# A Model of Conditional Altruism

Wilma Erhardt* and Ksenia Kosolapova**

Stockholm School of Economics

**Abstract**

Altruism is a concept that so far does not fit into mainstream economic theory. However, it has recently gained more attention in an attempt to justify non-selfish behavior of *homo economicus*, which is widely observed in economic experiments as well as social interactions. We develop a model of conditional altruism, in the framework of Prisoners' Dilemma, which allows to explain situations where people with altruistic social preferences do not behave altruistically. We show that incorporation of altruistic preferences into the utility function of individuals does not guarantee that they behave altruistically. Depending on their environment, even altruistically oriented individuals may decide to defect and the desired cooperation in the society may not be obtained and/or sustained. We also perform a series of computer simulations in order to strengthen our theoretical analysis and findings about the factors that influence people's decision to cooperate. We also assess the stability of the outcomes of the social interaction under consideration. We argue that issues such as beliefs, expectations, and mutual trust are of major importance for economic interactions.

**Tutor**: Jörgen Weibull
**Examiner:** Karl Wärneryd
**Opposition:** Fredrik Paues
**Presentation:** December 11, 2008

*80350@student.hhs.se
**80349@student.hhs.se

# Acknowledgement

We would like to thank our tutor Jörgen Weibull wholeheartedly. We thank him especially for inspiring discussions, supportive encouragement and advising cross-border phone conferences.

We also would like to express our gratitude to Tore Ellingsen, who gave us the right hint at the right time.

# Contents

# 1 Introduction

> "How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it. Of this kind is pity or compassion, the emotion which we feel for the misery of others, when we either see it, or are made to conceive it in a very lively manner. That we often derive sorrow from the sorrow of others, is a matter of fact too obvious to require any instances to prove it; for this sentiment, like all the other original passions of human nature, is by no means confined to the virtuous and humane, though they perhaps may feel it with the most exquisite sensibility. The greatest ruffian, the most hardened violator of the laws of society, is not altogether without it" (Smith (1759)).

In 1759 Adam Smith, the great-grandfather of economics, published his book *The Theory of Moral Sentiments*, and he already described, what we call altruism today. However, from 1759 until today, altruism did not experience a straightforward development, on the contrary, from its rare appearances in literature one may conclude that it has been forgotten for a long time.

Today altruism is not an uncommon concept anymore (Fehr and Fischbacher (2003)). Although only 10 years ago, Sen (1998) had to state that "many economic models tend to proceed as if the assumption of universal pursuit of self interest is the only motivation that can be legitimately presumed in serious economic analysis". He further noticed that it was assumed that the *homo economicus* directly followed the *homo neanderthalensis* and "turned everyone [...] into 'smooth-faced gentlemen, tickling commodity'". People interact on a daily basis with each other and economics tries to model such situations to predict future outcomes, however, as Ben-Ner and Puttermann (1998) express it, "it is difficult to reconcile some game-theoretic predictions with observed behaviors unless models of preferences are extended to include elements conventionally excluded from them".

Nowadays many economists recognize the fact that the *homo economicus* is only made-up to simplify the use of various models but does not exist in the real world. Whenever people interact with each other, when they trade, for example, values and moral behavior play an important role besides the economic rational. Referring once more to Sen's words, he claims that the assumption of pure self-interest is not more 'elementary' than assuming other values (Sen (1998)).

A society consists of many different types of individuals, like altruists, egoists, indifferent people, and many more. It is generally considered that if altruistic people exist in a population, they can positively influence a society towards more favorable outcomes.

However, for models of repeated Prisoners' Dilemma games, it has been shown that altruism can have negative effects on cooperation (Nakao (2008)). We are not interested in the direct negative effects of altruism on cooperation but rather what happens in the society given that individuals prefer to cooperate, though, behaving in line with their true preferences would result in a personal loss. Due to norms or established rules of the game, they may be forced to adjust their behavioral habits.

We would like to contribute to the understanding of social preferences and add a small puzzle piece to the overall picture of altruistic behavior into economic theory. We consider

our work as a relevant reference for any economic interaction if expectations, beliefs, or mutual trust play a role. One such example would be the case of contracting. We are living in the world of incomplete information and cannot expect to obtain full information about our counterpart and his or her potential behavior in the social situations we face. Therefore, the assumptions made in such situations may be incorrect or biased and, thus, lead to the failure of trust or even failure of contract between the parties involved.

We begin our work by describing the theoretical background, and hereby we highlight the development of altruism in economics by referring to concepts of inequity aversion, pure and paternalistic altruism, and direct and indirect reciprocity. In *Section 3* we introduce the model and its preliminaries. With *Section 4* follows the analysis of the model. *Section 5* includes some thoughts on the stability of the equilibria we find. Further, in *Section 6* we present some computer-based simulations for our model. And finally, we complete our work with concluding remarks in *Section 7*.

# 2 Theoretical Background

> "Increasing our knowledge of the nature of economic change requires that we utilize the only laboratory we have - the past. But to understand the past we must impose order on the myriad facts that have survived to explain what has happened, and doing so requires theory" (North (1998)).

Experimental economics provides evidence that people cooperate more often than they are expected to under the assumption of pure self-interest, and that individuals make voluntary contributions in social dilemma situations (*e.g.* Marwell and Ames (1981), Güth *et al.* (1982), Orbell *et al.* (1988) or Cronson (2007)). This behavior is known as altruism. A person is considered to be altruistic if her utility increases with the improved wellbeing of other people (Fehr and Schmidt (2001)). Not only do Andreoni and Miller (1998) conclude that altruistic behavior exists and is consistent under rationality assumption, but they also claim heterogeneity of preferences among individuals.

The concept of altruism does not only include human ability and desire to help and support others, but also is tightly connected with willingness to punish those who do not behave accordingly (*e.g.* Fehr and Gächter (2002)). Andreoni (1988) proposes that giving is consistent with social norms and that these norms tend to be enforced by punishing deviants. Falk *et al.* (2005) take into consideration both altruistic and spiteful punishment. According to the authors, the latter exists if an increase of the payoff difference between the player and the punished individual is observed. In contrast, they state that in case of altruistic punishment "cooperators increase their expenditures for punishment if the impact of a given investment in punishment causes a lower payoff reduction for the punished individual". Therefore the authors claim fairness preferences to be based on individualized payoff information and motives for revenge. They suggest that fairness theories should not only refer to the idea that players want to minimize payoff inequalities, since this would leave a considerable share of the cooperators' sanctions unexplained.

Another mechanism to mention in this context is the temptation to free ride. Fischbacher *et al.* (2001) find that in a public good game a non-negligible fraction of subjects free ride regardless of others' contribution. Even those who are conditionally cooperative display a

bias in the self-serving direction and contribute less than the others do on average. The authors, therefore, conclude that positive and stable contributions to the public good are very unlikely, that is, free riding will be pervasive when players interact anonymously despite the fact that conditional cooperators dominate in the population.

Furthermore, one can also observe that altruism is not without ulterior motives, that is, people are not willing to give without expecting to receive something in return. Experiments indicate that the willingness to help seems to be "highly dependent on the behavior of others. If people do not think that others are doing their fair share, then their enthusiasm for sacrificing for others is greatly diminished" (Rabin (1993)). Consequently, Rabin (1993) incorporates the concept of fairness into game theory. He claims "the same people who are altruistic to other altruistic people are also motivated to hurt those who hurt them". He explains this behavior with psychological evidence, which has shown that "people do not seek uniformly to help other people; rather, they do so according to how generous these other people are". His model incorporates psychological games, however, it can be generally applied to standard economic analysis.

In literature, altruism is often classified into two different categories, namely, pure altruism and paternalistic altruism. Pure altruism does not depend on any social conditions or norms. It is usually expressed as one person's preference for another person's material or psychic benefit (Konow (2006)). On the other side, an individual is considered to be paternalistically altruistic if he is not selfish but does not respect the preferences of his opponent (Jacobsson *et al.* (2007)).

Levine (1998) introduces a model of altruism and spitefulness. He aims to explain situations that cannot be explained with the assumption of the *homo economicus*, *i.e.,* a selfish player caring only about his own monetary income. In the model, Levine's player $i$ receives $u_i$, some direct utility, plus an adjusted $u_j$, the indirect utility, accounting for the opponent's payoff. The adjustment depends on a coefficient of altruism $\alpha$, within the interval $-1 < \alpha_i < 1$ and an element of fairness $\lambda$ within $0 \leq \lambda \leq 1$. "Players' weights on opponents' monetary payoffs depend both on their own coefficient of altruism or spite, and on what they believe their opponents' coefficients to be. In particular, a more positive weight is placed on the money received by an opponent who is believed to be more altruistic, and a more negative weight on one that is believed to be more spiteful". The model is defined as follows:

$$v_i = u_i + \sum_{j \neq i} \frac{\alpha_i + \lambda \alpha_j}{1 + \lambda} u_j$$

He applies the model to various experiments and concludes that the theory of altruism seems to fit with the experimental results, what cannot be said about the theory which incorporates *homo economicus* only.

When talking about altruism, it is very important to mention the concept of reciprocity, which people consider in any situation that implies repeated interaction. Reciprocity summarizes all actions by an individual that are based on previous actions by another individual, for instance, the return gift of an initial gift. This situation is very different from a self-interested exchange, where each transfer is provided only under the condition that the opponent provides something in return, and hence the opponent's behavior is not a gift but a necessity (Kolm (2006)). Fehr and Schmidt (1999) assume that a player is altruistic towards other players if their material payoffs are below an equitable benchmark. Further, they suggest

that the player is envious if the material payoffs of the other players exceed this certain level. They, therefore, conclude that in a public good game with punishment, a small fraction of inequity averse players is sufficient to credibly threaten free riders with punishment. This induces selfish players to contribute to the public good. However, Falk and Fischbacher (2001) come to another conclusion and suggest that reciprocal behavior is mainly driven as a response to kindness, and not as a desire to reduce inequality.

Further, another important issue regarding reciprocation is highlighted in the work of Fehr and Schmidt (2001). The authors state that the assumption that some people are fair-minded and have the desire to reciprocate does not imply that these people will always behave 'fairly'. In some environments, *e.g.*, in competitive markets or in public good games without punishment, fair-minded actors will often behave as if they are purely self-interested. Likewise, a purely self-interested person may often behave as if she is strongly concerned about fairness. Thus, the behavior of fair-minded and purely self-interested actors depends on the strategic environment in which they interact and on their beliefs about the fairness of their opponents. It is important for the players to estimate the underlying intentions of their opponents and to be aware of the environment they are located in. Furthermore, cooperation between individuals requires the ability to infer each other's mental state to form shared expectations over mutual gains and to make cooperative choices for these gains to be realized (McCabe *et al.* (2001)).

In addition to direct reciprocity, there is also a concept of indirect reciprocity. Under indirect reciprocity, researchers summarize situations, where an individual helps someone, who thereby gains more than the individual's help costs. If the help is reciprocated on the next occasion with a different individual, each individual has a net benefit. This implies that donors do not obtain a return from the recipient, but from a third party. It can be observed that donors provide help if the recipient has helped others in the past (Milinski *et al.* (2001)). This works if the cost of an altruistic act is set by a raised 'score', or status, which increases the chance to subsequently become the recipient of an altruistic act (Nowak and Sigmund (1998)). Sigmund (1998) concludes that cooperation is channelled towards the 'valuable' members of the community and he even suggest that the idea of indirect reciprocity can be applied in second-order defection, which is third-party punishment.

Indirect reciprocity itself is usually subdivided into two different categories. Nowak and Sigmund (2005) divide indirect reciprocity into 'upstream reciprocity', where reciprocal behavior is based on prior experiences (if A helps B then B helps C) and 'downstream reciprocity', where reciprocal behavior is based on reputation (if A helps B then C helps A). Stanca (2007) calls the same situations 'generalized indirect reciprocity' and 'social indirect reciprocity'. He implicitly includes damaging behavior into his definitions. For him positive or negative generalized indirect reciprocity "is a behavior to adopt a helpful or harmful action towards someone else, at one's own material cost, because some other person's intentional behavior was perceived to be helpful or harmful to oneself". And positive or negative social indirect reciprocity "is a behavior to adopt a helpful or harmful action towards someone else, at one's own material cost, because that person's intentional behavior was perceived to be helpful or harmful to some other person".

Indirect reciprocity requires the possibility to estimate someone else's behavior in the past and his potential attitude in the future. Under image scoring, researchers summarize methods that people use to form some intuition about their counterparts' possible manner. Nowak

and Sigmund (1998) declare that the "probability of knowing the image of the recipient must exceed the cost-to-benefit ratio of the altruistic act". They even argue that the development of indirect reciprocity was a crucial step for the evolution of human societies. Wedekind and Milinski (2000) reveal that image scoring promotes cooperative behavior in situations where direct reciprocity is unlikely. Hence, they provide experimental evidence for cooperation through indirect reciprocity in groups of human players.

The research about the concept of indirect reciprocity is not closed by a long shot. Researchers still are not sure whether altruism is really the driver behind this behavior. Stanca (2007) recommends that empirical studies should attempt to "identify what determines the perceived kindness of an action in determining of reciprocal behavior". He claims that reciprocity cannot be explained by models that focus only on the outcomes of the actions one is responding to. Therefore, he suggests that theoretical models of reciprocal behavior should also "take into account intentions and, in particular, consider explicitly the type of motivation driving an action". Finally, he concludes that "generalized reciprocity may represent a more general mechanism leading to cooperation than direct and indirect reciprocity, which require individual recognition and specific social memory". Tullberg (2004) proceeds even further and divides indirect reciprocal behavior into four categories; two of which "pertain to interaction between individuals and two of which involve social systems". The conclusion is that two of these categories, reciprocal reputation and institutionalized reciprocity, are strongly linked to reciprocity, whereas the other two categories, generous reputation and metaphysical reward, are likely to involve only an element of illusionary reciprocity and a substantial degree of altruism. He therefore requires that there should be a "strict separation between reciprocity and altruism, instead of using the term 'indirect reciprocity' as a wide gray zone". He argues that real indirect reciprocity, *i.e.*, reciprocal reputation and institutionalized reciprocity, is socially valuable, and that altruism, sometimes presented as indirect reciprocity, is more of "an obstacle than an asset to a democratic society".

The foregoing concepts of inequity aversion, pure and paternalistic altruism, direct and indirect reciprocity are important and popular areas of behavioral economics. The overview of the economic literature above implies the existence of widespread research possibilities when incorporating social preferences. We regard social preferences to be fundamental in our work. We argue that material self-interest is not the only line of behavior individuals can exhibit. Although self-interest is a mainstream assumption in economic theory, a consideration that people are able to possess more heterogeneous preferences enriches the scope of motives and incentives individuals employ and brings a fresh perspective to the understanding of human cooperation. In our work we basically follow the idea Levine (1998) suggested: individuals' preferences might differ in how other players' material payoffs are evaluated. We argue that people like people who are nice and, thus, care about those people's material payoffs. We complicate this reasoning by saying that cooperation is only feasible between two people who are not only nice and eager to cooperate potentially, but who are actually behaving cooperatively. In our model we assume that there exist selfish and altruistic players, however, we do not address the issue of why players should be altruistic or selfish. Neither image scoring nor reputation are taken into account in our model. We also exclude possibilities for punishment. Based on these premises, we develop a model together with our tutor, with which we demonstrate the behavior of conditional altruists in different situations.

Now we will discuss our model in detail.

# 3 The Model

We consider a simple model where two individuals are randomly matched and are involved in a two-person interaction. Consider a $2 \times 2$ Prisoners' Dilemma game with the following payoff matrix:

$$\Pi_0 = \begin{pmatrix} \pi_{cc} & \pi_{cd} \\ \pi_{dc} & \pi_{dd} \end{pmatrix}. \tag{1}$$

The player has two strategies: to cooperate, $C$, or to defect, $D$. Depending on the player's own strategy as well as her opponent's strategy the payoffs differ. To ensure that this payoff matrix stands for Prisoners' Dilemma we introduce the following condition:

$$\pi_{dc} > \pi_{cc} > \pi_{dd} > \pi_{cd}. \tag{2}$$

So far this game is a text-book Prisoners' Dilemma and now we will sophisticate it regarding the homogeneity of individuals who are to play it. From now on we no longer consider all individuals having exactly the same preferences. We introduce heterogeneity for preferences in order to differentiate between two types of individuals: selfish and conditionally altruistic.

The reason for which we assume heterogeneity in the population is the following: we assume some individuals may be more motivated by economic considerations, like monetary payoffs, others by social considerations, like their status, their reputation, or altruistic considerations. Therefore, we cannot expect all individuals to care equally about others (Fershtmann and Weiss (1998)).

We define a selfish player as one who cares only about her own payoff and does not take into consideration the payoff her opponent receives. A conditional altruist in our model is defined as an individual whose attitude to the opponent's payoff is dependent on the opponent's preferences. Namely, conditional altruists are those who care about payoffs of other conditional altruists. The utility function of a conditional altruist incorporates her opponent's weighted material payoff given this opponent is also conditionally altruistic. This idea is very much in line with the work of Levine (1998).

As mentioned before, the introduced types of individuals have different preferences and now we will generalize the payoff matrix discussed earlier, so that it represents the preferences of both types of individuals. Consider a $2 \times 2$ Prisoners' Dilemma game, where $\alpha \in [0, 1]$ is a weight assigned to the opponent's payoff:

$$\Pi_\alpha = \begin{pmatrix} \pi_{cc} + \alpha\pi_{cc} & \pi_{cd} + \alpha\pi_{dc} \\ \pi_{dc} + \alpha\pi_{cd} & \pi_{dd} + \alpha\pi_{dd} \end{pmatrix} \tag{3}$$

The value of $\alpha$ can be interpreted as the degree of altruism a conditional altruist holds.

We will analyze a population, where two different types of individuals exist. A selfish individual belongs to Type 0 and cares, by definition, only about her own payoff. Hence, all such individuals always choose strategy $D$. By contrast, an individual that is conditionally altruistic, also cares, to a degree $\alpha$, about her opponent's payoff if and only if the opponent is also a conditional altruist. If she meets a selfish player then she assigns $\alpha = 0$ to the payoff of her opponent. Conditional altruists are named Type $\alpha$, where $\alpha$ is the degree of conditional altruism. Generally, $\alpha \in [0, 1]$.

For our following example we decide to choose $\alpha = 1$, implying that a conditional altruist values the payoff her opponent receives, who is also conditionally altruistic, as much as she

values the payoff she receives herself. We will refer to these two types as Type 0 and Type 1, respectively. Let $p \in [0,1]$ be the population share of conditional altruists, and assume that this share is known by all individuals.

Let us now look at the decision-problem of a conditional altruist, when randomly matched with another individual from the population. With the probability $(1-p)$, the opponent is of Type 0 and will play $D$. With the probability $p$, the opponent is of Type 1. Let $x \in [0,1]$ be the probability that such an opponent will play $C$. Then, referring to (3), the expected utility function to our decision-maker, the conditional altruist, is:

$$U_C = (1-p)\,\pi_{cd} + p \cdot [x\,(1+\alpha)\,\pi_{cc} + (1-x)\,(\pi_{cd} + \alpha\pi_{dc})] \tag{4}$$

when using strategy $C$ and

$$U_D = (1-p)\,\pi_{dd} + p \cdot [x\,(\pi_{dc} + \alpha\pi_{cd}) + (1-x)\,(1+\alpha)\,\pi_{dd}] \tag{5}$$

when using strategy $D$. Hence, strategy $C$ is optimal for our decision-making individual of Type 1 if and only if $U_C \geq U_D$, or, in more detail, if and only if

$$p \cdot x \cdot (1+\alpha) \cdot [\pi_{cc} + \pi_{dd} - \pi_{cd} - \pi_{dc}] \geq p \cdot \alpha \cdot (\pi_{dd} - \pi_{dc}) + \pi_{dd} - \pi_{cd}. \tag{6}$$

Conversely, strategy $D$ is optimal for our decision-making individual of Type 1 if and only if the reversed weak inequality holds:

$$p \cdot x \cdot (1+\alpha) \cdot [\pi_{cc} + \pi_{dd} - \pi_{cd} - \pi_{dc}] \leq p \cdot \alpha \cdot (\pi_{dd} - \pi_{dc}) + \pi_{dd} - \pi_{cd}. \tag{7}$$

If both inequalities hold, (the right-hand side is equal to the left-hand side), then our decision-maker is indifferent between strategies $C$ and $D$:

$$p \cdot x \cdot (1+\alpha) \cdot [\pi_{cc} + \pi_{dd} - \pi_{cd} - \pi_{dc}] = p \cdot \alpha \cdot (\pi_{dd} - \pi_{dc}) + \pi_{dd} - \pi_{cd}. \tag{8}$$

We will call a conditional altruist who plays $D$ a bitter altruist. Let $q$ be the probability that a randomly drawn individual from the population will play $C$. With the residual probability, $(1-q)$, the opponent is, thus, either a selfish individual or a bitter altruist. In fact, in our model $p \in [0,1]$, represents the type distribution in the population and $q \in [0,1]$ stands for the behavior distribution. Every player has some belief about the proportion of conditional altruists $p$ in the group. This belief about their true proportion in the population is correct and everybody knows that everybody else has the same belief. Each player also has an initial expectation of $q$, the share of cooperative people, *i.e.* how many individuals are actually choosing to play $C$. Let us also note that the condition $p \geq q$ always holds. Whenever $p = q$, all altruists choose to cooperate, and whenever $p > q$ some altruistic players behave non-cooperatively, hence they are bitter altruists.

A further condition to our model is that players do not know the preferences of the individuals they meet. They have to form some expectations about the type of the counterpart they meet and estimate the probability of observing a certain behavior. They cannot be sure that their expectations reflect reality and they are therefore uncertain about the optimal behavior for themselves.

We will call a pair $(p,q) \in [0,1]^2$ a population state.

**Definition 1** *A population state* $(p, q) \in [0, 1]^2$ *is self-fulfilling if and only if* $px = q$*, for some* $x$ *that is optimal given* $(p, q)$*.*

In other words, for a state to be self-fulfilling we require that $q$ is consistent with $p$ in the sense that $px = q$ for some $x$ that is optimal given $(p, q)$. Intuitively, a self-fulfilling population state stands for a situation when a Type 1 individual knows the true proportion of conditional altruists in the society, $p$, and she is correct about how many of them are cooperative, that is her evaluation of $x$ is correct. As long as knowing $(p, q)$ does not lead to the awareness that for a given proportion of conditional altruists, $p$, a certain behavior of the population, $q$, can actually be observed in reality, the chosen combination $(p, q)$ is not self-fulfilling.

Let us now refer back to the inequalities (6) and (7) discussed before. We compiled these conditions to equation (8), the situation when a Type 1 individual is indifferent between her two pure strategies. From this condition and $px = q$, we get the following equation:

$$q(1+\alpha)(\pi_{cc} - \pi_{cd} - \pi_{dc} + \pi_{dd}) = p\alpha(\pi_{dd} - \pi_{dc}) - \pi_{cd} + \pi_{dd}. \tag{9}$$

We now refer back to inequality (6), where Type 1 individual decides to play $C$:

$$q(1+\alpha)(\pi_{cc} - \pi_{cd} - \pi_{dc} + \pi_{dd}) \geq p\alpha(\pi_{dd} - \pi_{dc}) - \pi_{cd} + \pi_{dd}. \tag{10}$$

Finally, for inequality (7) that is Type 1 individual finds strategy $D$ optimal, the following is true:

$$q(1+\alpha)(\pi_{cc} - \pi_{cd} - \pi_{dc} + \pi_{dd}) \leq p\alpha(\pi_{dd} - \pi_{dc}) - \pi_{cd} + \pi_{dd}. \tag{11}$$

# 4 Analysis

## 4.1 Model of Conditional Altruism with Degree of Altruism $\alpha = 1$

Let us consider the following payoff matrix as a baseline:

$$\Pi_0 = \begin{pmatrix} 2 & 0 \\ 3 & w \end{pmatrix} \text{ with } 0 < w < 2. \tag{12}$$

In line with the general case inequality (10), for this particular example the optimal strategy for Type 1 is $C$ if and only if:

$$q(1+\alpha)\,(w-1) \geq w + p\alpha\,(w-3)\,. \tag{13}$$

We now will focus on the case when Type $\alpha$ is fully altruistic, that is she has $\alpha = 1$ in the event she meets her own kind. This implies that a Type 1 player values the payoff another Type 1 player receives as much as her own.

Therefore, for this particular example the optimal strategy for Type 1 is $C$ if and only if:

$$2q\,(w-1) \geq w + p\,(w-3)\,. \tag{14}$$

Let us keep in mind that two types of individuals exist in our population, who are homogeneous in preferences and intentions within their types. The true share of conditional altruists in the society is $p$. Further we assume that the true $p$ is common knowledge in the population. We now distinguish three different cases for different values of $w$: **Case I** with $0 < w < 1$, **Case II** with $w = 1$ and **Case III** with $1 < w < 2$.

In all following graphs, the self-fulfilling population states are shown as dashed lines.

### 4.1.1 Case I

We will now focus on $w \in (0, 1)$. From this condition we can define $q$ as a function of $p$ for the given payoff matrix:

$$q(p) = \frac{(3-w)p - w}{2(1-w)}. \tag{15}$$

The function $q(p)$ specifies that $q$ is consistent with a given $p$, and we note that $q(p)$ is linearly increasing in $p$. For further analysis of the player's behavior we choose a particular value for $w$ from the interval under consideration, $0 < w < 1$, that is, $w = 0.5$. This gives us the following realization of the constraint function $q(p)$:

$$q(p) = 2.5p - 0.5. \tag{16}$$

One can notice $q(0) = -0.5$, and $q(1) = 2$. See *Figure 1* below for the illustration.

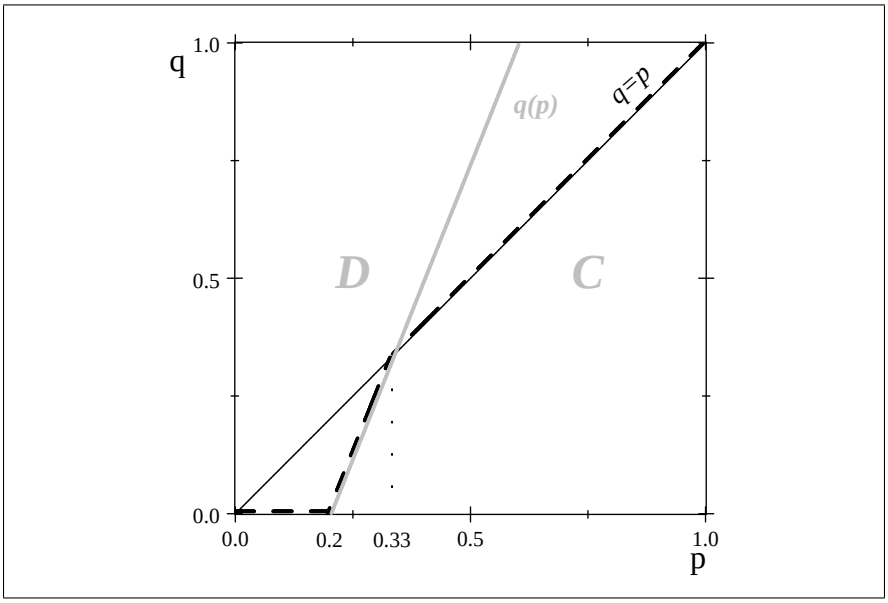


Figure 1: Constraint function $q(p)$ and self-fulfilling population states for $w = 0.5$ and $\alpha = 1$

The line $q(p)$ differentiates between the strategies for a Type 1 player, whether she chooses to cooperate or to defect. The area left to $q(p)$ indicates that the player chooses strategy $D$, whereas the area right to the constraint implies the choice in favor of strategy $C$. Any population state on the line $q = p$ implies that conditional altruists always cooperate and selfish people always defect.

The probability that a conditional altruist decides to be cooperative, $x$:

$$\Pr(C|\text{Type 1}) = \frac{q(p)}{p}. \tag{17}$$

Similarly, the probability that the same player chooses to defect, $(1-x)$:

$$\Pr(D|\text{Type 1}) = 1 - \frac{q(p)}{p}. \tag{18}$$

The probability that a randomly chosen player cooperates:

$$\Pr(C) = p \cdot \Pr(C|\text{Type 1}) = p \cdot \frac{q(p)}{p} = q(p), \tag{19}$$

and using the same logic, probability that a randomly drawn player from the population defects:

$$\begin{aligned} \Pr(D) &= (1-p) + p \cdot \Pr(D|\text{Type 1}) \\ &= (1-p) + p \cdot (1 - \frac{q(p)}{p}) \\ &= 1 - q(p). \end{aligned} \tag{20}$$

Let us keep in mind that $q$ stands for the behavior distribution, or the probability that a randomly chosen individual in the population of $N$ individuals will play strategy $C$, and $q$ should not be mixed up with $x$, the probability of a player to be cooperative given that she is a conditional altruist. But each Type 1 individual knows if $x = 1$ then the best reply for her is $C$ for given $(p, q)$, and vice versa, if $x = 0$ then the best reply for her is to defect. However if $x \in (0, 1)$, optimal strategy is a mixed one. That is, in the population either some Type 1 individuals play $C$ and others play $D$ all the time, or Type 1 individuals shuffle two pure strategies and sometimes a Type 1 player chooses $C$ and sometimes $D$.

By examining different combinations of $p$ and $q$ we will now consider the self-fulfilling population states. We distinguish between 3 intervals for $p \in [0, \frac{1}{5}]$, and $p \in (\frac{1}{5}, \frac{1}{3})$, and $p \in [\frac{1}{3}, 1]$.

Let us probe a real combination for values $p$ and $q$ so that $p \in [0, \frac{1}{5}]$ for $p \geq q$. Let us start with examining the population state $(p, p)$ which represents any point along the diagonal. This population state implies that all conditional altruists are cooperative, that is $x = 1$. However, for this state to be self-fulfilling, given $x = 1$, an individual of Type 1 should actually choose cooperation over defection. A best reply for the given $(p, q)$ can be derived from the constraint $q(p)$ given $p$. This calculation for $q(p)$ shows that for $p$ values in this interval a conditional altruists wants to defect. Cooperation would never be observed. We conclude that the diagonal does not represent self-fulfilling population states for the interval under consideration.

We will now proceed with another population state, namely, any state which satisfies $p \in [0, \frac{1}{5}]$ and $q \in (0, p]$. Such a population state is self-fulfilling if cooperation is played with probability $x$. However, $q(p)$ is negative for $p \in [0, \frac{1}{5}]$ and, thus, condition $px = q$ for given $(p, q)$ cannot be satisfied. These states are no self-fulfilling population states either.

The self-fulfilling population states for $p \in [0, \frac{1}{5}]$ lie along the abscissa. For a population state $(p, 0)$ to be self-fulfilling, it should be optimal for a conditional altruist to choose defection over cooperation when $x = 0$. This is exactly what we observe. In such a population state all conditional altruists use strategy $D$, just as all selfish individuals, and hence $q = 0$. Given that $x = 0$, $q$ is also equal to zero, therefore, $px = q$ for given $(p, 0)$ is satisfied. All conditional altruists are bitter altruists.

Now we will provide similar reasoning for $p \in [\frac{1}{3}, 1]$. If one takes a combination $(p, q)$ from this interval, it is necessarily a combination below or on $q = p$. A population state $(p, 0)$ is self-fulfilling if it is optimal for conditional altruists to choose strategy $D$ when $x = 0$. But an individual who is of Type 1 prefers $C$ over $D$ based on the constraint $q(p)$ for this interval of $p$ and never plays $D$. Therefore, the only self-fulfilling population states are $(p, p)$. This is true indeed, since if $x = 1$, then it is optimal for Type 1 to play $C$, that is the condition $px = q$ is satisfied. In such a population state, all conditional altruists use strategy $C$ and, thus, $q = p$.

The most interesting interval is for $p \in (\frac{1}{5}, \frac{1}{3})$. The self-fulfilling population states lie upon $q(p)$. In a self-fulfilling population state with $0 < q < p$, conditional altruists are indifferent between two pure strategies when $x = \frac{q}{p}$, and this $x$ is optimal if they play strategy $C$ with probability $x$ at each encounter. We take a combination $(p, q)$ which lies above the constraint $q(p)$, see point $A$ in *Figure 2* below. Any combination above the line $q = p$ is not feasible due to $p \geq q$. Further, a Type 1 individual decides to defect whenever the combination of $p$ and $q$ is above $q(p)$ and below $q = p$ and she decides to cooperate if $(p, q)$ lies below the constraint $q(p)$. For any given $(p, q)$ the optimal $x$ values lie within the range $x \in (0, 1)$ for the self-fulfilling population state.

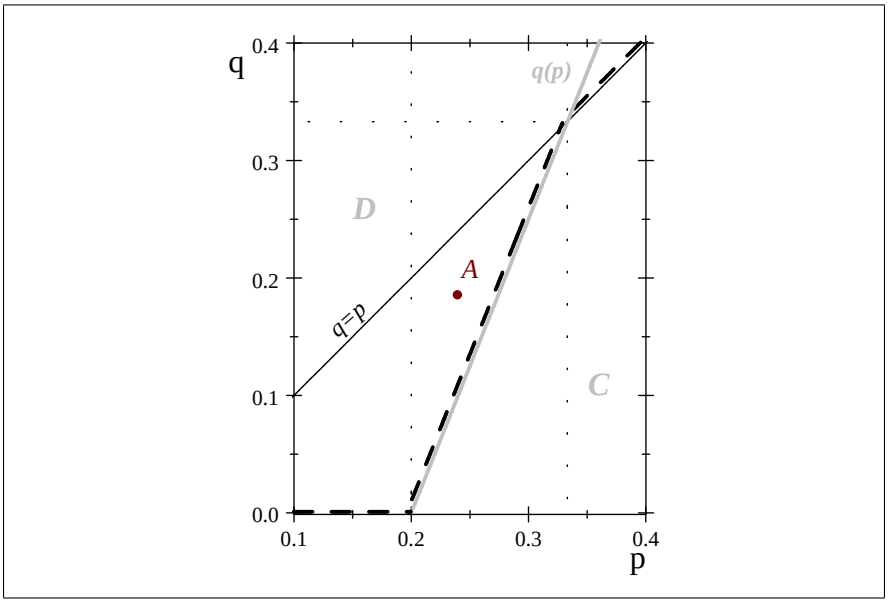


Figure 2: Detailed view of constraint function $q(p)$ and self-fulfilling population states for $w = 0.5$ and $\alpha = 1$

**Notation 1** *We will call a population state $(p, p)$ a favorable or cooperative self-fulfilling population state and, similarly, a population state $(p, 0)$ a unfavorable or defective self-fulfilling population state.*

### 4.1.2 Case II

In this case, $w = 1$ and the constraint function $q(p)$ is represented by a vertical line due to the payoffs chosen in the baseline matrix (12) and is described by

$$p = \frac{1}{2}. \tag{21}$$

The illustration is shown in the graph below:

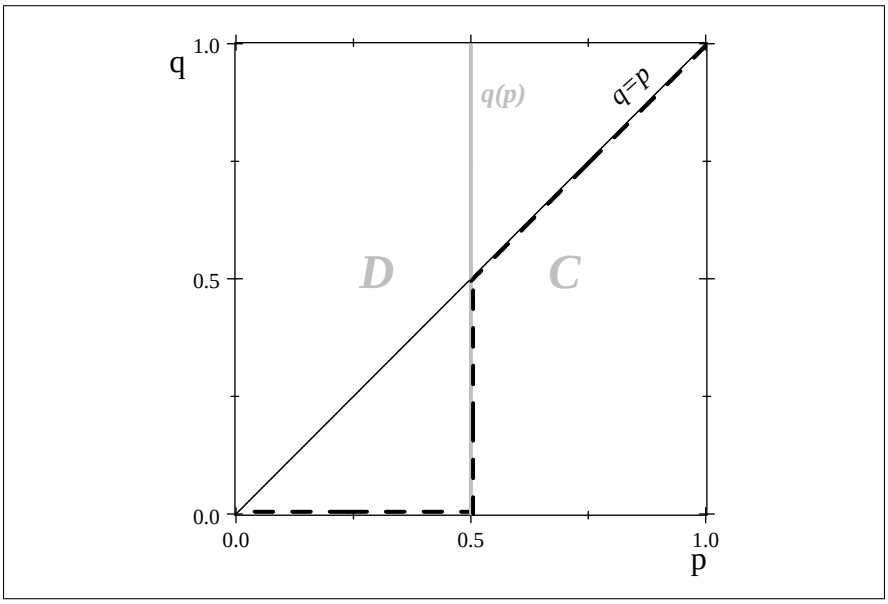


Figure 3: Constraint function $q(p)$ and self-fulfilling population states for $w = 1$ and $\alpha = 1$

Now we turn to the discussion of self-fulfilling population states. As before there is a need to distinguish among several intervals for $p$: $p \in [0, \frac{1}{2}]$, $p = \frac{1}{2}$ and $p \in [\frac{1}{2}, 1]$, where we will examine the various population states explicitly.

For $p \in [0, \frac{1}{2}]$, according to the constraint $q(p)$ and independent from the value of $q$, Type 1 individuals always decide to play strategy $D$. For any $p < \frac{1}{2}$, no conditional altruist can commit herself to play $C$, hence $q = 0$. For $px = q$ to hold, $x = 0$, therefore only combinations of $(p, q)$ on the abscissa are self-fulfilling. For $w = 1$ and $p < \frac{1}{2}$, all conditional altruists are bitter altruists.

For $p = \frac{1}{2}$ there is a range of $q$ values, $q \in [0, \frac{1}{2}]$, for which players' expectations are self-fulfilling. Similar to **Case I,** the self-fulfilling population states lie upon $q(p)$. Type 1 individual chooses to play a mixed strategy, because she is indifferent between her pure strategies. This implies that strategy $C$ should be optimal to play with a probability $x$ for a given population state $(\frac{1}{2}, q)$, $0 < q < p$, to be self-fulfilling. Any population state $(\frac{1}{2}, q)$ is self-fulfilling if optimal $x \in (0, 1)$ is such that $x = 2q$ given $(\frac{1}{2}, q)$. For $p = \frac{1}{2}$ any $q \in [0, \frac{1}{2}]$ is feasible, therefore, there exist many self-fulfilling population states for this $p$ value.

A population state $(p, p)$ is self-fulfilling if it is optimal for conditional altruists to choose strategy $C$ when $x = 1$. In such a population state, all conditional altruists use strategy $C$ and, thus, $q = p$. For $p \in [\frac{1}{2}, 1]$ the self-fulfilling population states lie on $q = p$ exactly due to this reasoning. For any $p \in (\frac{1}{2}, 1]$ and any $q$, according to the constraint $q(p)$, Type 1 players decide to cooperate. This implies for $x = 1$, $p = q$, therefore only combinations of $(p, q)$ on

the diagonal through the origin are self-fulfilling. Within this interval, no other population state can be found to be self-fulfilling.

### 4.1.3 Case III

In this case the relevant $w \in (1, 2)$. Hence for $\alpha = 1$ and the chosen payoffs of the baseline matrix (12), the constraint function is given by the following equation:

$$q(p) = \frac{(w-3)\,p + w}{2\,(w-1)}. \tag{22}$$

The function $q(p)$ specifies the $q$ value that is consistent with a given $p$ value and is linearly decreasing in $p$. For further analysis of the player's behavior we choose a particular value for $w$ from the interval under consideration that is, $w = 1.5$. This gives us the following realization of the constraint function $q(p)$:

$$q(p) = 1.5 - 1.5p. \tag{23}$$

One can notice $q(0) = 1.5$, and $q(1) = 0$.

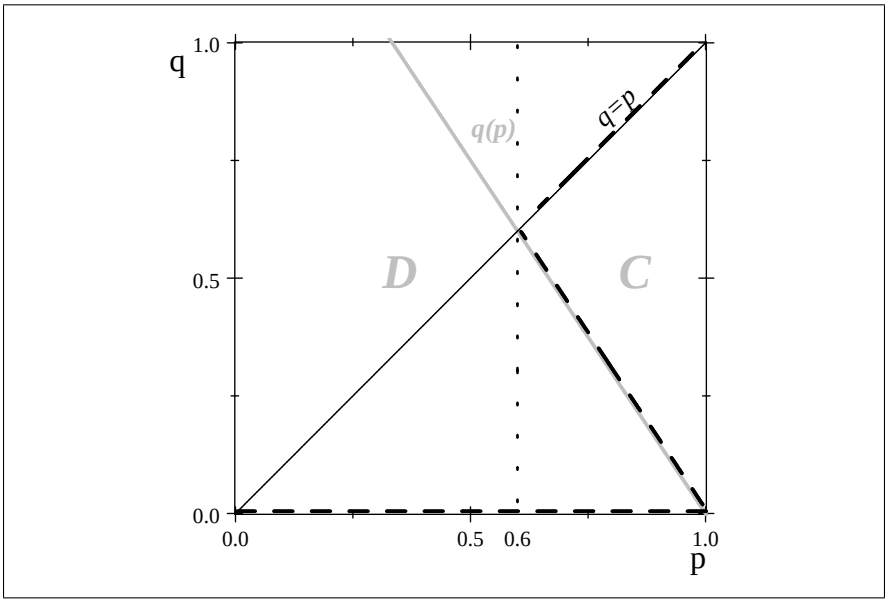


Figure 4: Constraint function $q(p)$ and self-fulfilling population states for $w = 1.5$ and $\alpha = 1$

This is the most interesting case, where three self-fulfilling population states are simultaneously feasible for one particular value of $p$. The distinguished intervals are $p \in [0, \frac{3}{5})$, which has only one population state for a given $p$, and $p \in [\frac{3}{5}, 1]$, which has three possible self-fulfilling population states simultaneously for a given $p$.

For interval $p \in [0, \frac{3}{5})$, the argumentation is the same as in the previous two cases. The self-fulfilling population states lie along the abscissa. For a population state $(p, 0)$ to be self-fulfilling it should be optimal for a conditional altruist to choose defection over cooperation when $x = 0$. In such a population state all conditional altruists use strategy $D$, just as all selfish individuals, and hence $q = 0$. Given that $x = 0$, $q$ is also equal to zero, therefore, $px = q$ for given $(p, 0)$ is satisfied. And again all conditional altruists are bitter altruists.

We consider now the second interval. For $p \in [\frac{3}{5}, 1]$ there exist two additional self-fulfilling population states: $q(p)$ and $q = p$. The occurrence of three possible self-fulfilling population states at the same time is due to the decreasing $q(p)$. We can confirm this issue, when we test for several combinations of $(p, q)$ within this interval. If one takes a combination $(p, q)$ from this interval this is necessarily a combination below or on $q = p$. A population state $(p, 0)$ is self-fulfilling if it is optimal for conditional altruists to choose strategy $D$ when $x = 0$. And indeed, an individual who is of Type 1 decides to play $D$ based on the constraint $q(p)$ for $x = 0$. Likewise, in a self-fulfilling population state with $0 < q < p$, conditional altruists are indifferent between the two pure strategies when $x = \frac{q}{p}$, and this $x$ is optimal if they play $C$ with probability $x$. For $p \in [\frac{3}{5}, 1]$, Type 1 individual decides to cooperate whenever the combination of $p$ and $q$ is above $q(p)$ and below $q = p$, and she decides to defect if $(p, q)$ lies below the constraint. For any given $(p, q)$ the optimal $x$ values lie within the range $x \in (0, 1)$ for the self-fulfilling population state. And, thirdly, there are self-fulfilling population states in $(p, p)$, too. This can be substantiated, because it is true that for $x = 1$ it is optimal for Type 1 to play $C$, that is the condition $px = q$ is satisfied and all conditional altruists use strategy $C$, thus $q = p$. Hence, we can conclude that for $w \in (1, 2)$ and $p \in [\frac{3}{5}, 1]$, there exist three self-fulfilling population states.

## 4.2 Model of Conditional Altruism with Degree of Altruism $\alpha < 1$

We assume that Type 1 players value their opponent's payoff as much as their own, therefore $\alpha$ is supposed to be one. However, this assumption does not seem to hold in reality. For example, for the behavior of siblings, Hamilton's rule offers a more appropriate value for $\alpha$ being around $\frac{1}{2}$ (*e.g.* Alger and Weibull (2008), Stark and Wang (2004) and Hamilton (1964)).

Until now there is no agreement among researchers about the degree of altruism for a general population. To be as broad as possible, we suggest that a reasonable value is $\alpha \in (0, 1)$. Therefore, we now analyze the impact of a decrease in $\alpha$, *i.e.*, $\alpha < 1$. We consider different $\alpha$ values and analyze their impact on the constraint function $q(p)$. We assume $0 < w < 2$ and proceed with the three different cases defined in the previous section: **Case I** with $0 < w < 1$, **Case II** with $w = 1$ and **Case III** with $1 < w < 2$.

The population still consists of selfish players and conditional altruists. Selfish individuals remain Type 0, with a population share of $(1 - p)$ and the dominant behavior strategy to defect. To differentiate from the previous analysis, the conditional altruist belongs to Type $\alpha$. It is assumed that at a given point in time only one value of $\alpha$ exists, *i.e.* it is not possible that Type $\alpha$ players have different $\alpha$ values simultaneously, but the players are homogeneous within their types. Therefore, only $\alpha = 0$ for Type 0 players and one other value of $\alpha$ for Type $\alpha$ players are present at the same time.

Generally for all three cases, it can be said that when $\alpha$ decreases, the expanse for defection increases, and the area where Type $\alpha$ players choose to cooperate decreases. This seems to be reasonable and in line with theory, as with a decreasing $\alpha$, the players decrease the weight they put on their opponent's payoff and therefore, they care less about the value their opponent receives and more about their own payoff. The smaller the area where cooperation is still possible, the higher the probability that Type $\alpha$ individuals behave as bitter altruists.

### 4.2.1 Case I

The first case takes into consideration the interval $w \in (0,1)$ and we choose $w = 0.5$ for a demonstration purpose. As we can see in the graph below, the value of $\alpha$ is responsible for the slope of the constraint, *i.e.*, with a lower $\alpha$, the curve gets flatter rotating clockwise around point $(-\frac{1}{5}, -1)$. The smaller $\alpha$, the smaller the area where Type $\alpha$ individuals decide to cooperate according to the constraint $q(p)$. For $w = 0.5$, the lowest value for $\alpha$, where cooperation is still feasible, is for $\alpha > 0.2$. For this value, the constraint intersects with the unit-square only in one point, $(1, 0)$. In *Figure 5* below, the constraint $q(p)$ is plotted for $w = 0.5$ and different values of $\alpha$.*

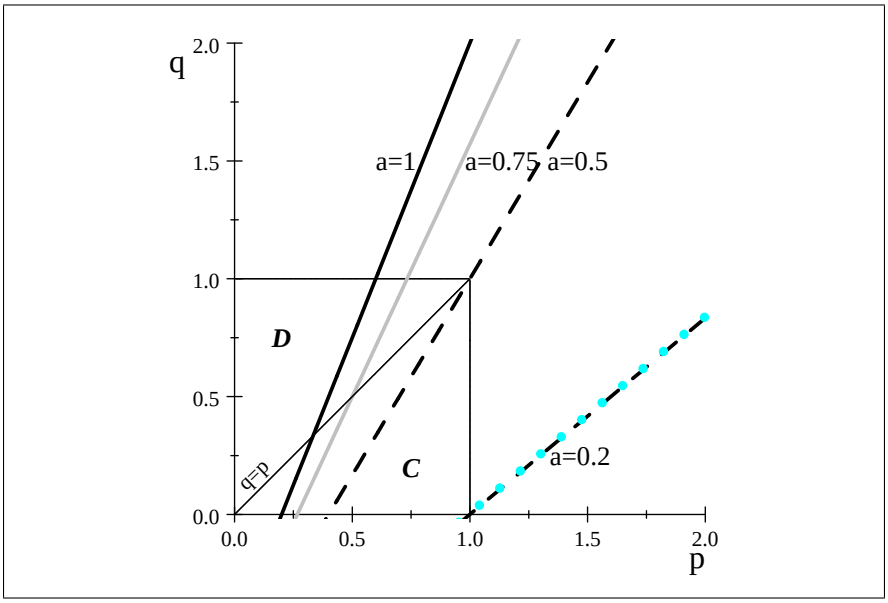


Figure 5: Constraint function $q(p)$ for $w = 0.5$ and different $\alpha$ values

For comparison with $\alpha = 1$ we choose $\alpha = 0.4$. The graph of the constraint for $w = 0.5$ and $\alpha = 0.4$ can be seen in *Figure 6* below.

*All "a" in the graph are intended to be $\alpha$

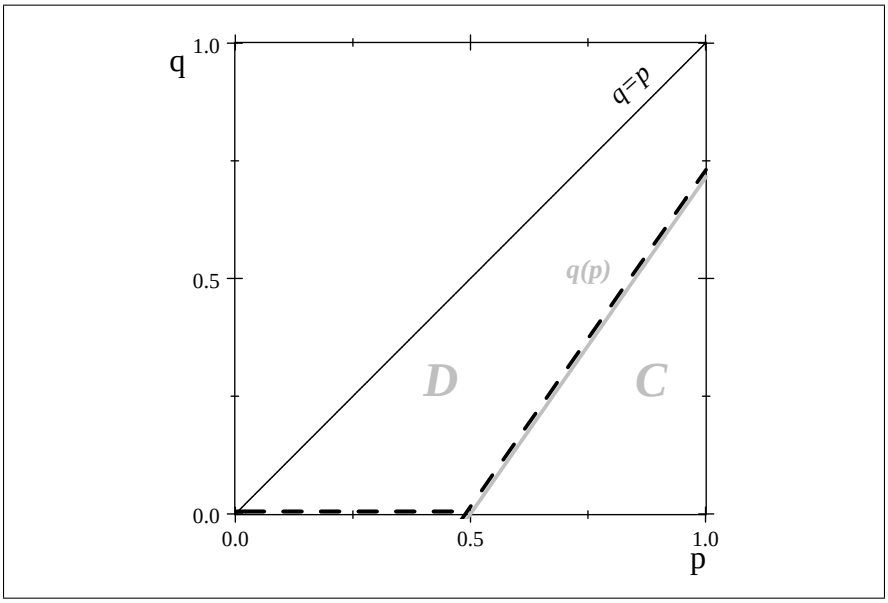


Figure 6: Constraint function $q(p)$ and self-fulfilling population states for $w = 0.5$ and $\alpha = 0.4$

In *Figure 6*, the constraint $q(p)$ intersects with the unit-square twice, once at $\frac{1}{2}$ on the abscissa and again in $(1, \frac{5}{7})$. As the second intersection is on the right border of the unit-square, there is no intersection for $q = p$ and $q(p)$ within the area under consideration. Hence, at any given $p$, for $\alpha = 0.4$, there always exist bitter altruists in the population, *i.e.* conditional altruists who decide to defect, and $q \in \left[0, \frac{5}{7}\right]$. Because $q < p$, $x = 1$ is never optimal for any $p$, and $q = p$, as a self-fulfilling population state can never be reached. The intersection of $q(p)$ with $p = 1$ at $(1, q < 1)$ is true for all $q(p)$ with $w = 0.5$ and $0.2 \leq \alpha < 0.5$.

Furthermore, the area where Type $\alpha$ players decide to cooperate decreases significantly compared to the situation when $\alpha = 1$. In addition, there are only two different self-fulfilling population states possible, depending on the value of $p$. For $p \in [0, \frac{1}{2}]$, the self-fulfilling population states are located on the abscissa, where $x = 0$ and it is optimal for conditional altruists to choose strategy $D$, hence $q = 0$. For $p \in (\frac{1}{2}, 1]$, the population states can be found on the constraint, where the individuals are indifferent between the two pure strategies.

### 4.2.2 Case II

This is the case for $w = 1$. For $\alpha = 1$, the constraint function $q(p)$ is $p = \frac{1}{2}$ and divides the relevant area into two equal parts. In *Figure 7* below the constraint $q(p)$ is plotted for $w = 1$ and different values of $\alpha$.*

*All "a" in the graph are intended to be $\alpha$

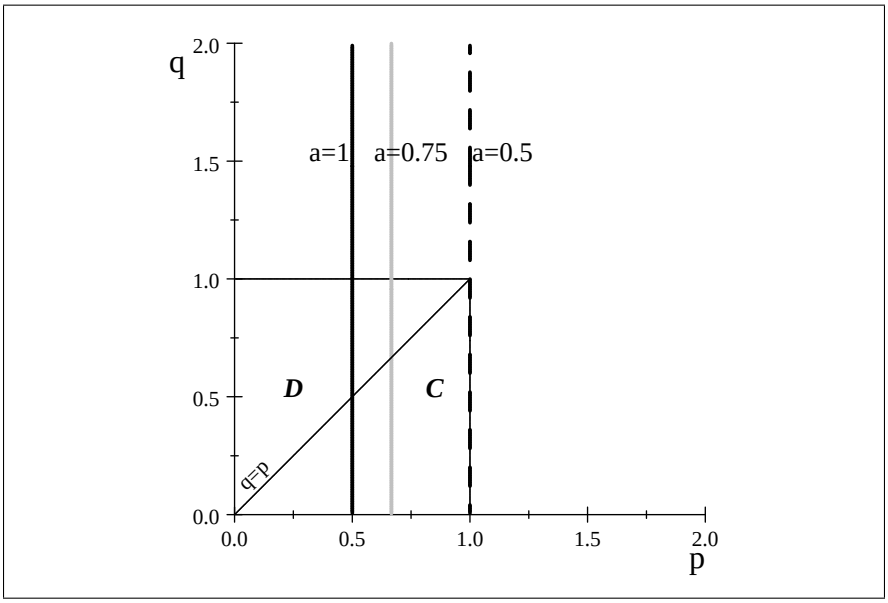


Figure 7: Constraint function $q(p)$ for $w = 1$ and different $\alpha$ values

Cooperation is only feasible for $\alpha \in [\frac{1}{2}, 1]$. For any $\alpha < 0.5$ and $w = 1$, the constraint fails to intersect with the feasible area for our model. Comparing to case $w = 1$ and $\alpha = 1$, when $\alpha$ decreases, the constraint remains a vertical line, $p = \frac{1}{2\alpha}$, and moves to the right, implying that both the interval for $p$, where the self-fulfilling population states are located on the abscissa, and the interval for $q$, where the self-fulfilling population states are along $p = \frac{1}{2\alpha}$, increase, while the interval for $p$ where the self-fulfilling population states are located on $q = p$ decreases. This can be observed in *Figure 8* below, where the constraint function $q(p)$ and the self-fulfilling population states are plotted for $w = 1$ and $\alpha = 0.75$.

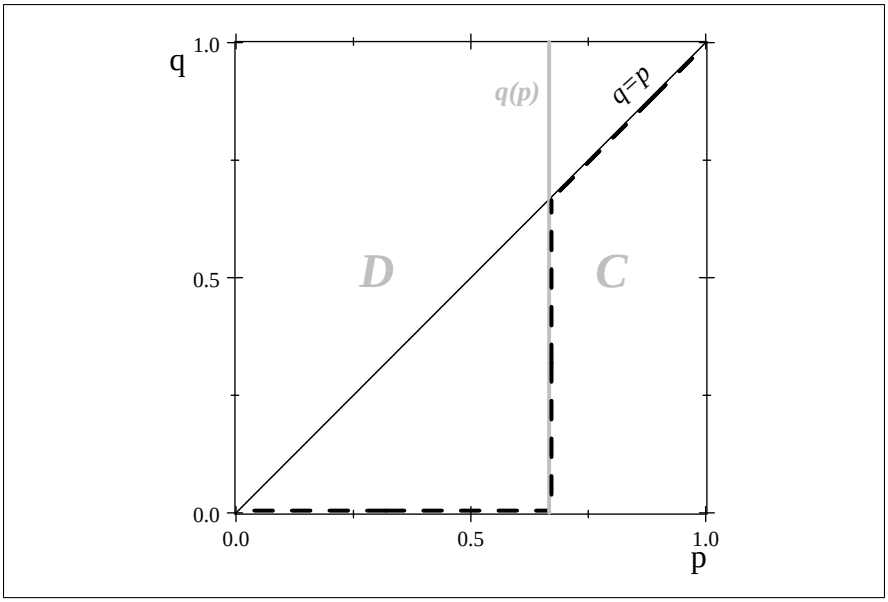


Figure 8: Constraint function $q(p)$ and self-fulfilling population states for $w = 1$ and $\alpha = 0.75$

### 4.2.3 Case III

In **Case III** $w \in (1, 2)$ and we choose $w = 1.5$. Type $\alpha$ players decide to defect at any $(p, q)$ left to the constraint $q(p)$, and they want to cooperate at any $(p, q)$ to the right to the constraint. In *Figure 9* below the constraint $q(p)$ is plotted for $w = 1.5$ and different values of $\alpha$.* We can see that with a lower $\alpha$, the curve is getting flatter again, rotating this time anti-clockwise around the point $(-1, 3)$. The lowest value for $\alpha$, where cooperation is still feasible for $w = 1.5$, is for $\alpha = 0.5$. Under these conditions, the constraint intersects only in one point, $(1, 1)$, with the unit-square. In this point all people in the population belong to Type $\alpha$ and all decide to behave cooperatively according to the constraint $q(p)$, thus, $q = p = 1$. Again the feasible range of $\alpha$ values depends on $w$, and generally for $w = 1.5$ the only $\alpha$ possible for cooperation are $\alpha \in [\frac{1}{2}, 1]$.

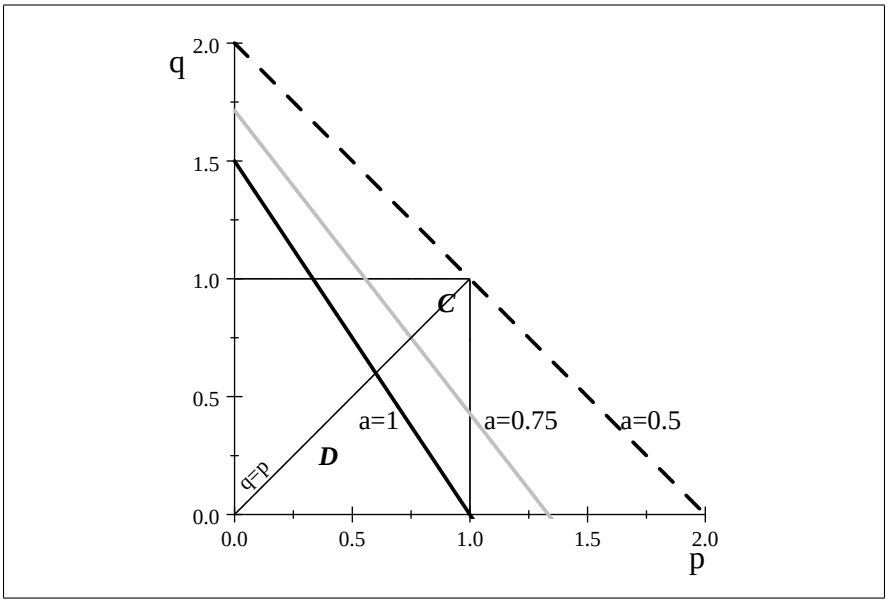


Figure 9: Constraint function $q(p)$ for $w = 1.5$ and different $\alpha$ values

We now consider one particular situation from this interval: $w = 1.5$ and $\alpha = 0.75$. The constraint function and the self-fulfilling population states are plotted in *Figure 10* below.

*All "a" in the graph are intended to be $\alpha$

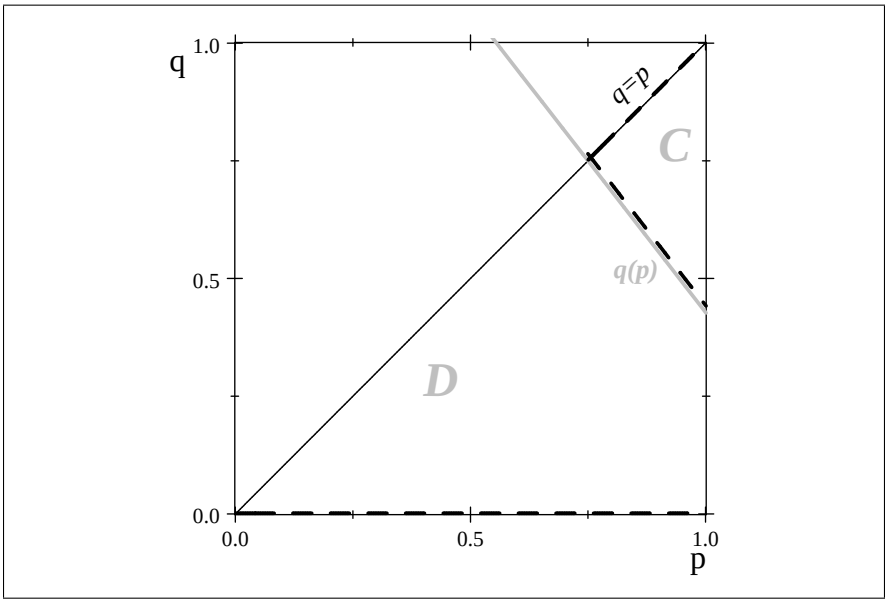


Figure 10: Constraint function $q(p)$ and self-fulfilling population states for $w = 1.5$ and $\alpha = 0.75$

The constraint $q(p)$ crosscuts the unit-square in the points $(\frac{5}{9}, 1)$ and $(1, \frac{3}{7})$, and intersects with the diagonal line $q = p$ at $(\frac{3}{4}, \frac{3}{4})$. The area, where Type $\alpha$ players choose to cooperate, is again much smaller compared to the situation when $\alpha = 1$.

Regarding the self-fulfilling population states, this situation is not very different from our initial one with $\alpha = 1$. For each $p$ within the interval $p \in [0, \frac{3}{4})$, there is only one feasible self-fulfilling population state: $q = 0$. For $p \in [\frac{3}{4}, 1]$ the Type $\alpha$ individuals could end up in three different self-fulfilling population states. It is possible for the population to reach $q = p$, the favorable state, where all Type $\alpha$ decide to play $C$; also to attain the condition where Type $\alpha$ individuals decide to play a mixed strategy; or to arrive at the unfavorable population state, where all Type $\alpha$ are bitter altruists and decide to defect as Type 0 individuals do.

Generally one can conclude that for **Case II** and **III**, $1 \leq w < 2$, any $\alpha$ value below 0.5 is not feasible, however, for **Case I**, $0 < w < 1$, even lower values of $\alpha$ are feasible. This seems to be consistent with theory; $w$ stands for the payoff a player receives if both players defect. The lower this payoff, compared to the other payoffs in the payoff matrix, the more cooperative behavior is valued. Although, the degree of altruism of Type $\alpha$ player might be very low, there still exist $p$ values, for which she might prefer cooperation over defection. This implies that for $0 < w < 1$ at a given $p$, Type $\alpha$ players are able to attribute a lower $\alpha$ to the payoff their opponent receives, and still could decide to cooperate, compared to $1 \leq w < 2$. We can therefore interfere, that given $w$, there is only a certain range of $\alpha$ values feasible.

# 5 Stability

In this section we address the issue of stability and selection of the self-fulfilling population states. We will focus on the most interesting case in our opinion. This is **Case III** with $\alpha = 1$ (see *Section 4.1.3* **Case III**). Once again, there are three possible self-fulfilling population

states in the case under consideration: favorable $(p, p)$, unfavorable $(p, 0)$ and a population state $(p, q(p))$.

To examine the stability of the self-fulfilling population states, we decide to introduce noise or mutations into our model. One natural way is to introduce newcomers into the population (Kandori *et al.* (1993)). These newcomers possess limited knowledge about the population. We assume that they know the true type distribution, $p$, in the society, but they do not know which behavior $q$ to expect. In other words, they are unaware about the current population state. In order to form a belief about $q$, they can sample some individuals in the population without replacement and observe their current behavior. Based on this information, newcomers decide on their strategy for the next stage game. We assume for simplicity that the strategy once chosen cannot be changed. However, it is obvious that if a newcomer's sample was not representative of the overall population, it might lead her to choose the wrong strategy. The newcomer can be fortunate and obtain a representative sample of the population, but if not, then she can cling to the strategy she would not choose, had she observed true $q$.

So we are interested to know whether it is actually possible for such newcomers to disturb the current population state. Namely, if we refer to the case under consideration, given the system is in a cooperative self-fulfilling population state, we are interested to know whether defecting Type 1 newcomers would be able to affect the system. In other words, can the dominant strategy for all Type 1 players in the population become to defect rather than to cooperate, moving the system to the $(p, 0)$ self-fulfilling population state?

The answer to this question is that such a change depends on the initial $p$ as well as on the procedure of initial $q$ estimation by the newcomers. We are asking, whether a change of the existing population state would be possible, if a series of newcomers have unrepresentative samples? If this were the case which self-fulfilling population state would be the final state, hence, the absorbing one? Is the existing population state strong enough to withstand the mutations, represented by the injection of newcomers? We argue that in fact, this is not the case and it can be disturbed.

With this question, we will now address each self-fulfilling population state of **Case III** for $p$ values, where three self-fulfilling population states are possible, and $\alpha = 1$.

**Proposition 1** *Both cooperative and defective self-fulfilling population states are stable in a certain limited sense: small perturbations in q-estimates do not lead the population away. However, for a given p the defective self-fulfilling population state is an absorbing state, that is, once the system enters the absorbing state it can never leave it.*

The easiness of achieving the absorbing state $(p, 0)$ depends on the initial population state, on how many players are sampled as a proportion of the population, and on the true proportion of conditional altruists. The system can start in the favorable self-fulfilling population state, but over time, a switch to the unfavorable one is possible. However, the system cannot be disturbed when locked in the unfavorable self-fulfilling population state: the sampling procedure will always ensure that the newcomer cannot choose the wrong strategy when the current population state is close enough to it. Particularly, the rational newcomer will never choose to cooperate since her sample will always be representative of the population behavior.

**Proposition 2** *A self-fulfilling state* $(p, q(p))$ *with* $0 < q < p$ *is unstable.*

The constraint $q(p)$ serves as a 'divider' between $q$ values that will drift up (those above the $q(p)$ curve), and those that will drift down (those below the $q(p)$ curve). Hence, the slightest shift in individuals' $q$-estimate will lead away from this population state.

**Proposition 3** *The larger the difference* $(p - q(p))$*, the less unstable is the population state* $(p, p)$*.*

The higher $p$, the true proportion of conditional altruists, the higher is the stability of the potentially cooperative population state. As $p$ increases, the system needs more time to get away from the cooperative self-fulfilling population state to the absorbing state.

Please note, *Proposition 3* itself is a general result, although it is only partially valid for the simulations we employ due to a particular realization of the parameters we use to perform the simulations. We explain this point in greater detail later on.

We will now perform a series of computer simulations and demonstrate the essence of the propositions above.

# 6 Simulations

As we already mentioned, the stability issue is studied through the introduction of mutations to the system. We are now conducting a series of simple computer simulation experiments in order to observe the dynamics of social interaction for our model. The method we choose to perform the simulations is not the only one applicable to this situation, however, we made the decision based on simplicity reasons. The results are presented in the *Appendix*.

Suppose there is a population of $N$ players in which $p$ is common knowledge. Let us assume that at each period there is a newcomer to the population, who replaces an established player. Newcomers appear in the system one at a time period. The new player knows true $p$, and believes in a certain $q$, but she has some doubts about her estimation of $q$. Initially, a newcomer has to gather and analyze information about the strategies chosen by others and form her best reply for the next stage game.

The newcomer tries to estimate $q$ based on some sample of players from the population. Suppose that she samples $k$ players from the population without replacement. The sampling procedure itself is not part of the game, instead the sample helps her to estimate $q$. However, she knows that her sample is possibly not representative, but this is the only means she can rely on. The player is interested in comparing her observed $q$ value with the threshold she uses to decide on her strategy. This threshold value of $q$ can be calculated by estimating $q(p)$ for the known $p$ value, given $\alpha$ and $w$. Let $q_{thres}$ stand for this threshold value. The decision rule is the following: if the observed $q > q_{thres}$, then the player chooses to cooperate, and if $q < q_{thres}$, then she prefers to defect.

Let us keep in mind that we assume the strategy for a newcomer, once chosen, cannot be changed. Another assumption is that a newcomer is of the same type as the player she replaces and this is common knowledge. This implies that the type distribution $p$ in the population remains the same, however, we are interested in how, if at all, the behavior

distribution $q$ can change over time in the group. Since selfish individuals have only one strategy to choose from, we will replace only conditional altruists.

Prior to discussing the results from simulations for each proposition, we describe the initial data we used. As we have already said, we produce simulations for **Case III** discussed in *Section 4.1.3*. We employ the favorable self-fulfilling population state, $p = q$ and $p = 0.7$. For performing the simulations we take $w = 1.5$ and $\alpha = 1$. Furthermore, we decide to consider a small group of people, namely, $N = 10$. Regarding the sampling procedure, we take $k = 3$.

Now we will explain each proposition stated above with the corresponding figures in the *Appendix*.

*Proposition 1* tells us that, no matter the initial population state, the system always achieves the absorbing state, which is the unfavorable self-fulfilling population state. With our particular simulation technique, this general result is limited for $p < 0.8$. This is due to how we conduct the simulations. Our mutations are limited to a single individual, which cannot be divided into smaller pieces and is, therefore, a rather high perturbation for the population of ten individuals. This special situation will be explained in detail later on.

As for the stability issue, one can observe from *Figure 1* and *Figure 2* in the *Appendix* that the time required for the system to achieve the unfavorable self-fulfilling population state depends on the initial state. In the particular examples shown, it takes up to 55 periods for a system to get from the favorable to unfavorable self-fulfilling population state, but only about five periods if the initial state was close enough to the unfavorable self-fulfilling population state. Basically we are not able to show with our simulations the stability of the self-fulfilling state $(p, p)$, since our mutations are not small in fact. However, referring back to *Figure1* in the *Appendix,* one can notice that the system actually floats around $(0.7, 0.7)$ for a while, which may indicate the stability of the favorable population state.

*Proposition 2* continues on the differences of stability of the various self-fulfilling population states: pure self-fulfilling population states, $(p, p)$ and $(p, 0)$, are more stable than $(p, q(p))$. *Figures 3* and *4* in *Appendix* reflect these findings. Namely, these are the graphs which show the time needed to reach the absorbing state $(p, 0)$, which is $(0.7, 0)$ in our example. For the initial population state above the constraint $q(p)$, at first, the system relocates to the favorable self-fulfilling population state and settles around it for a while, but, eventually, moves downward to its absorbing state, also floating a while at the unstable population state on $q(p)$ (intervals in the graph with a flat slope). *Figure 4* in *Appendix* confirms that for the initial population state 'below' but close enough to the self-fulfilling population state $(p, q(p))$, the system cannot shift persistently towards the favorable self-fulfilling population state. It can only move downwards and finally arrive at the unfavorable self-fulfilling population state.

To be able to present *Proposition 3* better in the simulations, we decide on three different values of $p$, $p = 0.6$, $p = 0.7$ and $p = 0.8$. Each simulation is run for hundred periods, thousand times. We should clearly identify that for our particular simulation technique, there are two possible absorbing states for $p \geq 0.8$, $(p, p)$ and $(p, 0)$. However, the general result is that the only absorbing population state is the unfavorable self-fulfilling population state. Taking this into consideration, we can now refer to *Figure 5, Figure 6* and *Figure 7* in the *Appendix* These are frequency distribution charts for the average time of the system to the absorbing state, based on the thousand times repetition of each simulation, for each of the chosen $p$ values. We observe that the distribution has less positive skew for higher

$p$, implying that more time is required on average to achieve the unfavorable self-fulfilling population state. For $p = 0.8$ we observe that the system stays in the favorable self-fulfilling population state. The system gets more difficult to be disturbed, the more individuals of Type 1 exist.

The situation for $p = 0.8$ is rather special in our simulations (*Figure 7* in the *Appendix*). In fact, for our simulations we can say that for $p \geq 0.8$, the absorbing state is the favorable population state, given that the initial population state is $(p, p)$. The reasoning is the following: for the newcomer, it is not possible to observe enough defective players during sampling so that she decides to defect herself. We construct the sampling procedure in such a way that the belief is formed quiet correctly in the beginning and the initial population state cannot be that much disturbed to be abandoned.

Basically, according to our sampling procedure of $k = 3$, a newcomer is restricted to the four $q$-estimates, $q = \left\{0, \frac{1}{3}, \frac{2}{3}, 1\right\}$. While if $p = 0.8$ then only $q = \left\{\frac{1}{3}, \frac{2}{3}, 1\right\}$ can be obtained through sampling. The lowest potentially observable value of $q$, $q = \frac{1}{3}$, is not low enough for the newcomer to choose the defective strategy. Namely, the constraint $q(p)$, for the considered $p$, ensures that this player would always be cooperative. Once again, this implies that, *e.g.* when $p = 0.9$, the newcomer must sample three defective players in order to switch to strategy $D$ herself. Obviously, this is not possible, if the population starts out from the favorable population state of $p = 0.9$. By contrast, if $p = 0.7$, then a sample of two defectors and one cooperator is sufficient to lead her to switch to strategy $D$ herself.

To keep in mind, generally there is only one absorbing state, the unfavorable self-fulfilling population state, no matter the initial population state. Then, the most important conclusion to draw is that even though the entire population might consist of conditional altruists or the majority of them are conditional altruists, small mutations may finally lead to the unfavorable self-fulfilling population state, though every one would prefer to stay in the initial one. We are aware that our simulations are rather simple to show this general result. We believe that a richer and more sophisticated simulation would confirm the result (one can think of a simulation with an increased sample size up to $N = 100$ keeping the sampling technique of $k = 3$ in order to allow for smaller perturbations). We leave it out for further research.

# 7 Conclusion

With the model we developed, we are able to show that the behavior of conditional altruists, those who care about people with similar values, depends on the share of conditional altruists in the population, expectations about individuals' behavior, the assigned payoffs, and the degree of altruism these individuals possess. We conclude that, despite favorable initial conditions for the cooperation in the population, the defective strategy may become dominant for every individual. Additionally, regarding the stability of the self-fulfilling population states, we argue that the only possible rational choice for the individuals in the long-run is defection, which is in line with Prisoners' Dilemma outcome.

Moreover, by taking into consideration different values for the degree of altruism, $\alpha \in (0, 1]$, we generalize and validate our model. We show that the likeliness for people actually behaving cooperatively increases with $\alpha$.

We understand that our model is extremely simplified and is certainly not able to rep-

resent reality tightly. Nevertheless, we consider it as a stage for further understanding of the nature of humans' cooperation or noncooperation. We consider further research by introducing players with different degrees of altruism, which would allow higher differentiation of individual's preferences and hence reflect the real world more closely. Furthermore, the design of the simulations for our theoretical model could be improved. And, additionally, we cannot overestimate the value of conducting a real-life experiment for the purpose of assessing both conclusions and the external validity of the model.

Eventually, let us round off with an apt quotation, which not only serves as a keynote for our work but also conveys the aptitude for the relevant implications of the topic we discuss:

> "there can be no true understanding of social organization without a clear understanding of economic arrangements; that there is no clean separability of institutions into economic and noneconomic; and that institution including firms and markets, both affect and are affected by values" (Ben-Ner and Puttermann (1998)).

# References

Alger, I. and Weibull, J. (2008), "The Fetters of the Sib: Weber Meets Darwin." Working paper, Stockholm School of Economics.

Andreoni, J. (1988), "Why Free Ride? Strategies and Learning in Public Goods Experiments." *Journal of Public Economics* **37**(3), 291-304.

Andreoni, J. and Miller, J. (1998), "Giving According to GARP: An Experimental Test of the Rationality of Altruism." Workingpaper, University of Wisconsin and Carnegie Mellon University.

Ben-Ner, A. and Puttermann L.(1998), "Values and Institutions in Economic Analysis." In Ben-Ner, A. and Puttermann L. (eds.), *Economics, Values, and Organization.* Cambridge: Cambridge University Press.

Cronson, R. T. A. (2007), "Theories of Commitment, Altruism and Reciprocity: Evidence from Linear Public Goods Games." *Economic Inquiry* **45(**2), 199–216.

Falk, A., Fehr, E. and Fischbacher, U. (2005), "Driving Forces Behind Informal Sanctions." *Econometrica* **73(**6), 2017–2030.

Falk, A. and Fischbacher, U. (2001), "A Theory of Reciprocity." CESifo Working Paper No. 457, Center for Economic Studies & Ifo Institute for Economic Research, Munich.

Fehr, E. and Fischbacher, U. (2003), "The nature of human altruism." *Nature* **425**(6960), 785-791.

Fehr, E. and Gächter, S. (2002), "Altruistic Punishment in Humans." *Nature* **415**(1), 137-140.

Fehr, E. and Schmidt, K. M. (1999), "A Theory of Fairness, Competition and Cooperation." *Quarterly Journal of Economics* **114**(3), 817-868.

Fehr, E. and Schmidt, K. M. (2001), "Theories of Fairness and Reciprocity - Evidence and Economic Applications." Working Paper No. 75, University of Zurich.

Fershtmann, C. and Weiss, Y. (1998), "Why Do We Care What Others Think About Us?" In Ben-Ner, A. and Puttermann L. (eds.), *Economics, Values, and Organization.* Cambridge: Cambridge University Press.

Fischbacher, U., Gächter, S. and Fehr, E. (2001), "Are people conditionally cooperative? Evidence from a public goods experiment." *Economics Letters* **71**(3), 397 –404.

Güth, W., Schmittberger, W. and Schwarze, B. (1982), "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization* **3**(4), 367-88.

Hamilton, W. D. (1964), "The Genetical Evolution of Social Behavior. Parts I and II." *Journal of Theoretical Biology* **7**(1), 1–52.

Jacobsson, F., Johannesson, M. and Borgquist, L. (2007), "Is Altruism Paternalistic?" *The Economic Journal* **117**(520), 761–781.

Kandori, M., Mailath, G. J., and Rob, R. (1993), "Learning, mutation and long run equilibria in games." *Econometrica* **61**(1), 29-56.

Kolm, S.-C. (2006), "Introduction to the Economics of Giving, Altruism and Reciprocity." In Kolm, S.-C. and Ythier, J. M. (eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity, Volume 1.* Elsevier B.V..

Konow, J. (2006), "Mixed Feelings: Theories and Evidence of Warm Glow and Altruism." MPRA Paper No. 2727, Munich Personal RePEc Archive, Munich.

Levine, D. K. (1998), "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics* **1**(3), 593-622.

Marwell, G. and Ames, R. (1981), "Economists Free Ride, Does Anyone Else?: Experiments on the Provision of Public Goods, IV." *Journal of Public Economics* **15**(3), 295-310.

McCabe, K., Houser, D., Ryan, L., Smith, V. and Trouard, T. (2001), "A Functional Imaging Study of Cooperation in Two-Person reciprocal Exchange." *Proceedings of the National Academy of Sciences* **98**(20), 11832–11835.

Milinski, M., Semmann, D., Bakker, T. C. M. and Krambeck, H.-J. (2001), "Cooperation Through Indirect Reciprocity: Image Scoring or Standing Strategy?" *Proceedings of the Royal Society B: Biological Sciences* **268**(1484), 2495-2501.

Nakao, K. (2008), "Can Altruism Hinder Cooperation?" *Economics Bulletin* **4**(26), 1-6.

Nowak, M. A. and Sigmund, K. (1998), "Evolution of Indirect Reciprocity by Image Scoring / The Dynamics of Indirect Reciprocity." Interim Report IR-98-040, International Institute for Applied Systems Analysis, Laxenburg.

Nowak, M. A. and Sigmund, K. (2005), "Evolution of indirect reciprocity." *Nature* **437**(7063), 1291-1298.

North, D. C. (1998), "Where Have We Been and Where Are We Going?" In Ben-Ner, A. and Puttermann L. (eds.), *Economics, Values, and Organization.* Cambridge: Cambridge University Press.

Orbell, J. M., Dawes, R. M. and van de Kragt, A. J. C.(1988), "Explaining Discussion Induced Cooperation." *Journal of Personality and Social Psychology* **54**(5), 811-19.

Rabin, M. (1993), "Incorporating Fairness into Game Theory and Economics." *American Economic Review* **83**(5), 1281-1302.

Sen, A. (1998), "Foreword." In Ben-Ner, A. and Puttermann L. (eds.), *Economics, Values, and Organization.* Cambridge: Cambridge University Press.

Sigmund, K. (1998), "Complex Adaptive Systems and the Evolution of Reciprocation." Interim Report IR-98-100, International Institute for Applied Systems Analysis, Laxenburg.

Smith, A. (1759), *The Theory of Moral Sentiments.* Oxford: Clarendon Press.

Stanca, L. (2007), "Measuring Indirect Reciprocity: Whose Back Do We Scratch?" Working Paper No. 131, Department of Economics, University of Milan - Bicocca.

Stark, O. and Wang, Y. Q. (2004), "On the Evolutionary Edge of Altruism: A Game-Theoretic Proof of Hamilton's Rule for a Simple Case of Siblings." *Journal of Evolutionary Economics* **14(**1**)**, 37–42.

Tullberg, J. (2004), "On Indirect Reciprocity - The Distinction Between Reciprocity and Altruism, and a Comment on Suicide Terrorism." *The American Journal of Economics and Sociology* **63(**5), 1193-1212.

Wedekind, C. and Milinski, M. (2000), "Cooperation Through Image Scoring in Humans." *Science* **288**(5467), 850-852.

# Appendix

## Proposition 1

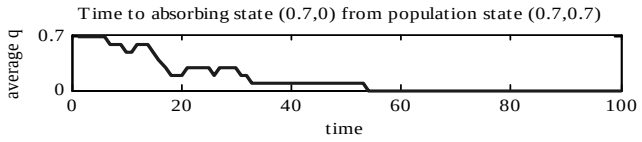


Figure 1: From pure cooperation to pure defection for the intial state $(0.7, 0.7)$.

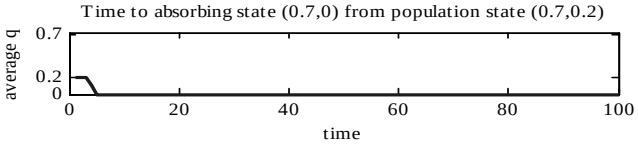


Figure 2: To the defecting strategy from initial population state $(0.7, 0.2)$.

## Proposition 2

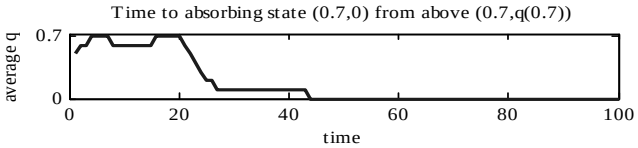


Figure 3: From mixed to defecting strategy from initial population state $(p, q(p) + \varepsilon)$

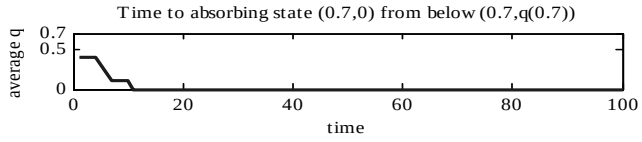


Figure 4: From mixed to defecting strategy from initial population state $(p, q(p) - \varepsilon)$

## Proposition 3

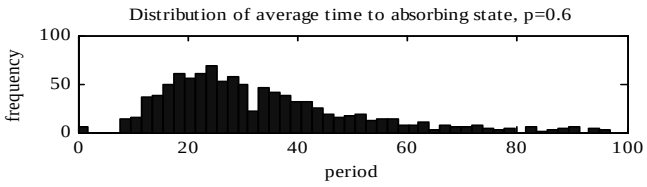


Figure 5: Distribution of average time to absorbing state for $p = 0.6$

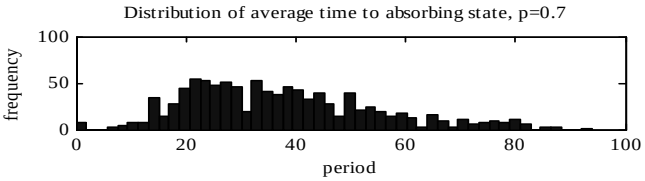


Figure 6: Distribution of average time to absorbing state for $p = 0.7$

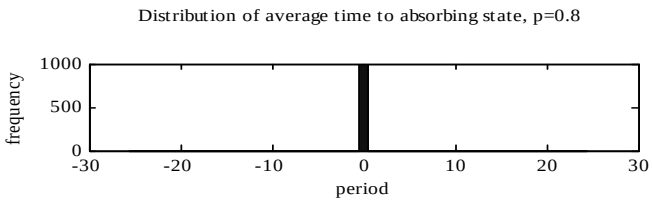


Figure 7: Distribution of average time to absorbing state for $p = 0.8$