

# The Paradox of Cooperation Benefits

2009

## Abstract

It seems obvious that when benefits of cooperation are increasing, the share of cooperators (if there are any) in the population also increases. It is well documented that positive assortment between cooperative types, for instance, in spatially structured populations, provide better conditions for the evolution of cooperation than complete mixing in most social dilemmas. This study demonstrates, however, that while under most conditions, an increase in the benefits of cooperation also increases the share of cooperators assuming positive assortment, under a specified range of payoff values, when at least two payoff parameters are modified, a reverse relation holds. The conditions for this paradox are determined for two-person social dilemmas: the Prisoner's Dilemma, the Hawks and Doves game, and the Stag Hunt game, assuming global selection and positive assortment.

keywords: altruism, evolution of cooperation, spatially structured social dilemmas, Price equation, Prisoner's Dilemma, Hawks and Doves, Stag Hunt

## 1 Introduction

Cooperation in single-shot two-person social dilemmas is a difficult puzzle that has attracted many theorists in the past. If the interaction is not repeated, there is no place for reciprocity, reputation, image scoring, or other similar mechanisms that support cooperation. Without doubt, there is very limited place for the evolution of cooperation in single-shot social dilemma situations considering complete mixing (random interactions), meaning that individuals are paired with one partner from all possible pairs with a uniform probability. In

social dilemmas, cooperators can be exploited by defectors, which leads to the extinction of cooperation in evolutionary terms.

The exception is the Hawks and Doves game, in which in evolutionary equilibrium, cooperators (doves) establish a share in the population (*Maynard Smith, 1982*). In the Hawks and Doves game (Table 1), increasing the payoffs for mutual cooperation ( $R$ ) will increase the proportion of cooperators in equilibrium (cf. *Maynard Smith, 1982*). A smaller temptation payoff ( $T$ ), a smaller payoff for mutual defection ( $P$ ), and a higher sucker's payoff ( $S$ ) will also increase the proportion of cooperators in evolutionary equilibrium. Hence, the role of cooperation benefits is straightforward: an increase always contributes to a larger share of cooperators in the population. Similarly, larger benefits for defection (hawks) will always decrease the share of unconditional cooperators in the population.

INSERT TABLE 1 HERE

In case cooperators are more likely to meet each other than by pure chance, the benefits of cooperation go also more likely to cooperators. This is a segmentation effect that can make contribution a viable strategy (*Becker, 1976; Axelrod and Hamilton, 1981; Queller, 1985; Bowles and Gintis, 1998; Doebeli and Hauert, 2005*). For instance, cooperators are more likely to meet cooperators, if interactions and reproduction takes place in a spatial structure or in a social network (*Ohtsuki et al, 2006; Wang et al, 2008*). In most social dilemmas, the introduction of a spatial arrangement modifies equilibrium conditions in favor of cooperation (*Doebeli and Hauert, 2005*). Most in the center of research interest, a spatial structure promotes cooperation in the Prisoner's Dilemma (*Nowak and May, 1992; Nowak and May, 1993; Hubermann and Glance, 1993; Nowak et al, 1994*).

In contrast with the spatial Prisoner's Dilemma, the spatial structure has no positive effect in the Hawks and Doves game if local interaction is coupled with local competition for reproduction (*Doebeli and Hauert, 2005*). Population viscosity, where neighbors compete with each other to occupy nearby spaces for their offspring, unsurprisingly has a drawback for segmented clusters of cooperators. The counteract of positive assortment and local competition is well studied in viscous populations (*Taylor, 1992a; Taylor, 1992b; Wilson et al, 1992; Queller, 1994; van Baalen and Rand, 1998; Doebeli and Hauert, 2005; Grafen and*

*Archetti*, 2008). Despite the high relevance of this problem, a necessary first step to investigate effects of cooperation benefits on the share of cooperators with positive assortment, is to neglect local competition and assume (for the sake of simplicity) selection at the global level.

Hamilton's rule (1964), which has been the basis of one of the most important explanations for the evolution of altruism, can be used to determine the effect of cooperation benefits assuming positive assortment. If  $\alpha$  denotes the frequency to which benefits of altruism accrue to other altruists compared with average population members, and  $b$  denotes the benefits of altruism to the partner, and  $c$  indicates the cost to the altruist, then altruists will increase their share if their inclusive fitness  $\alpha b - c$  is greater than zero. Clearly, an increase in the benefits  $b$  will always make it more likely that this requirement of Hamilton's rule is met, indicating that larger cooperation benefits always increase the share of cooperators in the population when positive assortment is considered.

The general result of *Ohtsuki et al* (2006) shows striking similarity with Hamilton's rule (*Grafen*, 2007). They show that when interaction is not random, but rather determined by social networks, natural selection favors cooperation. The rule of thumb they find is that the benefit of the altruistic act  $b$ , divided by the cost  $c$ , should exceed the average number of neighbors  $k$ , which means  $b/c > k$ . This result holds for all type of graphs, including cycles, spatial lattices, other regular graphs, random and scale-free networks. This result also implies that increasing the benefits ( $b$ ) of cooperation in a given structure will always support the survival of cooperators.

On the other hand, *Németh and Takács* (2007) have demonstrated in a simulation study that assuming spatial interaction and global selection, altruism benefits might have a paradoxical effect on the proportion of altruists in the population. They investigated a knowledge transfer interaction, in which passing the knowledge is costly and an altruistic help cannot be reciprocated because knowledge is dichotomous and cannot be lost. Their results show that altruists gain a share in the population, but this share decreases as the value of knowledge in terms of extended lifetime increases. This study highlights that it is misleading to draw conclusions too early from intuition and from Hamilton's rule, and conclude that an increase in cooperation benefits always contribute to a larger share of contributors in the

population. On the contrary, as a major contribution of this study, with a simple analysis we intend to derive that such paradoxes occur for all types of social dilemma games. With a thorough and systematic analysis of games with positive assortment, we will demonstrate under what conditions one can find paradoxical effects of cooperation benefits. Our findings not only imply that an investment in extending cooperation benefits might backfire, assuming at a medium level of positive assortment, but also that less investment in cooperation benefits can contribute to the spread of cooperation in a population where individuals of the same genotype meet more likely than they would meet by chance.

## 1.1 The use of the Price equation

The starting point of our analysis is the Price equation (*Price*, 1970) that has been used in its simple and its general form to a wide range of evolutionary phenomena (*Frank*, 1995; *van Veelen*, 2005), including also to a rederivation of Hamilton's rule (*Grafen*, 1985; *Queller*, 1985). We will use several simplifications as we consider a haploid population interacting in pairs, where the genetic component of cooperation is at a single locus with two possible values: cooperation and defection. We will use the Price equation first without payoff restrictions to determine under what conditions (1) defection or (2) cooperation is an evolutionary stable strategy (*Maynard Smith and Price*, 1973) and under what conditions there is a (3) mixed equilibrium of defectors and cooperators. Consider two groups: cooperators and defectors, having characteristics  $z_1 = 1$  and  $z_2 = 0$ . As the characteristic values do not change from the parent to the child generation ( $\Delta z_i = 0$ ), we can use the simplified Price equation:

$$w\Delta z = cov(w_i, z_i)$$

where  $z_i$  are the characteristic values of different groups of the population,  $w_i$  are their absolute fitness (per capita number of offspring),  $z$  is the average characteristic value, and  $w$  is the average fitness.

That in our special case can be further transformed to:

$$w\Delta z = z(1 - z)(w_1 - w_2) \tag{1}$$

### 1.1.1 Equilibrium

In equilibrium,  $\Delta z = 0$ , which yields three different solutions:

1.  $z = 0$ . The equilibrium proportion of cooperators is zero.
2.  $z = 1$ . The equilibrium proportion of defectors is zero.
3.  $w_1 = w_2$ . A mixed equilibrium of cooperators and defectors. At this point, the average number of offsprings of cooperators and of defectors are the same, which means that population ratios do not change, hence the equilibrium. This equilibrium only exists if  $0 < z^* < 1$ , where  $z^*$  is the equilibrium proportion of cooperators for the solution  $w_1(z) = w_2(z)$ .

### 1.1.2 Stability

Let us take the partial derivative of equation (1) over  $z$ :

$$\frac{\partial w}{\partial z} \Delta z + w \frac{\partial \Delta z}{\partial z} = \frac{\partial(z - z^2)}{\partial z} (w_1 - w_2) + z(1 - z) \left( \frac{\partial w_1}{\partial z} - \frac{\partial w_2}{\partial z} \right)$$

In equilibrium we can simplify to:

$$\frac{\partial \Delta z}{\partial z} = (1 - 2z) \frac{w_1 - w_2}{w} + \frac{z(1 - z)}{w} \left( \frac{\partial w_1}{\partial z} - \frac{\partial w_2}{\partial z} \right)$$

An equilibrium is stable if

$$\frac{\partial \Delta z}{\partial z} < 0.$$

Now we can take a look at the three equilibria again:

1.  $z = 0$

$$\frac{\partial \Delta z}{\partial z} = \frac{w_1 - w_2}{w} < 0$$

So this equilibrium is stable if  $w_1 < w_2$ . A cooperator cannot penetrate the population because its fitness falls below the fitness of defectors. This means that defection is an evolutionary stable strategy (ESS).

2.  $z = 1$

$$\frac{\partial \Delta z}{\partial z} = -\frac{w_1 - w_2}{w} < 0$$

So this equilibrium is stable if  $w_1 > w_2$ . This means that cooperation is an ESS.

3.  $w_1 = w_2$

$$\frac{\partial \Delta z}{\partial z} = \frac{z(1-z)}{w} \left( \frac{\partial w_1}{\partial z} - \frac{\partial w_2}{\partial z} \right) < 0$$

So this equilibrium is stable if  $\frac{\partial w_1}{\partial z} < \frac{\partial w_2}{\partial z}$ . If the ratio of cooperators increases, their fitness falls below the fitness of defectors, so their ratio decreases back. Similarly, if the ratio of cooperators decreases, their fitness exceeds the fitness of defectors, so their ratio increases back.

## 1.2 ESS in social dilemmas

When looking at replicator dynamics based on reproductive fitness, defection is the evolutionary stable strategy (ESS) in the Prisoner's Dilemma, and it is also an ESS in the Stag Hunt game (*Maynard Smith, 1982; Doebeli and Hauert, 2005*). On the other hand, in the Hawks and Doves game, replicator dynamics converge to a mixed stable equilibrium at which both cooperation and defections strategies are present (*Maynard Smith, 1982*). These textbook results are displayed on Figure 1a-c for the sake of comparison.

INSERT FIGURE 1 HERE

## 2 Populations with positive assortment

### 2.1 General equilibrium conditions

Let us now study populations in which the interaction probability of two individuals of the same genotype might be different from the probability of interacting with an average individual. Let us denote the fitness of individuals with genotype  $G$  interacting with another individual with genotype  $H$  by  $w_{GH}$ . We will denote the level of positive assortment by  $\alpha$ . This is the probability that an average individual interacts with another individual of its own genotype instead of interacting with a randomly selected partner (*Cavalli-Sforza and Feldman, 1981*). Thus  $\alpha = 0$  means random interaction or complete mixing, in which individuals of the same genotype only meet each other as can be expected based on their share in the population, and  $\alpha = 1$  means that individuals meet only members of their own genotype.

The average fitness of cooperators and defectors are given as:

$$w_1 = \alpha w_{CC} + (1 - \alpha)[z w_{CC} + (1 - z)w_{CD}]$$

$$w_2 = \alpha w_{DD} + (1 - \alpha)[z w_{DC} + (1 - z)w_{DD}]$$

Using the classical notations  $P = w_{DD}$ ,  $R = w_{CC}$ ,  $S = w_{CD}$ ,  $T = w_{DC}$ :

$$w_1 = \alpha R + (1 - \alpha)[zR + (1 - z)S]$$

$$w_2 = \alpha P + (1 - \alpha)[zT + (1 - z)P]$$

The three equilibria in this case are:

1.  $z = 0$  is stable if  $\alpha(R - S) < P - S$ .
2.  $z = 1$  is stable if  $\alpha(T - P) > T - R$ .
3.  $z^* = \frac{P - \alpha R - (1 - \alpha)S}{(1 - \alpha)(R + P - S - T)}$  exists if  $0 < z^* < 1$  and is stable if  $R - S < T - P$ .

These comply with *Bergstrom's* (2003) results.

## 2.2 Prisoner's Dilemma with positive assortment

The three equilibria in the Prisoner's Dilemma with positive assortment with

$T > R > P > S$  payoffs are:

1.  $z = 0$  is stable if  $\alpha < \frac{P - S}{R - S}$ .
2.  $z = 1$  is stable if  $\alpha > \frac{T - R}{T - P}$ .
3.  $z^* = \frac{P - \alpha R - (1 - \alpha)S}{(1 - \alpha)(R + P - S - T)}$  exists if  $\frac{P - S}{R - S} \leq \alpha \leq \frac{T - R}{T - P}$  and is stable if  $R - S < T - P$ .  
 $(z^*|_{\alpha = \frac{P - S}{R - S}} = 0; z^*|_{\alpha = \frac{T - R}{T - P}} = 1; \frac{\partial z^*}{\partial \alpha} > 0)$

In the Prisoner's Dilemma with positive assortment, close to complete mixing ( $\alpha < \frac{P - S}{R - S}$ ) all

cooperators die out, and close to perfect assortment ( $\alpha > \frac{T - R}{T - P}$ ) all defectors die out. For

intermediate cases of  $\frac{P - S}{R - S} \leq \alpha \leq \frac{T - R}{T - P}$  we have a stable mixed equilibrium with  $z$

monotonously rising from 0 to 1 as  $\alpha$  is rising from  $\frac{P - S}{R - S}$  to  $\frac{T - R}{T - P}$  (cf. Figure 2).

INSERT FIGURE 2 HERE

It is interesting to note that a homogenous population of cooperators is an evolutionary

stable equilibrium when positive assortment is above a threshold level of  $\frac{T - R}{T - P}$ . This is also

one of the two thresholds that the continuation probability has to exceed in the iterated Prisoner's Dilemma for a trigger strategy Tit for Tat (TFT) being in equilibrium (Axelrod, 1984). The similarity is a bit more than pure resemblance: in both cases, the condition describes the minimum probability that a cooperator has an interaction with another cooperator.

In the single-shot Prisoner's Dilemma we now deal with, in an evolutionary horizon, if  $\frac{P-S}{R-S} > \frac{T-R}{T-P}$ , then the mixed equilibrium is unstable and above the critical value of  $z^*$ , the population evolves to full cooperation, and below the critical value of  $z^*$ , the population evolves to full defection (cf. Figure 3).

INSERT FIGURE 3 HERE

### 2.3 Hawks and Doves with positive assortment

The three equilibria in the Hawks and Doves game with positive assortment with  $T > R > S > P$  payoffs are:

1.  $z = 0$  is never stable.
2.  $z = 1$  is stable if  $\alpha > \frac{T-R}{T-P}$ .
3.  $z^* = \frac{P-\alpha R-(1-\alpha)S}{(1-\alpha)(R+P-S-T)}$  exists if  $\alpha \leq \frac{T-R}{T-P}$  and is always stable.  
 $(0 < z^*|_{\alpha=0} = \frac{S-P}{S-P+T-R} < 1; z^*|_{\alpha=\frac{T-R}{T-P}} = 1; \frac{\partial z^*}{\partial \alpha} > 0)$

In the Hawks and Doves game with positive assortment, close to perfect assortment ( $\alpha > \frac{T-R}{T-P}$ ), all defectors die out; and for  $\alpha \leq \frac{T-R}{T-P}$ , there is a stable mixed equilibrium with  $z$  monotonously rising from  $\frac{S-P}{S-P+T-R}$  to 1 as  $\alpha$  is increasing from 0 to  $\frac{T-R}{T-P}$  (cf. Figure 4).

INSERT FIGURE 4 HERE

### 2.4 Stag Hunt with positive assortment

The three equilibria in the Stag Hunt game with positive assortment with  $R > T > P > S$  payoffs are:

1.  $z = 0$  is stable if  $\alpha < \frac{P-S}{R-S}$ .
2.  $z = 1$  is always stable.

3.  $z^* = \frac{P - \alpha R - (1 - \alpha)S}{(1 - \alpha)(R + P - S - T)}$  exists if  $\alpha \leq \frac{P - S}{R - S}$  and is never stable.

$$(0 < z^*|_{\alpha=0} = \frac{P - S}{P - S + R - T} < 1; z^*|_{\alpha=\frac{P - S}{R - S}} = 0; \frac{\partial z^*}{\partial \alpha} < 0)$$

The mixed equilibrium in the Stag Hunt game with positive assortment exists if  $\alpha \leq \frac{P - S}{R - S}$  but it is not stable. A tiny move below the equilibrium point results in all cooperators dying out. Similarly, a tiny move above the equilibrium point has the consequence that all defectors die out. Besides, if  $\alpha \geq \frac{P - S}{R - S}$  then only full cooperation is an ESS; otherwise both full cooperation and full defection are stable equilibria (cf. Figure 5).

INSERT FIGURE 5 HERE

## 3 The Price of Cooperation

### 3.1 General guidelines of a comparative analysis

In this section, we determine the conditions under which increasing the benefits of cooperation results in a decrease in the proportion of contributors in equilibrium; or decreasing the benefits of cooperation results in an increase in the proportion of contributors in equilibrium in social dilemmas with positive assortment. As the effects of changing a single parameter are trivial, we highlight that paradoxes occur when two (or more) payoffs are modified.

### 3.2 Prisoner's Dilemma with positive assortment

#### 3.2.1 Boundary conditions (Prisoner's Dilemma)

We do not find any surprises when analyzing the effect of changes in single parameter values of the Prisoner's Dilemma with positive assortment on the proportion of cooperators in equilibrium. As one can already directly imply from the payoff matrix (Table 1), an increase in  $R$  or  $S$  are beneficial for cooperation, and an increase in  $T$  or  $P$  are beneficial for defection. Hence, the effects of the change in single payoff parameters on  $z^*$  are trivial. For instance, when the reward for mutual cooperation  $R$  is increased, the proportion of cooperators in the mixed ESS of the Prisoner's Dilemma with positive assortment also increases. Paradoxes might occur, however, when at least two parameter values, one that

favors cooperation and one that favors defection (for instance,  $R$  and  $T$ ) are modified at the same time. In case  $\Delta T > \Delta R$ , where we intuitively would expect the decrease of cooperators in ESS, the proportion of cooperators actually rises in a certain range of  $\alpha$  (an example is shown in Figure 6).

INSERT FIGURE 6 HERE

One can see from Figure 2 that a mixed equilibrium is ESS in the Prisoner's Dilemma with positive assortment for  $\frac{P-S}{R-S} \leq \alpha \leq \frac{T-R}{T-P}$ . Let us now introduce the notation  $\alpha_L = \frac{P-S}{R-S}$  and  $\alpha_H = \frac{T-R}{T-P}$  for the boundaries. The effects of single payoff parameters on the boundaries are self-explanatory, but help us to determine the conditions under which paradoxes occur. For this, at least two parameters should change at the same time. Table 2 shows all possible pairs of parameter changes which may result in a paradox. In Appendix I, all cases denoted with an asterisk in Table 2 are explored.

INSERT TABLE 2 HERE

Consider the case when the temptation reward  $T$  is increased. This clearly favors defection, because it has an unambiguous impact on  $z^*$  (see previous section) and also on  $\alpha_H$ . Let us also increase at the same time  $R$  that favors cooperation (for the other cases see Appendix I). If  $\Delta T > \Delta R$ , and especially if  $\Delta T$  is larger than  $\Delta R$  with an order of magnitude ( $\Delta T \gg \Delta R$ ), we would intuitively expect that the proportion of cooperators in mixed equilibrium is decreasing. This is, however, not the case for all values of  $\alpha$ .

The explanation is that if a mixed equilibrium exists, the increase in  $T$  has only an impact on the upper bound, but not on the lower bound of the range of mixed equilibrium.<sup>1</sup> On the other hand, a tiny increase in  $R$  already has the consequence that a mixed equilibrium will exist also for  $\alpha$  values lower than the original lower bound of  $\alpha_L$ . This is displayed on Figure 8. We see a paradox in the range between  $\alpha'_L$  and  $\alpha_T$ . Between  $\alpha'_L$  and  $\alpha_L$  the full defection equilibrium has been replaced by a mixed equilibrium and between  $\alpha_L$  and  $\alpha_T$  the proportion of cooperators in mixed equilibrium has increased.<sup>2</sup>

---

<sup>1</sup>Changing  $P$  and  $R$  is a special case, since both parameters have an effect on both boundaries of  $\alpha$ . And yet, as can be seen in Appendix I, this case may also yield a paradox. See Figure 7.

<sup>2</sup>Let's consider two cases, when  $\frac{\Delta T}{\Delta R} \rightarrow \infty$ :

- $\Delta R$  is fixed and  $\Delta T \rightarrow \infty$ :

The range of  $\alpha$  in which a paradox occurs approaches  $[R', R]$  which has a fixed length, but  $\Delta z^*$  approaches

INSERT FIGURE 7 HERE

INSERT FIGURE 8 HERE

### 3.2.2 The paradox of cooperation benefits in the Prisoner's Dilemma with positive assortment

Consider the situation when  $\alpha_L < \alpha_H$  that is  $\frac{P-S}{R-S} < \frac{T-R}{T-P}$ . This means that a mixed equilibrium exists in the Prisoner's Dilemma game with positive assortment (see Figure 2). When the temptation reward  $T$  is increased (to  $T'$ ) and  $R$  or  $S$  are increased (to  $R'$  and  $S'$ ) or  $P$  is decreased (to  $P'$ ), then a paradox occurs in the following range of  $\alpha$ :

$$\alpha'_L = \frac{P' - S'}{R' - S'} < \alpha \leq \frac{P - S}{R - S} = \alpha_L$$

where a mixed equilibrium will be ESS while originally full defection was the only stable equilibrium (see Figure 7). In addition, the proportion of cooperators in the mixed equilibrium increases in the range:

$$\alpha_L = \frac{P - S}{R - S} < \alpha < \alpha_T$$

where  $\alpha_T$  can be obtained from:

$$z^* = \frac{P - \alpha_T R - (1 - \alpha_T)S}{(1 - \alpha_T)(R + P - S - T)} = \frac{P' - \alpha_T R' - (1 - \alpha_T)S'}{(1 - \alpha_T)(R' + P' - S' - T')}$$

that yields:

$$\alpha_T = \frac{\frac{P-S}{R+P-S-T} - \frac{P'-S'}{R'+P'-S'-T'}}{\frac{R-S}{R+P-S-T} - \frac{R'-S'}{R'+P'-S'-T'}} \quad (2)$$

Consider now the situation when  $\alpha_L > \alpha_H$  that is  $\frac{P-S}{R-S} > \frac{T-R}{T-P}$ . This means that there is no mixed equilibrium (see Figure 3). When the temptation reward  $T$  and the cooperation reward  $R$  are both increased and  $\Delta T > \Delta R$ , and especially if  $\Delta T \gg \Delta R$ , then conditions are zero, so the paradox disappears in limit value.

- $\Delta R \rightarrow 0$  and  $\Delta T$  is fixed:

The length of the range of  $\alpha$  in which a paradox occurs approaches zero, so the paradox disappears in limit value.

seemingly more advantageous for defection. Paradoxically, in a certain range of  $\alpha$ , conditions become more favorable for cooperation (see Figures 9 and 10).

INSERT FIGURE 9 HERE

Figure 9 indicates the case, when  $\alpha'_L > \alpha'_H$ . For

$$\alpha'_L = \frac{P-S}{R'-S} < \alpha \leq \frac{P-S}{R-S} = \alpha_L$$

only full cooperation is possible (if the initial  $z$  is greater than the critical  $z^*$  value), while originally full defection was also a stable equilibrium.

For

$$\alpha_T < \alpha < \alpha'_L = \frac{P-S}{R'-S}$$

the critical value of  $z^*$  is decreased, which is clearly favorable for cooperation: if the initial ratio of cooperators is random (marked with an x on Figure 9), a lower  $z^*$  means a higher chance of reaching full cooperation.

INSERT FIGURE 10 HERE

Figure 10 indicates the case, when  $\alpha'_L < \alpha'_H$ , so the new upper and lower bounds are swapped back.

For

$$\alpha_H < \alpha < \alpha'_L$$

full defection is the only stable equilibrium instead of full cooperation and full defection depending on the initial  $z$ . This is favorable for defection.<sup>3</sup>

For

$$\alpha'_H < \alpha < \alpha_L$$

full cooperation is the only stable equilibrium instead of full cooperation and full defection depending on the initial  $z$ . This is favorable for cooperation, thus paradoxical.<sup>4</sup>

---

<sup>3</sup>If  $\alpha'_L < \alpha_H$ , the equilibrium between the two is mixed instead of the original full defection, which is favorable for cooperation, thus paradoxical.

<sup>4</sup>If  $\alpha_L < \alpha'_H$ , the equilibrium between the two is mixed instead of the original full cooperation, which is favorable

For

$$\alpha'_L < \alpha < \alpha'_H$$

the mixed equilibrium  $z^*$  becomes ESS instead of full cooperation and full defection. As  $\alpha$  approaches  $\alpha'_H$ , the change in the parameters gets more favorable for cooperation, but there is no unambiguous interval where the effect of change is paradoxical.

### 3.3 Hawks and Doves with positive assortment

#### 3.3.1 Boundary conditions (Hawks and Doves)

A similar analysis can be carried out for the Hawks and Doves game with positive assortment. We will avoid overlapping discussions. It is easy to see, for instance, that the effects of  $T$ ,  $R$ ,  $P$  and  $S$  on  $z^*$  are the same as in the Prisoner's Dilemma with positive assortment (see Section 3.2.1).

One can see from Figure 4 that a mixed equilibrium is ESS in the Hawks and Doves game with positive assortment for  $\alpha \leq \alpha_H = \frac{T-R}{T-P}$ . The effects of single payoff parameters on this boundary are the same as in the Prisoner's Dilemma.

#### 3.3.2 The paradox of cooperation benefits in the Hawks and Doves game with positive assortment

Let's increase  $T$  and  $R$  (to  $T'$  and  $R'$ ), so that  $\Delta T > \Delta R$ . A paradox occurs in the following range of  $\alpha$ :

$$\alpha'_H = \frac{T' - R'}{T' - P} < \alpha < \frac{T - R}{T - P} = \alpha_H$$

where mixed equilibrium gives place to full cooperation. In addition, the proportion of cooperators in the mixed equilibrium increases in the range:

$$\alpha_T < \alpha < \alpha'_H = \frac{T' - R'}{T' - P}$$

where  $\alpha_T$  is defined in equation (2). See Figure 11.

---

for defection, so there is no paradox.

INSERT FIGURE 11 HERE

### 3.4 Stag Hunt with positive assortment

#### 3.4.1 Boundary conditions (Stag Hunt)

The effects of  $T$ ,  $R$ ,  $P$  and  $S$  on  $z^*$  are the same as in the Prisoner's Dilemma except for their signs. The effects of  $T$ ,  $R$ ,  $P$  and  $S$  on  $\alpha_L$  are the same as in the Prisoner's Dilemma.

#### 3.4.2 The paradox of cooperation benefits in the Stag Hunt game with positive assortment

Let's increase  $T$  and  $R$  (to  $T'$  and  $R'$ ), so that  $\Delta T' > \Delta R$ . A paradox occurs in the following range of  $\alpha$ :

$$\alpha'_L = \frac{P - S}{R' - S} < \alpha \leq \frac{P - S}{R - S} = \alpha_L$$

where only full cooperation will be stable while originally both full cooperation and full defection were stable depending on the initial  $z$ . In addition, the critical value  $z^*$  above which full cooperation can develop decreases in the range:

$$\alpha_T < \alpha < \alpha'_L = \frac{P - S}{R' - S}$$

where  $\alpha_T$  is defined in equation (2). See Figure 12.

INSERT FIGURE 12 HERE

Finally, note that although we have modified  $T$  and  $R$  in our examples, other changes in the payoff parameters also induce similar paradoxes. In short, for a paradox to occur it is necessary that at least two payoff values change at the same time. The pairs of payoffs that at least need to change and the direction of change are indicated in Table 2. More than two changes also result in paradoxes, but these cannot be interpreted in such a straightforward way as our examples. As we determined  $\alpha_T$  in equation (2) in a general way, this could help the reader to derive other paradoxes. Furthermore, we have listed paradoxical results in Appendix I for the non-trivial cases of parameter changes in Table 2.

## 4 Evolution of positive assortment

It has been demonstrated previously by other research that when individuals are able to select their interaction partners, it increases cooperation in the population (partner selection: *Yamagishi et al*, 1994; *Yamagishi and Hayashi*, 1996; or exit: *Schuessler*, 1989; *Vanberg and Congleton*, 1992). Furthermore, when cooperation cannot evolve in networks with high connectivity, an additional mechanism of topological co-evolution ensures the survival of cooperation (*Santos et al*, 2006).

Some studies have also highlighted that humans have a cognitive capacity to guess with a good accuracy who are cheaters or defectors (*Yamagishi et al*, 2003). This trait might have been evolved as a result of remembering cheater characteristics (*Cosmides*, 1989; *Cosmides and Tooby*, 1992) or just as a result of remembering characteristics (either cheaters or cooperators) that are less frequent in the population (*Barclay*, 2008).

A possible extension of our model could be to let the positive assortment parameter  $\alpha$  evolve endogenously. This is equivalent to introducing an evolvable trait that enables individuals to recognize and select their interaction partners with certain accuracy. As *Wilson and Dugatkin* (1997) notes, it is likely that the cognitive prerequisites for assortative interactions are often satisfied. If defectors are able to recognize the type of others, it is reasonable to assume that they do not choose an interaction partner of their own type (cf. *Bergstrom*, 2003). They rather choose cooperators; because this provides them higher payoffs. In this model extension  $\alpha$  denotes the positive assortment of cooperators and  $\beta$  denotes the negative assortment of defectors. We assume that a certain individual is randomly selected and based on the  $\alpha$  ( $\beta$ ) parameter of this individual, an interaction partner is chosen. The interaction partner is forced to play, thus her  $\alpha$  ( $\beta$ ) parameter does not influence whether the interaction takes place or not.

The average fitness of cooperators and of defectors can be expressed as:

$$w_C = z \frac{[\alpha + (1 - \alpha)z]2R + (1 - \alpha)(1 - z)S}{z} + (1 - z) \frac{[\beta + (1 - \beta)z]S}{z}$$

$$w_D = z \frac{(1 - \alpha)(1 - z)T}{1 - z} + (1 - z) \frac{[\beta + (1 - \beta)z]T + (1 - \beta)(1 - z)2P}{1 - z}$$

Unsurprisingly, evolution selects for  $\alpha = 1$  ( $\frac{\partial w_C}{\partial \alpha} > 0$ ). The ESS value of  $\beta$  depends on  $T$  and

$P \left( \frac{\partial w_D}{\partial \beta} = (1-z)(T-2P) \right)$ . If  $T > 2P$ , evolution selects for  $\beta = 1$  (defectors prefer exploiting cooperators by selecting them). If  $T < 2P$ , evolution selects for  $\beta = 0$  (defectors prefer interacting with themselves by selecting randomly). In the latter case, however,  $w_C > w_D$  for every  $z$ , so  $z^* = 1$  is the evolutionary stable equilibrium. In the former case ( $T > 2P, \beta = 1$ ), there are more subcases:

- if  $S = 0$ 
  - if  $2R > T$ , then  $z^* = 1$
  - if  $2R < T$ , then  $z^* = 0$
  - if  $2R = T$ , then  $z^*$  does not exist
- if  $S > 0$ 
  - if  $2R \geq T$ , then  $z^* = 1$
  - if  $2R < T$ , then  $z^* = \frac{S}{T+S-2R}$

If  $S > 0$  and  $T > 2R$ , then  $z^* = \frac{S}{T+S-2R}$ . If we increase  $T$  and  $R$  simultaneously, and  $\Delta R < \Delta T < 2\Delta R$ , then the ratio of cooperators increases, which is a paradox.

## 5 Discussion

This study has demonstrated that raising the rewards of cooperation might play against the success of cooperative behavior in populations where cooperators are more likely to meet other cooperators than by chance. We found this new and counter-intuitive result for all social dilemmas: in the Prisoner's Dilemma, in the Hawks and Doves game, as well as in the Stag Hunt game.

It is important to note, however, that there is no paradox if only one payoff parameter is modified. Increasing purely the rewards for mutual cooperation ( $R$ ), for instance, always provides improved conditions for the survival of cooperation. To obtain a paradoxical result, it is necessary that at least two payoff parameters are changed. We have displayed, for instance, the case when both the temptation reward ( $T$ ) and the reward for mutual cooperation ( $R$ ) have been increased in the Prisoner's Dilemma with positive assortment such that  $\Delta T \gg \Delta R$ . This is a situation that is favorable for defectors. We have

demonstrated, however, that in this case, there is always a nonzero range of positive assortment when the equilibrium proportion of cooperators increases.

Furthermore, the paradox only occurs at a certain range of positive assortment. This means that at certain probabilities that describe how much more likely two cooperators meet each other than by chance, rewards of cooperation backfire and increase the proportion of defectors in the population. Similarly, it applies only to a limited range of positive assortment that payoffs that favor defection backfire and increase the proportion of cooperators in the population. Typically, when cooperators too often or too rarely meet each other, then increasing the rewards of cooperation will not diminish their chances of survival. In this paper we have shown the exact conditions when the increase of rewards of cooperation and when the increase of rewards of defection contributes to a paradoxical change in the equilibrium proportion of cooperators and defectors in the Prisoner's Dilemma, in the Hawks and Doves game, and in the Stag Hunt game.

We have also clarified why the paradox occurs. A mixed proportion of cooperators and defectors is evolutionary stable in a certain range of positive assortment in all social dilemmas. The boundary conditions of mixed equilibrium unequivocally determine the equilibrium proportion of cooperators within the boundaries. The boundary conditions, however, are not affected by all payoff values. The lower bound is independent from the temptation reward ( $T$ ) and the upper bound is independent from the sucker's reward ( $S$ ). This implies that a change of  $T$  or  $S$  and another parameter will leave place for paradoxical results. Furthermore, paradoxes can also occur when both boundaries change ( $P$  and  $R$  are modified).

The paradox we found in this paper is highly counterintuitive in the light of previous theoretical results on the evolution of cooperation. Hamilton's rule (1964) asserts that altruists will spread in a population if  $ab - c > 0$ . This means that if the benefits of cooperation are increasing, it always benefits altruists (cooperators) and results in their dissemination (see Appendix II). In this paper we found justification for this result and found no paradoxes in a special case of payoff structure that has restrictions on payoffs using a single benefit and cost parameter (see e.g. *Doebeli and Hauert, 2005: Table 1*). We highlighted that the counterintuitive cases have been overlooked previously due to the

simplified representation of social dilemma games. On the contrary, we determined the conditions of paradoxical situations in which the increase in cooperation benefits result in a lower share of cooperators in the population using the four payoff parameters ( $T$ ,  $R$ ,  $P$ ,  $S$ ) of the more general description of social dilemmas (see e.g. *Axelrod*, 1984). Our result might seem narrow, because paradoxes only occur for a certain parameter range, but they provide a general warning for the wide body of research on the evolution of cooperation that parameter restrictions in social dilemmas can result in a loss of important insights. The fact that in most empirical social dilemma situations, we find cooperators and defectors co-existing, underline the relevance of these results. In empirical cases, matching is not random, but is biased towards meeting of the same types (see e.g. *Ohtsuki et al*, 2006). This might be voluntary (homophily) or unconscious, as it is the case in spatially structured populations. This paper has highlighted paradoxical effects of cooperation benefits concerning these empirically highly relevant situations. As empirical situations are always more complex than simple models, it is difficult to justify that such paradoxical effects frequently occur in nature. Although there might be other explanations, there are some documented cases that are at least partly in line with our theoretical findings.

Human societies vary in their level of assortativity, but interactions rarely occur randomly. In societies with lower degree of market integration, less cooperation is experienced (*Henrich et al*, 2001; 2004). In Ultimatum Game experiments, laboratory experiments often find that larger stakes resulted in no or only minor changes in behavior (*Hoffman et al*, 1994; *Fehr and Tougareva*, 1995; *Slonim and Roth*, 1998; *Cameron*, 1999; *Fehr and Schmidt*, 2005).

In the classical Prisoner's Dilemma experiments of *Rapoport and Chammah* (1965), consistent with later findings in the literature (cf. *Ledyard*, 1995), modifying single payoff parameters has resulted in the expected changes in cooperation rates. In some cases, where there are more than one payoff difference between the two games compared, however, unexpected differences occur in cooperation rates. For instance, the comparison of Games II ( $T = 10, R = 1, P = -9, S = -10$ ) and IV ( $T = 2, R = 1, P = -1, S = -2$ ) reveal that modifying  $P$  in favor of defection and two other parameters ( $T, S$ ) in favor of cooperation equally within the same experiment does not result in higher, but in lower cooperation rates (Game II: 77%, Game IV: 66%). A similar paradox with three payoff differences occur for

the comparison of Games I ( $T = 10, R = 9, P = -1, S = -10$ ) and IV (Game I: 73%;  
*Rapoport and Chammah*, 1965: 37).

The results imply that interventions that invest less in cooperation benefits lead to a larger proportion of cooperators if interactions take place at a certain level of positive assortment, but also imply that environmental systems that reward cooperators to a lesser extent can sustain more cooperation, if cooperators meet each other more often than what pure chance would dictate.

## References

- Axelrod, R. and Hamilton, W. D., 1981. The evolution of cooperation. *Science* 211, 1390-1396.
- Axelrod, R., 1984. *The evolution of cooperation*. New York, Basic Books.
- Barclay, P., 2008. Enhanced Recognition of Defectors Depends on Their Rarity. *Cognition*, 107(3): 817-828.
- Becker, G. S., 1976. Altruism, egoism, and genetic fitness: Economics and sociobiology. *J. Econ. Lit.* 14, 817-826.
- Bergstrom, T. C., 2003. The Algebra of Assortative Encounters and the Evolution of Cooperation. *International Game Theory Review*, 5(3): 211-228.
- Bowles, S. and Gintis, H., 1998. The moral economy of communities: Structured populations and the evolution of prosocial norms. *Evol. Human Behav.* 19, 3-25.
- Cameron, L. A., 1999. Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia. *Economic-Inquiry* 37(1), 47-59.
- Cavalli-Sforza, L.L. and Feldman, M.W., 1981. *Cultural transmission and evolution: A quantitative approach*. Princeton University Press, Princeton, NJ.
- Cosmides, L. and Tooby, J., 1992. Cognitive Adaptations for Social Exchange. In: Barkow, J. H.; Cosmides, L., and Tooby, J. (eds.): *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York

*etc.*

- , Oxford University Press.
- Cosmides, L., 1989. The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition* 31, pp. 187-276.
- Doebeli, M. and Hauert, C., 2005. Models of cooperation based on the Prisoner's Dilemma and the Snowdrift game. *Ecol. Lett.*, 8, 748-766.
- Eberhard, W. G., 1980. Horned Beetles. *Scient. Am.* March, 124-131.
- Fehr, Ernst and Schmidt, Klaus M., 2005. The Economics of Fairness, Reciprocity and Altruism - Experimental Evidence and New Theories. Discussion Paper 2005-20, Department of Economics, University of Munich, [http://epub.ub.uni-muenchen.de/726/1/Fehr-Schmidt\\_Handbook\(2005-Munichecon\).pdf](http://epub.ub.uni-muenchen.de/726/1/Fehr-Schmidt_Handbook(2005-Munichecon).pdf)
- Fehr, Ernst and Tougareva, Elena, 1995. Do Competitive Markets with High Stakes Remove Reciprocal Fairness? Experimental Evidence from Russia. Manuscript.
- Frank, S. A., 1995. George Price's contributions to evolutionary genetics. *J. Theor. Biol.* 175, 373-388.
- Grafen, A. and Archetti, M., 2008. Natural selection of altruism in inelastic viscous homogeneous population. *J. Theor. Biol.* 252: 694-710.
- Grafen, A., 1985. A geometric view of relatedness. *Oxford Surveys Evol. Biol.* 2, 28-90.
- Grafen, A., 2007. An inclusive fitness analysis of altruism on a cyclical network. *J. Evol. Biol.* 20, 2278-2283.
- Hamilton, W. D., 1964. The genetical evolution of social behavior: I. *J. Theor. Biol.* 7, 1-16.
- Hamilton, W.D., 1971. Selection of selfish and altruistic behavior in some extreme models. In J.F. Eisenberg and W.S. Dillon, editors, *Man and Beast: Comparative Social Behavior*, pages 57-91. Smithsonian Press, Washington, D.C.
- Hamilton, W.D., 1975. Inate social aptitudes of man: an approach from evolutionary genetics. In R. Fox, editor, *Biosocial Anthropology*, pages 133-155. Malaby Press, London.
- Henrich, Joseph; Boyd, Robert; Bowles, Samuel; Camerer, Colin; Fehr, Ernst; Gintis, Herbert, and McElreath, Richard, 2001. In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *American Economic Review* 91(Papers and Proceedings): 73-78.

- Henrich, J.; R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis (eds.), 2004. *Foundations of Human sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press.
- Hoffman, Elisabeth, Kevin McCabe, Keith Shachat, and Vernon Smith, 1994. „Preferences, Property Right, and Anonymity in Bargaining Games”, *Games and Economic Behavior* 7, 346-380.
- Hubermann, B. A. and Glance, N. S., 1993. Evolutionary games and computer simulations. *Proc. Natl Acad. Sci. U.S.A.* 90, 7712-7715.
- Ledyard, J. O., 1995. Public Goods: A Survey of Experimental Research. In: Kagel, J. H. and Roth, A. E. (eds.): *The Handbook of Experimental Economics*. Princeton NJ, Princeton University Press.
- Maynard Smith, J. and Price, G. R., 1973. The logic of animal conflicts. *Nature*, 246, 15-18.
- Maynard Smith, J., 1982. *Evolution and the Theory of Games*. Cambridge, Cambridge University Press.
- Németh, A. and Takács, K., 2007. The evolution of altruism in spatially structured populations. *J. Artificial Societies Soc. Simul.* 10(3), 4, <http://jasss.soc.surrey.ac.uk/10/3/4.html>.
- Nowak, M. A. and May, R. M., 1992. Evolutionary games and spatial chaos. *Nature* 359, 826-829.
- Nowak, M. A. and May, R. M., 1993. The spatial dilemmas of evolution. *Int. J. Bifurcation Chaos* 3, 35-78.
- Nowak, M. A., Bonhoeffer, S. and May, R. M., 1994. Spatial games and the maintenance of cooperation. *Proc. Natl Acad. Sci. U.S.A.* 91, 4877-4881.
- Ohtsuki, H., Hauert, C., Lieberman, E. and Nowak, M. A., 2006. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441, 502-505.
- Price, G. R., 1970. Selection and covariance. *Nature*, 227, 520-521.
- Queller, D. C., 1985. Kinship, reciprocity and synergism in the evolution of social behavior. *Nature* 318, 366-367.

- Queller, D. C., 1994. Genetic relatedness in viscous populations. *Evol. Ecol.* 8, 70-73.
- Rapoport, Anatol and Chammah, A. M., 1965. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbor (MI), University of Michigan Press.
- Santos, F. C., Pacheco, J. M., and Lenaerts, T., 2006. Cooperation Prevails When Individuals Adjust Their Social Ties. *PLoS Computational Biology*, 2(10): e140.
- Schuessler, R., 1989. Exit Threats and Cooperation Under Anonymity. *Journal of Conflict Resolution*, 33(4): 728-749.
- Slonim, R. and Roth, A. E., 1998. Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic. *Econometrica*, 66(3): 569-596.
- Taylor, P. D., 1992a. Inclusive fitness in a homogeneous environment. *Proc. R. Soc. Lond. B.* 249, 299-302.
- Taylor, P. D., 1992b. Altruism in viscous populations - an inclusive fitness approach. *Evol. Ecol.* 6, 352-356.
- van Baalen, M. and Rand, D. A., 1998. The unit of selection in viscous populations and the evolution of altruism. *J. Theor. Biol.* 193: 631-648.
- van Veelen, M., 2005. On the use of the Price equation. *J. Theor. Biol.* 237, 412-426.
- Vanberg, V. and Congleton, R., 1992. Rationality, Morality and Exit. *American Political Science Review*, 86: 418-431.
- Wang, S., Szalay, M. S., Zhang, C. and Csermely, P., 2008. Learning and Innovative Elements of Strategy Adoption Rules Expand Cooperative Network Topologies. *PLoS ONE* 3(4): e1917. doi:10.1371/journal.pone.0001917
- Wilson, D. S. and Dugatkin, L. A., 1997. Group Selection and Assortative Interactions. *Am. Naturalist* 149, 336-351.
- Wilson, D. S., Pollock, G. B. and Dugatkin, L. A., 1992. Can altruism evolve in purely viscous populations? *Evol. Ecol.* 6, 331-341.
- Yamagishi, Toshio and Hayashi, Nahoko, 1996. Selective Play: Social Embeddedness of Social Dilemmas. In: Liebrand, W. B. G. and Messick, D. M. (eds.): *Frontiers in Social Dilemma Research*. Berlin, Springer.

Yamagishi, T.; Hayashi, N., and Jin, N., 1994. Prisoner's Dilemma Networks: Selection Strategy Versus Action Strategy. In: Schulz, U.; Albers, W., and Mueller, U. (eds.): Social Dilemmas and Cooperation. Berlin, Springer Verlag.

Yamagishi, T.; Tanida, S., Mashima, R., Shimoma, E., and Kanazawa, S., 2003. You Can Judge a Book by Its Cover - Evidence that Cheaters May Look Different from Cooperators. *Evolution and Human Behavior*, 24(4): 290-301.

## Appendix I

In this appendix, cases in Table 2 (Prisoner's Dilemma with positive assortment and  $\alpha_L < \alpha_H$ ) marked with asterisk will be explored. For all pairs of parameters there are two cases, which are mirrored, so it is enough to analyze one of them. In all the cases below we refer to  $\alpha_T$ , which is defined in equation (2).

### Increasing T and R

- if  $(1 < \frac{T-P}{R-P} < \frac{\Delta T}{\Delta R}$   
then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H > 0$ : paradox for  $\alpha < \alpha_T$ , where beneficial for C (vs  $\Delta T > \Delta R$ ). See Figure 8 for this case.  $\frac{T-P}{R-P} < \frac{\Delta T}{\Delta R}$  can be deduced from  $\alpha'_H > \alpha_H$  that is  $\frac{T+\Delta T-R-\Delta R}{T+\Delta T-P} > \frac{T-R}{T-P}$ . Other results in the Appendix are obtained in a similar way.
- if  $1 < \frac{\Delta T}{\Delta R} < \frac{T-P}{R-P}$   
then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H < 0$ : paradox for every  $\alpha$ , because beneficial for C (vs  $\Delta T > \Delta R$ )
- if  $\frac{\Delta T}{\Delta R} < 1$   
then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H < 0$ : normal for every  $\alpha$

### Increasing T and decreasing P

- if  $1 < \frac{T-R}{R-P}$ :  
– if  $\frac{T-R}{R-P} < \frac{\Delta T}{|\Delta P|}$   
then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H > 0$ : paradox for  $\alpha < \alpha_T$ , where beneficial for C (vs  $\Delta T > |\Delta P|$ )

- if  $1 < \frac{\Delta T}{|\Delta P|} < \frac{T-R}{R-P}$   
then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H < 0$ : paradox for every  $\alpha$ , because beneficial for C (vs  $\Delta T > |\Delta P|$ )
- if  $\frac{\Delta T}{|\Delta P|} < 1$   
then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H < 0$ : normal for every  $\alpha$
- if  $\frac{T-R}{R-P} < 1$ :
  - if  $1 < \frac{\Delta T}{|\Delta P|}$   
then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H > 0$ : paradox for  $\alpha < \alpha_T$ , where beneficial for C (vs  $\Delta T > |\Delta P|$ )
  - if  $\frac{T-R}{R-P} < \frac{\Delta T}{|\Delta P|} < 1$   
then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H > 0$ : paradox for  $\alpha > \alpha_T$ , where beneficial for D (vs  $\Delta T < |\Delta P|$ )
  - if  $\frac{\Delta T}{|\Delta P|} < \frac{T-R}{R-P}$   
then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H < 0$ : normal for every  $\alpha$

## Increasing T and S

- if  $\Delta T > \Delta S$   
 $\Delta\alpha_L < 0$  and  $\Delta\alpha_H > 0$ : paradox for  $\alpha < \alpha_T$ , where beneficial for C (vs  $\Delta T > \Delta S$ )
- if  $\Delta T < \Delta S$   
 $\Delta\alpha_L < 0$  and  $\Delta\alpha_H > 0$ : paradox for  $\alpha > \alpha_T$ , where beneficial for D (vs  $\Delta T < \Delta S$ )

## Increasing R and P

- if  $\frac{\Delta P}{\Delta R} < \frac{P-S}{R-S} (< 1)$   
then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H < 0$ : normal for every  $\alpha$
- if  $\frac{P-S}{R-S} < \frac{\Delta P}{\Delta R} < 1$   
then  $\Delta\alpha_L > 0$  and  $\Delta\alpha_H < 0$ : paradox for  $\alpha < \alpha_T$ , where beneficial for D (vs  $\Delta P < \Delta R$ )
- if  $1 < \frac{\Delta P}{\Delta R} < \frac{T-P}{T-R}$

then  $\Delta\alpha_L > 0$  and  $\Delta\alpha_H < 0$ : paradox for  $\alpha > \alpha_T$ , where beneficial for C (vs

$$\Delta P > \Delta R)$$

- if  $(1 <) \frac{T-P}{T-R} < \frac{\Delta P}{\Delta R}$

then  $\Delta\alpha_L > 0$  and  $\Delta\alpha_H > 0$ : normal for every  $\alpha$

## Increasing R and decreasing S

- if  $1 < \frac{P-S}{R-P}$ :

- if  $\frac{P-S}{R-P} < \frac{|\Delta S|}{\Delta R}$

then  $\Delta\alpha_L > 0$  and  $\Delta\alpha_H < 0$ : paradox for  $\alpha > \alpha_T$ , where beneficial for C (vs

$$|\Delta S| > \Delta R)$$

- if  $1 < \frac{|\Delta S|}{\Delta R} < \frac{P-S}{R-P}$

then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H < 0$ : paradox for every  $\alpha$ , because beneficial for C (vs

$$|\Delta S| > \Delta R)$$

- if  $\frac{|\Delta S|}{\Delta R} < 1$

then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H < 0$ : normal for every  $\alpha$

- if  $\frac{P-S}{R-P} < 1$ :

- if  $1 < \frac{|\Delta S|}{\Delta R}$

then  $\Delta\alpha_L > 0$  and  $\Delta\alpha_H < 0$ : paradox for  $\alpha > \alpha_T$ , because beneficial for C (vs

$$|\Delta S| > \Delta R)$$

- if  $\frac{P-S}{R-P} < \frac{|\Delta S|}{\Delta R} < 1$

then  $\Delta\alpha_L > 0$  and  $\Delta\alpha_H < 0$ : paradox for  $\alpha < \alpha_T$ , where beneficial for D (vs

$$|\Delta S| < \Delta R)$$

- if  $\frac{|\Delta S|}{\Delta R} < \frac{P-S}{R-P}$

then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H < 0$ : normal for every  $\alpha$

## Increasing P and S

- if  $(1 <) \frac{R-S}{R-P} < \frac{\Delta S}{\Delta P}$

then  $\Delta\alpha_L < 0$  and  $\Delta\alpha_H > 0$ : paradox for  $\alpha > \alpha_T$ , where beneficial for D (vs

$$\Delta S > \Delta P)$$

- if  $1 < \frac{\Delta S}{\Delta P} < \frac{R-S}{R-P}$

then  $\Delta\alpha_L > 0$  and  $\Delta\alpha_H > 0$ : paradox for every  $\alpha$ , because beneficial for D (vs  $\Delta S > \Delta P$ )

- if  $\frac{\Delta S}{\Delta P} < 1$

then  $\Delta\alpha_L > 0$  and  $\Delta\alpha_H > 0$ : normal for every  $\alpha$

## Appendix II

Consider the special case where  $T = b$ ,  $R = b - c$ ,  $P = 0$ , and  $S = -c$ , hence

$T - R = P - S = c$ . In this special case, we have less parameters ( $b, c$ ) and these parameters can easier be interpreted as benefits and costs. In this case,  $z = 0$  is evolutionary stable if  $\alpha b - c < 0$ ,  $z = 1$  is evolutionary stable if  $\alpha b - c > 0$ , and there is no mixed equilibrium.

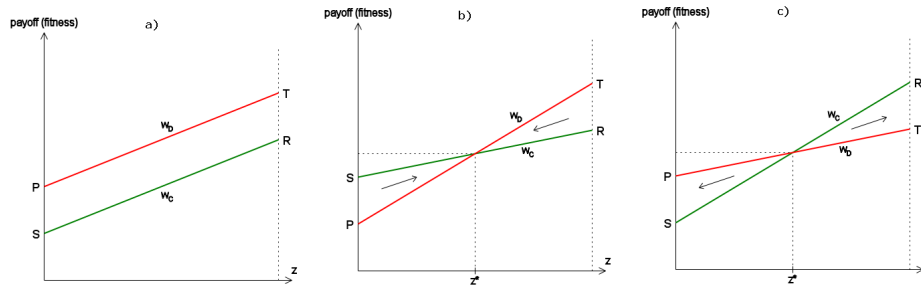
Hence, we have obtained a re-derivation of Hamilton's rule for this special case with positive assortment (a similar re-derivation is given in *Hamilton*, 1971; 1975; *Bergstrom*, 2003).

Furthermore, there are no paradoxical cases as the increase of  $c$  is always beneficial for defection and the increase of  $b$  is always beneficial for cooperation, and as expected, the simultaneous increase of  $c$  and  $b$  (if  $\Delta c > \Delta b$ ) also benefits defection.<sup>5</sup>

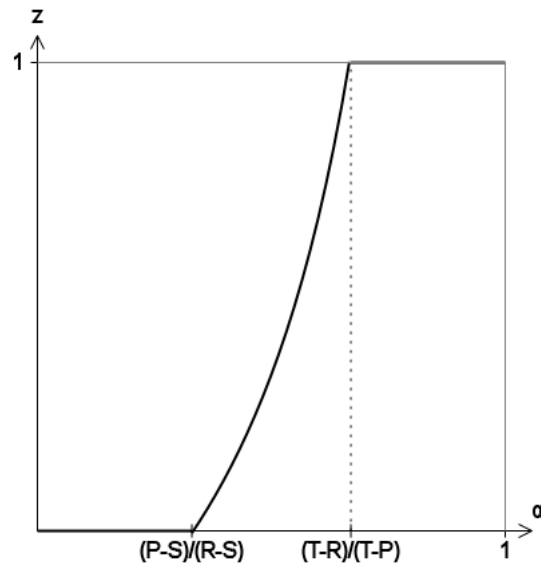
In short, the fundamental reason why Hamilton's rule leaves no space for paradoxes lies in the simplification of the social dilemma with a single cost ( $c$ ) and benefit ( $b$ ) parameter with the restriction of  $T - R = P - S = c$ . The simplified Prisoner's Dilemma game nicely applies to symmetric decisions of altruistic help, where altruists suffer costs, but the benefits of their altruistic act are solely enjoyed by their interaction partner (e.g. *Doebeli and Hauert*, 2005; *Ohtsuki et al*, 2006). This is, however, only a special case of Prisoner's Dilemma games. If the payoffs of the Prisoner's Dilemma are expressed as independent parameters with their ordinal ranking being fixed (see e.g. *Axelrod*, 1984), then we obtain the paradoxical results in which the increase in the benefits of cooperation can result in a lower proportion of cooperators in the population.

---

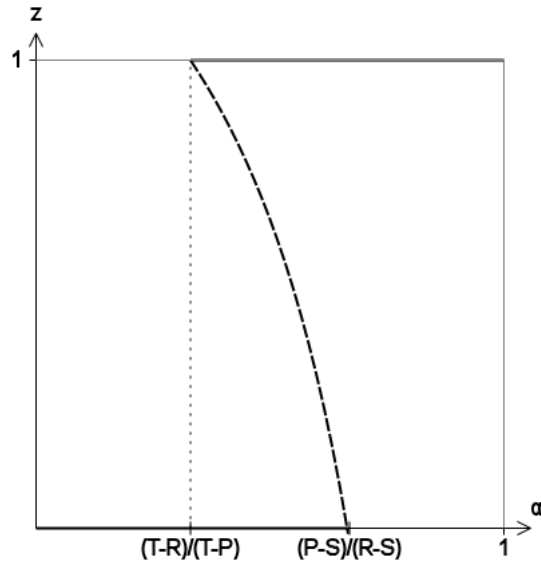
<sup>5</sup>A paradoxical case ( $\Delta c > \Delta b$  and increasing  $\alpha b - c$ ) might occur when  $b < c$ , but it implies that  $R < P$ , which cannot hold in a social dilemma game.



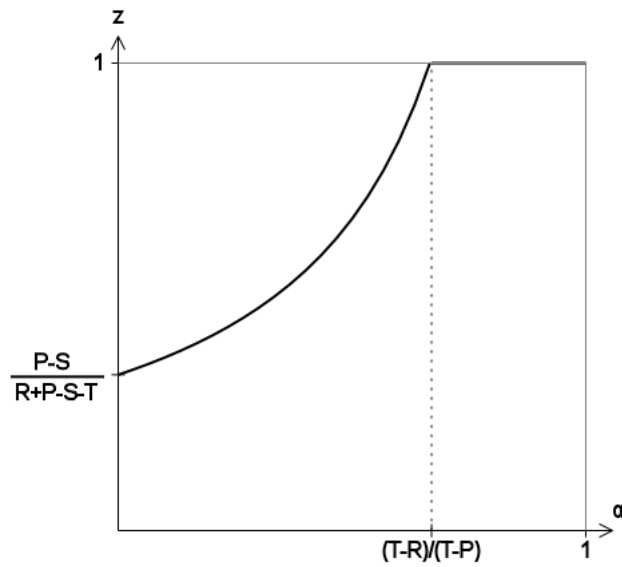
**Figure 1:** ESS in Social Dilemmas with Complete Mixing. a) There is no mixed equilibrium in the Prisoner's Dilemma. The only ESS is defection. b) The only ESS in the Hawks and Doves game is a mixed equilibrium. c) The mixed equilibrium is not an ESS in the Stag Hunt game.  $z^* = \frac{P - \alpha R - (1 - \alpha)S}{(1 - \alpha)(R + P - S - T)}$ . For the meaning of  $\alpha$  see section 2.1.



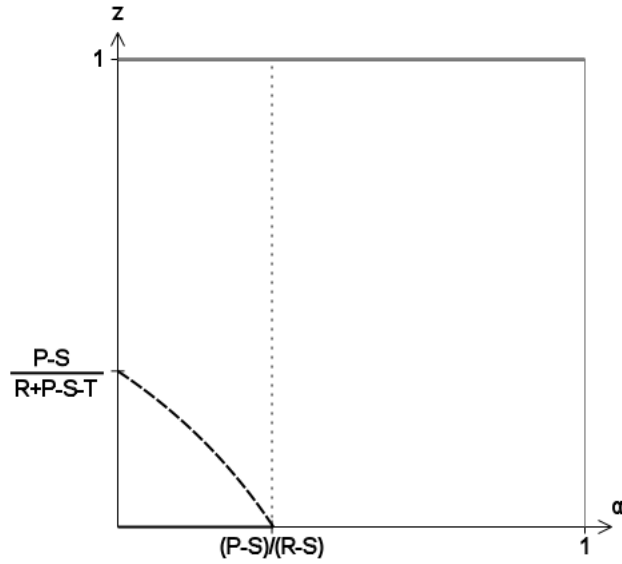
**Figure 2:** Evolutionary stable values of  $z$  as a function of  $\alpha$  in Prisoner's Dilemma with positive assortment if  $\frac{P-S}{R-S} < \frac{T-R}{T-P}$ .



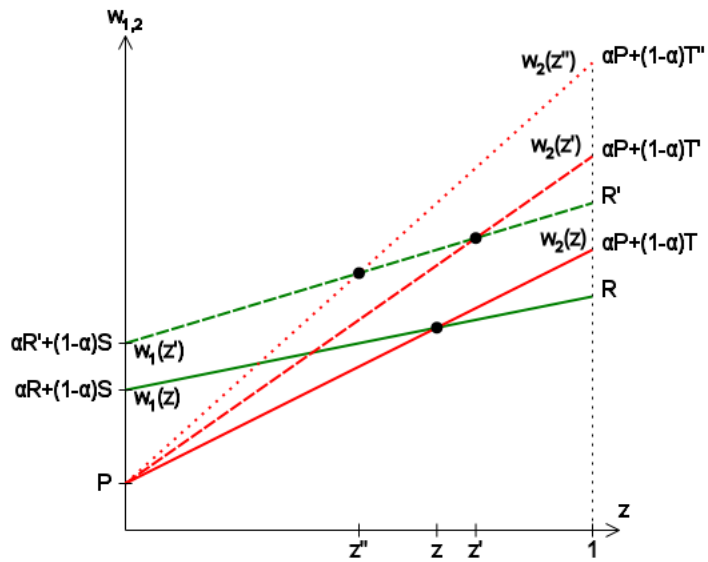
**Figure 3:**  $z(\alpha)$  in Prisoner's Dilemma with positive assortment if  $\frac{P-S}{R-S} > \frac{T-R}{T-P}$ . The solid lines indicate the stable equilibria of full cooperation and full defection. The dashed line indicates the unstable mixed equilibrium  $z^*$ .



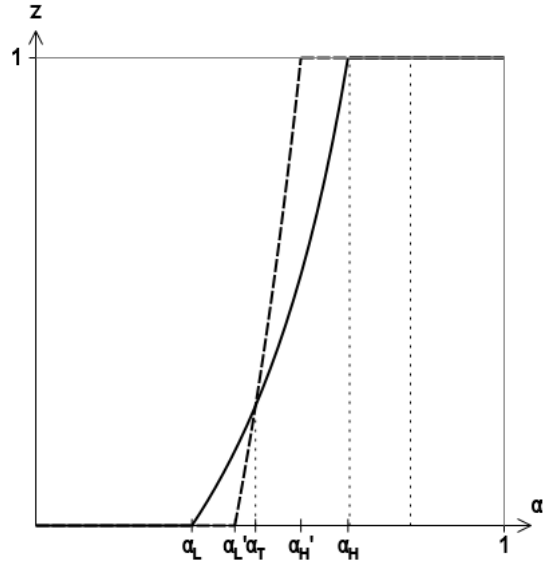
**Figure 4:** Evolutionary stable values of  $z$  as a function of  $\alpha$  in the Hawks and Doves game with positive assortment.



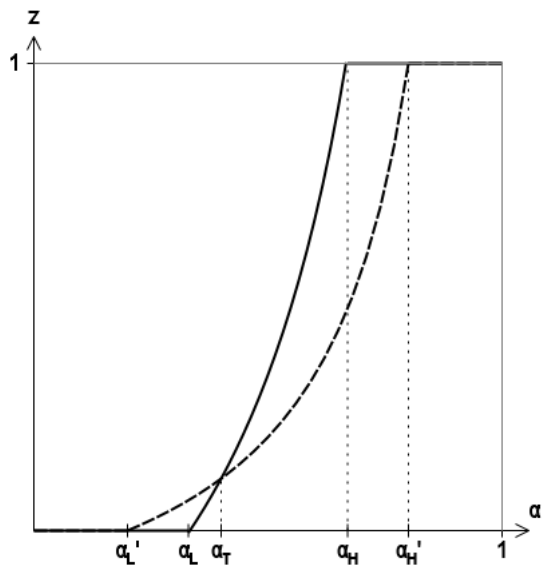
**Figure 5:**  $z(\alpha)$  in Stag Hunt with positive assortment. The solid lines indicate the stable equilibria of full cooperation and full defection. The dashed line indicates the unstable mixed equilibrium  $z^*$ .



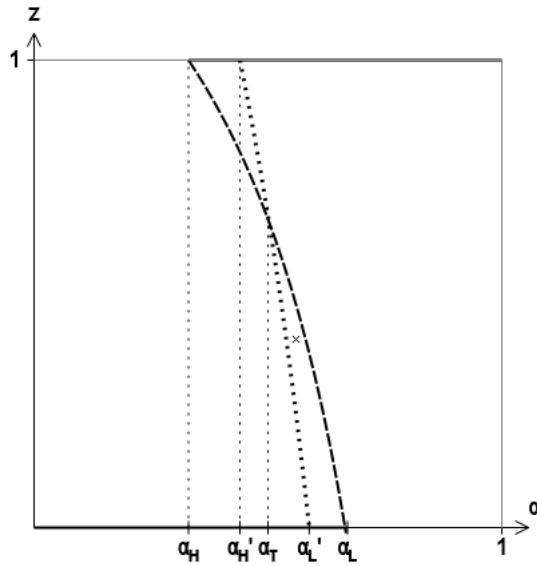
**Figure 6:** Mixed equilibria (ESS) in the Prisoner's Dilemma with positive assortment when  $R$  and  $T$  are increased. The solid lines indicate the original fitness functions. Dashed and dotted lines indicate the new fitness functions.  $w_1(z)$  in green and  $w_2(z)$  in red.  $z''$  is the normal,  $z'$  is the paradoxical case.



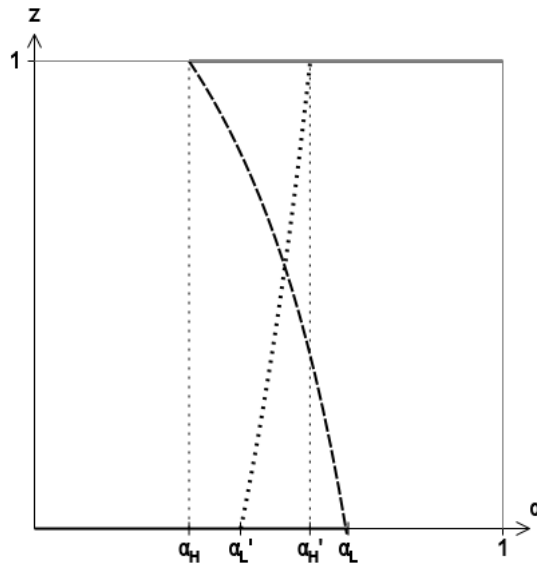
**Figure 7:** The paradox of cooperation benefits in the Prisoner's Dilemma with positive assortment and  $\alpha_L < \alpha_H$ . The solid line is the original  $z(\alpha)$  function. The dashed line shows the case when  $P$  and  $R$  are increased. The numerical values for this figure are:  $T = 7$ ,  $R = 3$ ,  $P = 1$ ,  $S = 0$ ,  $\Delta R = 1$ , and  $\Delta P = 0.7$  ( $\Delta R > \Delta P$ ).



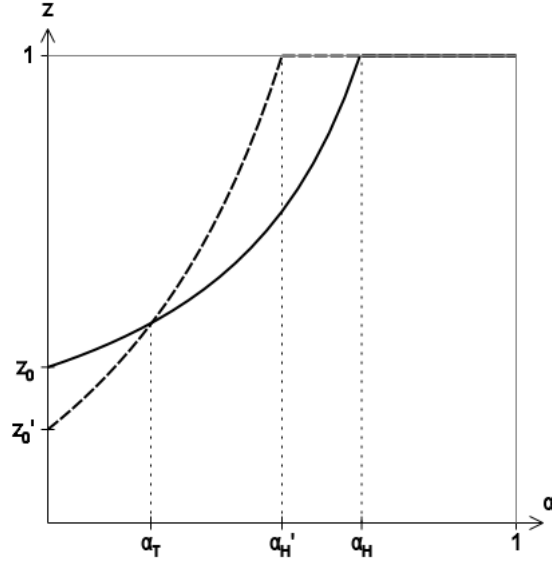
**Figure 8:** The paradox of cooperation benefits in the Prisoner's Dilemma with positive assortment and  $\alpha_L < \alpha_H$ . The solid line is the original  $z(\alpha)$  function. The dashed line shows the case when  $R$  and  $T$  are increased. The numerical values for this figure are:  $T = 7$ ,  $R = 3$ ,  $P = 1$ ,  $S = 0$ ,  $\Delta T = 14$ , and  $\Delta R = 2$  ( $\Delta T \gg \Delta R$ ).



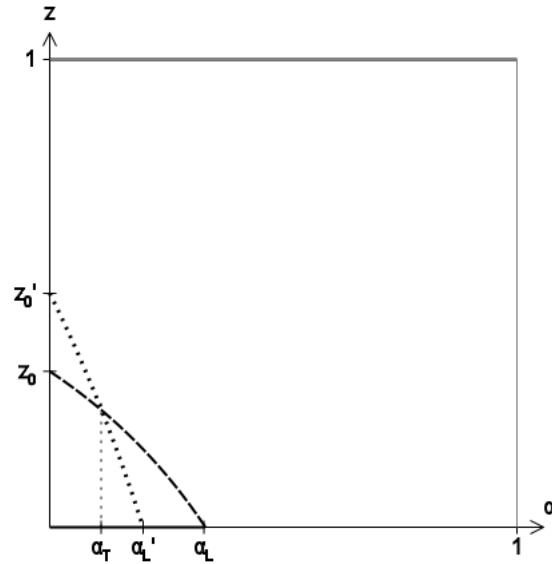
**Figure 9:** The paradox of cooperation benefits in the Prisoner's Dilemma with positive assortment and  $\alpha_L > \alpha_H$  and  $\alpha'_L > \alpha'_H$ . The solid lines indicate the original stable equilibria of full cooperation and full defection. The dashed line indicates the unstable mixed equilibrium  $z^*$ . The dotted line shows the case when  $R$  and  $T$  are increased (the mixed equilibrium is still unstable). The numerical values for this figure are:  $T = 1.75$ ,  $R = 1.5$ ,  $P = 1$ ,  $S = 0$ ,  $\Delta T = 0.5$ , and  $\Delta R = 0.2$  ( $\Delta T > \Delta R$ ). The  $x$  indicates a random initial ratio which evolves to full defection with the original parameters, and to full cooperation with the modified ones.



**Figure 10:** The paradox of cooperation benefits in the Prisoner's Dilemma with positive assortment and  $\alpha_L > \alpha_H$  and  $\alpha'_L < \alpha'_H$ . The solid lines indicate the original stable equilibria of full cooperation and full defection. The dashed line indicates the unstable mixed equilibrium  $z^*$ . The dotted line shows the case when  $R$  and  $T$  are increased (the mixed equilibrium becomes stable). The numerical values for this figure are:  $T = 1.75$ ,  $R = 1.5$ ,  $P = 1$ ,  $S = 0$ ,  $\Delta T = 2.34$ , and  $\Delta R = 0.77$  ( $\Delta T > \Delta R$ ).



**Figure 11:** The paradox of cooperation benefits in the Hawks and Doves game with positive assortment. The solid line is the original  $z(\alpha)$  function. The dashed line shows the case when  $R$  and  $T$  are increased. The numerical values for this figure are:  $T = 3$ ,  $R = 1$ ,  $P = 0$ ,  $S = 1$ ,  $\Delta T = 5$ , and  $\Delta R = 3$  ( $\Delta T > \Delta R$ ).



**Figure 12:** The paradox of cooperation benefits in the Stag Hunt game with positive assortment. The solid lines indicate the original stable equilibria of full cooperation and full defection. The dashed line indicates the unstable mixed equilibrium  $z^*$ . The dotted line shows the case when  $R$  and  $T$  are increased. The numerical values for this figure are:  $T = 1$ ,  $R = 3$ ,  $P = 1$ ,  $S = 0$ ,  $\Delta T = 3$ , and  $\Delta R = 2$  ( $\Delta T > \Delta R$ ).

	C	D
C	$R,R$	$S,T$
D	$T,S$	$P,P$

**Table 1:** Payoffs in Social Dilemma Games.  $T > R > P > S$  in the Prisoner's Dilemma,  $T > R > S > P$  in the Hawks and Doves game, and  $R > T > P > S$  in the Stag Hunt game.

parameters	$\alpha_L$	$\alpha_H$	paradox
T↑ R↑	↓ (C)	?	*
T↑ R↓	↑ (D)	↑ (D)	
T↓ R↑	↓ (C)	↓ (C)	
T↓ R↓	↑ (D)	?	*
T↑ P↑	↑ (D)	↑ (D)	
T↑ P↓	↓ (C)	?	*
T↓ P↑	↑ (D)	?	*
T↓ P↓	↓ (C)	↓ (C)	
T↑ S↑	↓ (C)	↑ (D)	**
T↑ S↓	↑ (D)	↑ (D)	
T↓ S↑	↓ (C)	↓ (C)	
T↓ S↓	↑ (D)	↓ (C)	**
R↑ P↑	?	?	*
R↑ P↓	↓ (C)	↓ (C)	
R↓ P↑	↑ (D)	↑ (D)	
R↓ P↓	?	?	*
R↑ S↑	↓ (C)	↓ (C)	
R↑ S↓	?	↓ (C)	*
R↓ S↑	?	↑ (D)	*
R↓ S↓	↑ (D)	↑ (D)	
P↑ S↑	?	↑ (D)	*
P↑ S↓	↑ (D)	↑ (D)	
P↓ S↑	↓ (C)	↓ (C)	
P↓ S↓	?	↓ (C)	*

**Table 2:** Effects of changing two parameters. ↑/↓ for increase/decrease, C/D for beneficial for cooperators/defectors, ? for ambiguous (depends on the exact values of these and other parameters). \* denotes cases, when there might be a paradox in certain ranges of the payoff parameters (and their changes), and \*\* denotes cases, when there is a paradox for certain  $\alpha$  values for all ranges of the payoff parameters.