1 = \delta_2 = 0$, the disputant layer plays a PDG regardless of the intervention strategy, so $F_1 = (0, 0)$ is the unique asymptotically stable equilibrium.

When $\delta_1 < 0$ and $\delta_2 > 0$, the intervention behaves as TM type, amplifying the dilemma strength in the disputant layer. Consequently, in the upper left part of the diagram, there are still only two available strategies: 1) defection for disputants and 2) silence for third parties. When $\delta_1 > 0$ and $\delta_2 < 0$, intervention manifests as PM type, weakening the dilemma strength of disputant layer. Particularly in $\delta_2 < -0.1$ domain, co-dominance of $C$ and $I$ emerge. Therefore, we showcase a bistable area that contains two stable equilibrium points. As shown in Fig. 4(b), which equilibrium point the system falls is closely related to the initial value of $C$ and $I$. It is worth noting that the interior fix point here is a saddle whose eigenvalues of characteristic function are real roots with opposite signs. From a geometric perspective, there exist orbits approaching and moving away from the saddle point simultaneously. Furthermore, a small region containing equilibrium points $F_1$ and $F_0$ is also triggered by PM type intervention. The result states that compared with TM type, PM type intervention is particularly good at stimulating cooperation.

Turning Attention to MIX Type: When $\delta_1 < 0$ and $\delta_2 < 0$ (lower left region), the game played by disputants under intervention shifts toward the SHG. Since the Nash equilibria of SHG are $(C, C)$ and $(D, D)$, when $\delta_2 < -0.1$, the system undoubtedly enters the bistable state. When $\delta_1 > 0$ and $\delta_2 > 0$ (upper right region), the game played by disputants under intervention shifts toward the SDG. Cooperation can be maintained only when $\delta_1 > 0.1$ and $U_2 > 0$, i.e., $F_1F_6$ bistable region. Otherwise, point $F_1$ will be the unique asymptotically stable equilibrium.

Furthermore, through fixing $\delta_1 = 0.8$, the results in Fig. 4(c) show that intervention with CC-pair dominance pattern keeps the same evolutionary orientation with cooperation. There exist bistable states of $F_1F_4$ and $F_1F_6$, and a monostable state of $F_1$. In particular, discontinuous and continuous phase transitions emerge, respectively, in disputant and third-party layers as $\delta_2$ increases. The results so far demonstrate that cooperation is promoted when intervention emerges in the third-party layer with a larger frequency. It is natural to ask whether a minority of interventions can stimulate significant increases in cooperation. We will address this doubt in the following parts.

2) CD-Pair Dominance Pattern: In the case of $A_1 = 0$, $A_2 = 8$, and $A_3 = 0$, (5) can be rewritten as

$$\pi_I = 16x(x - 1)$$

$$\pi_Q = \beta.$$  \hspace{1cm} (14)

There are two interior equilibrium points $F_7 = ([2 + \sqrt{\delta_2}/4], [S_1 - (\delta_1 + T_1 - 1)x_2]/([\delta_1 + \delta_2]x_2 - \delta_1)$) and $F_8 = ([2 - \sqrt{\delta_2}/4], [S_1 - (\delta_1 + T_1 - 1)x_3]/([\delta_1 + \delta_2]x_3 - \delta_1)$, where $x_2 = (2 + \sqrt{\delta_2}/4)$ and $x_3 = (2 - \sqrt{\delta_2}/4)$. We can get their stability conditions according to Theorem 7. Fig. 5(a) reveals the distribution of asymptotically stable equilibrium points in $\delta_1, \delta_2$ parameter space. As given by Theorems 1–4, $F_1 = (0, 0)$ is always stable, while equilibrium points $F_2, F_3$, and $F_4$ are unstable in these parameter settings. Furthermore, we also find regions where equilibrium points $F_6$ and $F_7$ are stable. In detail, when intervention behaves as PM type, there is a bistable region that satisfies the stability condition for point $F_7$. This indicates partial intervention can promote cooperation in the disputant layer. Here, equilibria are sensitive to the initial frequency of $C$ and $I$. As shown in Fig. 5(b), the system falls into equilibrium point $F_1$ when the initial frequencies of the cooperation and intervention are low, or $F_7$ when the initial conditions are applicable. The result reveals that PM type intervention can still promote cooperation under CD-pair dominance pattern. Another cooperation existence area is MIX type (upper right corner), i.e., the effect of intervention transforms the PDG into an SDG. Since intervention benefits more from the coexistence of cooperation and defection, $F_6$ becomes stable in a large proportion of this region.

Then, we showcase that a minority of interventions can stimulate a majority of cooperation in CD-pair dominance pattern. As shown in Fig. 5(c), cooperation in the disputant layer is higher than intervention when $\delta_2 \leq 0.02$. It answers our concerns about whether cooperation can be triggered by a small fraction of intervention. Furthermore, intervention can
Fig. 4. Phase diagram of cooperation and intervention in $\delta_1$-$\delta_2$ space under CC-pair dominance pattern. (a) Compared with TM type, PM type intervention is more conducive to the evolution of cooperation. For MIX type, when the game is transferred to an SHG under the intervention, the conflict layer is likely to change to a fully cooperative state. When the intervention effect is an SDG, the conflict layer can maintain a state of coexistence of cooperation and defection. (b) Within a bistable region, the initial frequency of cooperation and intervention plays a key role in the equilibrium that the system reaches. Solid and open dots represent stable and other fixed points, respectively. (c) Intervention keeps the same evolutionary orientation with cooperation. Parameters are fixed to $\delta_1 = 0.8$ and $\delta_2 = -0.5$ in panel (b), and $\delta_1 = 0.8$ in panel (c).

Fig. 5. Phase diagram of cooperation and intervention in $\delta_1$-$\delta_2$ space under CD-pair dominance pattern. (a) Cooperation is promoted when intervention behaves as PM and MIX types (right top corner). (b) Within a bistable region, which equilibrium the system falls depends on the initial frequency of cooperation and intervention. (c) Intervention dominates the third-party layer when cooperation and defection are sufficiently mixed. Parameters are fixed to $\delta_1 = 0.95$ and $\delta_2 = 0.02$ in panel (b), and $\delta_1 = 0.95$ in panel (c). Easily become dominant when cooperation and defection are sufficiently mixed if intervention receives more payoff from the CD-pair.

3) DD-Pair Dominance Pattern: In the case of $A_1 = 0$, $A_2 = 1$, and $A_3 = 4$, (5) can be rewritten as

$$\pi_I = 2(1 - x)(2 - x)$$

$$\pi_Q = \beta.$$  

Subsequently, we can derive an interior equilibrium point $F_7 = ([3 - \sqrt{5}/2], [S_1 + (S_1 + T_1 - 1)x_4]/[\delta_1 + \delta_2]x_4 - \delta_1)$, where $x_4 = (3 - \sqrt{5}/2)$. Its stability condition is obtained according to Theorem 7. Given $\delta_2 = 0.5$, we show how asymptotically stable equilibria change with $\delta_1$ in Fig. 6(a). With the increase of $\delta_1$, the asymptotically stable equilibrium moves from $F_3$ to $F_6$ and finally to $F_7$. It reveals that cooperation and intervention remain in opposite evolutionary orientations, i.e., cooperation (intervention) is promoted (prohibited) as the increase of $\delta_1$. In particular, a minority of interventions can stimulate a higher level of cooperation with suitable $\delta_1$. In a monostable state, the equilibrium is insensitive to the initial values [see Fig. 6(b)]. This is further evidenced by assigning the initial values of pair $(x, \phi)$ as $(0.1, 0.1)$, $(0.2, 0.2)$, $\cdots$, $(0.9, 0.9)$ in Fig. 6(c). After finite steps, the evolution of cooperation (top) and intervention (bottom) eventually reaches a unique stable state.

V. EXTENSION TO SQUARE LATTICES

Having seen the highly nontrivial interplay of intervention and cooperation in replicator dynamics (RDs), in this section, we will consider the effect of relaxing the infinitely large well-mixed hypothesis by allowing finitely large populations. In particular, unlike interactions between all actors (well-mixed populations), structures with local interactions
also play a crucial role in strategic conflict [8]. In doing so, there is a widely used updating rule in the literature. That is the Fermi rule [47], where each player imitates one of their opponent’s strategies with a probability given by the Fermi function. Note that RD is determinate, and the variation of the population is linear in the payoff difference. Similar to RD, the Fermi function is also a function of payoff difference. The strategies of players with high payoffs are more likely to spread. Unlike RD, the Fermi function is also a function of payoff difference. The Fermi rule [47], where each player imitates one of their opponent’s strategies with a probability given by the Fermi function 

The strategies of players with high payoffs are more likely to spread. Unlike RD, the Fermi function is also a function of payoff difference. The Fermi rule [47], where each player imitates one of their opponent’s strategies with a probability given by the Fermi function

\[ \text{Algorithm 1: Evolutionary Games With Third-Party Intervention} \]

```
Input: the payoff matrix, the step of MCS \( \Lambda \)
for each \( i \) on the square lattice do
  if \( i \in D \) then
    Initialize player \( i \) with a strategy from set \( S \) randomly;
  else
    Initialize player \( i \) with a strategy from set \( A \) randomly;
  end
  \( m \leftarrow 1; \)
  while \( t < \Lambda \) do
    \( m \leftarrow m + 1; \)
    \( t \leftarrow t + 1; \)
  end
end
```

This obtains payoff via (16) if it belongs to disputant layer, or via (17) if it belongs to third-party layer. Then, we randomly select one of \( i \)'s neighbors, say \( j \), and get its payoff in a similar way. Finally, \( i \) imitates \( j \)'s strategy with a probability determined by the Fermi function

\[ W_{S_i \leftarrow S_j}(P_i, P_j) = \frac{1}{1 + e^{-(P_j - P_i)/K}} \]

where \( K^{-1} \) represents the intensity of selection. Since it has been well studied [47], we parameterize it as 0.1. To ensure

---

Fig. 6. Co-evolution of cooperation and intervention in \( \delta_1 - \delta_2 \) space under DD-pair dominance pattern. (a) Intervention and cooperation maintain opposite evolutionary orientations. (b) Within a monostable region, the interior equilibrium is globally stable regardless of the initial conditions. (c) Equilibria of cooperation (top) and intervention (bottom) are insensitive to initial values. The parameter is fixed to \( \delta_1 = 0.9 \) in panels (b) and (c).
Fig. 7. Cooperation under PM type intervention is more prosperous than that under a TM type. Changing the IPP of intervention can effectively control the equilibrium of this coupled system. From the first to the third column, it shows the frequency of cooperation (top row) and intervention (bottom row) under CC-pair, CD-pair, and DD-pair dominance patterns, respectively. (a1) and (a2) Cooperation and intervention keep the same evolution orientation. (b1) Cooperation reaches optimal value even if only a mere fraction of intervention in the third-party layer. (b2) Intervention dominates the third-party layer when cooperation mixes sufficiently with defection. (c1) Cooperation and intervention evolve in opposite directions. In particular, regardless of the frequency of intervention, cooperation cannot dominate disputant layer. (c2) Intervention dominates the third-party layer when defection prevails. The color code represents the frequency of cooperation and intervention. Parameters are obtained as $T_1 = 1.1$, $S_1 = -0.1$, and $K = 0.1$.

the accuracy of the results, we calculate the average frequency of each strategy over 3000 MCS steps after entering a convergent state. Without the specific declaration, the square lattice consists of $200 \times 200$ players. Furthermore, to avoid finite size effects, we test scales of $100 \times 100$ and $300 \times 300$, with almost identical results.

### B. Phase Diagram

So far, we have revealed the evolutionary dynamics in well-mixed populations with intervention by third parties through MF theory. We are now attempting to explain how spatial structure affects the coupling between cooperation and intervention. To have a comprehensive overview, we provide the phase diagrams of $C$ and $I$ in $\delta_1-\delta_2$ space (see Fig. 7). Consistent with well-mixed populations, the parameter space where cooperation thrives varies with the IPP of intervention. Without the supervision of interveners, players in the disputant layer participate in PDG with $T_1 = 1.1$, $S_1 = -0.1$. Since PDG has been well studied on the square lattice, there is no doubt that cooperation disappears under these parameters. However, this situation will be changed if we consider third-party intervention. It is worth noting that interactions in the disputant layer entirely follow matrix $M_2$ if intervention dominates the third-party layer, whereas it follows matrix $M_1$ if the third-party layer evolves into a full $Q$ state. When the intervening IPP is CC-pair dominance, PM type is more likely to stimulate cooperation than TM type [Fig. 7(a1)]. Turning our attention to MIX type, there is a discontinuous phase transition in the lower left of panels (a1) and (a2), but a continuous phase transition in the upper right corners. Since intervention gains from CC-pair but loses from others, intervention keeps the same evolutionary orientation as cooperation, i.e., when cooperation thrives, intervention thrives; when cooperation declines, intervention declines. Following the CD-pair dominance pattern, interveners spring up as CD-pair increases. Therefore, intervention dominates the third-party layer if cooperation and defection are mixed sufficiently [see Fig. 7(b1) and (b2)]. When the IPP is DD-pair dominance, selecting intervention is better if there exists more DD-pair in the disputant layer [see Fig. 7(c1) and (c2)], revealing a completely opposite evolutionary orientation of $C$ and $I$.

### VI. CONCLUSION AND DISCUSSION

In this article, we develop a novel framework to address the co-evolution of cooperation and third-party intervention. Although evidence has proven that third parties play an inevitable role in the emergence and maintenance of cooperative behavior [48], [49], they have not addressed the emergence of intervention. Different from these studies that consider only one population, we model the interplay between human conflicts and third parties by a coupled system, including disputant and third-party layers. Another difference is that the intervention in this article is risky rather than considering intervening in a cost-effective way [36], [40]. We showcase
seven theorems and implement three special cases by considering gain-and-loss forms, including CC-, CD-, and DD-pair dominance patterns. Furthermore, according to the utility on the dilemma strategy between disputants, we propose three types of interveners: 1) peacemakers; 2) troublemakers; and 3) mixers. Instead of choosing to intervene, players in the third-party layer can also keep silent to avoid the risk of loss. Through the analysis of coupled replicator equations, we show that peacemakers are particularly effective at promoting cooperation. Interestingly, a mere fraction of intervention can stimulate higher cooperation in CD- and DD-pair dominance patterns. On the other hand, complete cooperation is not necessary for complete intervention. Moreover, we find monostable states such as co-extinction, co-dominance, and coexistence of cooperation and intervention, as well as bi-stable states. Then, by developing an evolutionary algorithm in large-scale square lattices, we reproduce the co-extinction, co-dominance, and coexistence of cooperation and intervention. Our research revealed the condition under which intervention emerges and how intervention controls the equilibrium in the conflict. Similar to the environment feedback [46], [50], feedback between third-party and player’s strategies provides the potential for studying the linkage between exogenous intervention and human behavior. Without intervention, cooperators in conflict with defectors are less likely to win the game with a larger dilemma strength. However, strong positive intervention (especially peacemaker) enables cooperation to dominate defection. This unveils the potential of third parties to control the evolution of cooperation.

Under this framework, several attractive avenues for future work still exist. One of the most concerning directions is evaluating the cost efficiency of this kind of risk-bearing intervention. Previous studies have proposed a promising framework for solving the cost-efficiency problem [36], [40]. To do so, we must formulate a scheme that takes into account both the cost of intervention and the benefit of this system, including the increased cooperation rate in the disputant layer and the increased payoff in the third-party layer. Moreover, it is crucial to consider implementing this scheme with a limited number of interveners, rather than relying on global intervention. On the other hand, to evaluate the utility of intervention, we need to consider samaritan interveners who do not change their behavior over time [51]. One possible way is to assign part of third parties as permanent interveners permanently. In light of this, it is natural to expect whether the location of a samaritan intervener (how many cooperators it corresponds to) has a significant impact on the evolution of cooperation [37]. Furthermore, with the framework proposed in this article, there is still room for improvement, such as the time-delay effect [52] in well-mixed populations and time scale [53], [54] in the different layers. Incorporating control theory into cooperative systems is also an exciting research direction [38], [55]. Although cooperation can be effectively promoted by third-party intervention, defection still dominates the network under sufficiently strong temptation. In order to investigate how to facilitate large-scale cooperation, we need to relax the hypotheses further and even organize human experiments.

ACKNOWLEDGMENT

The authors grateful to Shupeng Gao for useful discussions.

REFERENCES
