Journal of Theoretical Biology 355 (2014) 117–127

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Indirect reciprocity in three types of social dilemmas

Mitsuhiro Nakamura*, Hisashi Ohtsuki

Department of Evolutionary Studies of Biosystems, The Graduate University for Advanced Studies (SOKENDAI), Hayama, Kanagawa 240-0193, Japan

AUTHOR-HIGHLIGHTS

- We examine stability of indirect reciprocity in general simultaneous-move games.
- We study Snowdrift (SG), Stag Hunt (SH), and Prisoner's Dilemma (PD) games.
- Strong punishment via bad reputations for defectors is only necessary in SG and PD.
- Punishment for unconditional cooperators stabilizes reciprocation in all the three games.
- Social norms that unfairly favor reciprocators enhance cooperation in SH.

ARTICLE INFO

Article history:
Received 30 October 2013
Received in revised form
19 March 2014
Accepted 24 March 2014
Available online 8 April 2014

Keywords:
Evolutionary game theory
Prisoner's Dilemma game
Reputation
Snowdrift game
Stag Hunt game

ABSTRACT

Indirect reciprocity is a key mechanism for the evolution of human cooperation. Previous studies explored indirect reciprocity in the so-called donation game, a special class of Prisoner's Dilemma (PD) with unilateral decision making. A more general class of social dilemmas includes Snowdrift (SG), Stag Hunt (SH), and PD games, where two players perform actions simultaneously. In these simultaneous-move games, moral assessments need to be more complex; for example, how should we evaluate defection against an ill-reputed, but now cooperative, player? We examined indirect reciprocity in the three social dilemmas and identified twelve successful social norms for moral assessments. These successful norms have different principles in different dilemmas for suppressing cheaters. To suppress defectors, any defection against good players is prohibited in SG and PD, whereas defection against good players may be allowed in SH. To suppress unconditional cooperators, who help anyone and thereby indirectly contribute to jeopardizing indirect reciprocity, we found two mechanisms: indiscrimination between actions toward bad players (feasible in SG and PD) or punishment for cooperation with bad players (effective in any social dilemma). Moreover, we discovered that social norms that unfairly favor reciprocators enhance robustness of cooperation in SH, whereby reciprocators never lose their good reputation.

1. Introduction

In everyday life, your social image influences what you obtain. Helping someone raises your reputation in your community and others help you later when required. This is called indirect reciprocity, a key mechanism for explaining the evolution of cooperative behavior among unrelated individuals (Alexander, 1987; Sugden, 1986; Trivers, 1971). Indirect reciprocity based on reputation has been extensively investigated for decades through numerous theoretical studies (Berger, 2011; Brandt and Sigmund, 2005, 2006; Chalub et al., 2006; Fishman, 2003; Martinez-Vaquero and Cuesta, 2013; Masuda and Nakamura, 2012; Panchanathan, 2011; Panchanathan and Boyd, 2004; Sugden, 1986; Suzuki and Akiyama, 2005, 2007, 2008; and Uchida and Sasaki, 2013) and experimental tests (Bolton et al., 2005; Engelmann and Fischbacher, 2009; Milinski et al., 2001; Pfeiffer et al., 2012; Seinen and Schram, 2006; Sommerfeld et al., 2007; Wedekind and Milinski, 2000; Yoeli et al., 2013). The global success of humans in the past was partially dependent on the establishment of indirect reciprocity, as it was used to explore for more suitable partners for effective economic exchange instead of maintaining closed transactions in inefficient relationships (Greif, 1989; Kandori, 1992).

One important feature of indirect reciprocity is that it endogenously provides an incentive for actors to reward or punish other community members, which is achieved by controlling the actors' reputations that lead to the future rewards or punishments for the actors themselves. We can imagine numerous possibilities of rules to control the reputation of actors who behave differently in various social contexts; such rules are called social norms (Kandori, 1992; Ohtsuki and Iwasa, 2004). Some promising norms can stabilize cooperation in indirect reciprocity, but others cannot. Previous studies have systematically obtained successful social

* Corresponding author.
E-mail address: nakamuramh@soken.ac.jp (M. Nakamura).

http://dx.doi.org/10.1016/j.jtbi.2014.03.035
0022-5193/© 2014 Elsevier Ltd. All rights reserved.

norms in Prisoner's Dilemma scenarios when the reputation information is well-shared in a population (Ohtsuki and Iwasa, 2004, 2006), when it belongs to each individual (Brandt and Sigmund, 2004; Martinez-Vaquero and Cuesta, 2013), with the presence of costly punishment (Ohtsuki et al., 2009), with incomplete reputation information (Nakamura and Masuda, 2011), with multiple reputation states (Tanabe et al., 2013), and with group-level reputations (Masuda, 2012).

Most of the previous studies have investigated social norms for the so-called donation game, a variant of Prisoner's Dilemma with unilateral decision making (Sigmund, 2010). In the donation game, two individuals called donor and recipient participate in and only the donor can decide whether or not to help the recipient, *i.e.*, whether to benefit the recipient by making an investment. Because the donation game focuses on the unilateral behavior of a donor, it ignores many aspects that exist in reality. One such aspect is that the donation game is merely an instance of various social dilemmas. Reputation systems would also play an important role in various simultaneous-move games such as Snowdrift, Stag Hunt, and general Prisoner's Dilemma games. In these games, social norms may depend not only on an actor's choice but also on his/her co-player's choice. For example, how should we define goodness when an actor defects against a bad co-player that unexpectedly cooperates with the actor? Should the actor's defection be justified, even if the co-player shows reformation? Moreover, individuals could infer that a focal player's reputation should be bad when the player received punishment from another player who had established a high reputation. Can such possibility be stable in evolutionary scenarios? To the best of our knowledge, although two previous studies have investigated games other than the donation game, they have not done so exhaustively and not clarified the general characteristics of social norms for the simultaneous-move games (Kandori, 1992; Uchida, 2011).

The present study is directed toward completely exploring reputation systems in simultaneous-move games that comprise more extensive social situations than those in the donation game. We discover that diverse social norms stabilize reciprocation and realize cooperative and stable populations. These successful social norms vary for different types of social dilemmas. To suppress cheating in Prisoner's Dilemma and Snowdrift games, these norms have a common characteristic such that defection against good players is regarded as bad irrespective of the co-player's action. However, in the Stag Hunt game, defection against good players may be allowed, whereas social norms that unfairly favor reciprocators are required to achieve robustness of reciprocation; under these norms, reciprocators never lose their good reputation. It is also imperative to punish unconditional cooperators that help anyone, because they blindly support cheaters (Leimar and Hammerstein, 2001; Panchanathan and Boyd, 2003). There are two mechanisms to restrain unconditional cooperation. One method is to avoid distinguishing between cooperation and defection toward bad players, in which case unconditional cooperators pay an extra cost of helping bad players while reciprocators do not. The other method is to regard cooperation with a bad player as a bad deed, in which case unconditional cooperators are explicitly punished. We discover that the former mechanism is feasible in Prisoner's Dilemma and Snowdrift games, whereas the latter works for all three social dilemmas.

2. Model

We consider a large, well-mixed population in which players from time to time play a symmetric two-player simultaneous-move game. In a one-shot game, two players are sampled from the population in a uniform random manner. Each player selects an action, which is either cooperation (C) or defection (D). There are four possible outcomes of the game for a player: both players select C (the outcome is called reward; R), the focal player selects C and his/her co-player selects D (sucker; S), the focal player selects D and his/her co-player selects C (temptation; T), and both players select D (punishment; P). The payoff matrix of the game is given by

$$\begin{array}{cc} & \begin{array}{cc} \mathrm{C} & \mathrm{D} \end{array} \\ \begin{array}{c} \mathrm{C} \\ \mathrm{D} \end{array} & \begin{bmatrix} 1 & S \\ T & 0 \end{bmatrix} \end{array} \tag{1}$$

where the payoff of the focal player is 1, S, T, or 0 when the outcome is R, S, T, or P, respectively. Fig. 1 illustrates the outcomes of competitions (*e.g.*, replicator dynamics) between cooperators and defectors for the three types of social dilemmas contained in the payoff matrix (1) (Macy and Flache, 2002; Santos et al., 2006; Sigmund, 2010). In a two-dimensional payoff space, the region defined by $T>1>S>0$ yields a Snowdrift game (SG) that has one stable internal equilibrium at which the fraction $S/(S+T-1)$ of players are cooperators and the rest are defectors. The region $T>1>0>S$ yields a Prisoner's Dilemma game (PD) that has a unique stable equilibrium at which defectors dominate the population. It should be noted that the donation game, where the sum of the payoffs of outcomes S (one-sidedly paying cost of helping) and T (one-sidedly enjoying benefit of being helped) is always equal to the payoff of outcome R (both paying cost and enjoying benefit), is projected onto a half-line $S+T=1$ $(T>1)$ in the payoff space (solid red line in Fig. 1); the PD game defined here is more general than the donation game. The region $1>T>0>S$ yields a Stag Hunt game (SH) that has two pure stable equilibria at which cooperators and defectors each dominate the population. Because there is no dilemma when $1>T>0$ and $1>S>0$, we do not study this trivial region.

We employ a binary reputation model in which reputation states are either good (G) or bad (B) (*e.g.*, Nowak and Sigmund, 1998b; see Nowak and Sigmund, 2005; Sigmund, 2010, 2012). In a one-shot game, each of the two players selects an action (*i.e.*, C or D), which is a response to each co-player's reputation (*i.e.*, G or B). A rule that specifies when to use which action is called an action rule, and it is denoted by a. There are four possible action rules. A reciprocator cooperates with a good co-player and defects against a bad co-player, *i.e.*, $a(\mathrm{G})=\mathrm{C}$ and $a(\mathrm{B})=\mathrm{D}$. An unconditional cooperator always cooperates $(a(\mathrm{G})=a(\mathrm{B})=\mathrm{C})$ while an unconditional defector always defects $(a(\mathrm{G})=a(\mathrm{B})=\mathrm{D})$. A 'contrary' player cooperates with a bad co-player and defects against a good co-player $(a(\mathrm{G})=\mathrm{D}$ and $a(\mathrm{B})=\mathrm{C})$. Hereafter, we denote

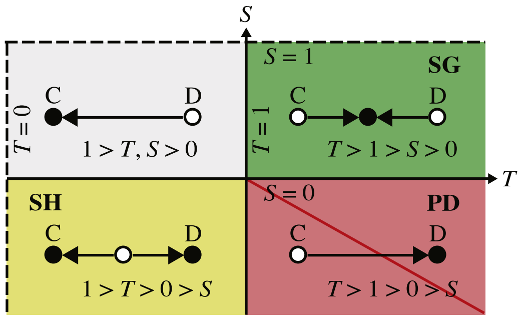

Fig. 1. Three types of social dilemmas. In the payoff space spanned by T and S, the game defined by the payoff matrix (1) is the Snowdrift game (SG) when $T>1>S>0$ (green region), the Prisoner's Dilemma game (PD) when $T>1>0>S$ (red region), and the Stag Hunt game (SH) when $1>T>0>S$ (yellow region). The standard donation game is on the solid red line $(S+T=1$ $(T>1))$. Schematic diagrams inside these regions represent dynamics in competitions between cooperators (C) and defectors (D). Arrows represent the direction of evolution. Solid and hollow circles represent stable and unstable rest points, respectively.