COLOGNE GRADUATE SCHOOL
in Management, Economics and Social Sciences

Vol 2 | No 5

CGS WORKING PAPER

A Welfare-Tradeoff-Ratio-Model of Social Preferences

Björn Hartig (University of Cologne)

Cologne Graduate School
in Management, Economics
and Social Sciences
Albertus-Magnus-Platz
50923 Köln
www.cgs.uni-koeln.de

University of Cologne

A Welfare-Tradeoff-Ratio-Model of Social Preferences

Bjoern Hartig*

September 28, 2011

Abstract

This paper introduces a model of social preferences featuring a single parameter representing an individual's disposition to share resources with others. The parameter reacts to observed behavior of others in a clearly defined manner. Therefore, the model allows the numerical analysis of reciprocal interaction. Based on evolutionary concepts, the model is characterized by a very basic utility maximization condition and it is consistent with and often predictive of the results of a multitude of different behavioral games and phenomenon. (JEL C71, C73, C90, C91, D03, D63, D64)

Keywords: other-regarding preferences, altruism, cooperation, evolution, reciprocity, welfare-tradeoff-ratio.

1 Introduction

There are currently several popular models in circulation covering different aspects of social preferences. Some highlight the prominence of the equal split (Fehr and Schmidt 1999; Bolton and Ockenfels 2000), some welfare concerns (Charness and Rabin 2002), others reciprocity (Levine 1998; Dufwenberg 2004; Falk and Fischbacher 2006). Additionally, there is a growing economic literature exposing various aversions and other behavioral irregularities (e.g. Bohnet and Zeckhauser 2004; Charness and Dufwenberg 2006; Dana et al. 2007). While not all of these models and theories are mutually exclusive per

*Cologne Graduate School of Management, Economics and Social Sciences, University of Cologne, Richard-Strauss-Strasse 2, 50931 Köln, Germany. E-mail: hartig@wiso.uni-koeln.de.

se, they sometimes lead to different predictions (e.g. Engelmann and Strobel 2004). Moreover, critics are mocking the apparent tendency to come up with a novel fix to the utility function for each newly discovered behavioral phenomenon as the "neo-classical repair shop" (Güth 1995). In the attempt to make economic models of social behavior more realistic, they criticize, the already implausible assumption of utility maximization is made even less plausible by further complicating the function with additional elements. In turn, though, economists usually counter that the "as-if"-approach in economics cares little for realism and all the more for mathematical analysis, prediction and verifiability, there is no obvious reason why considering the underlying cognitive processes of behavior must necessarily lead to inferior economic models. In fact, this approach appears to be the most promising one to develop an integrative theory of social preferences able to explain a large variety of social behavior.

The model presented in this paper attempts just that. It considers the underlying evolutionary mechanisms of altruism, reciprocity and cooperation without sacrificing mathematical traceability. The model incorporates the essence of other social preference models - the aforementioned prominence of the equal split, welfare concerns, and reciprocity - and it accommodates (and often predicts) not only the typical results of simple games like dictator and ultimatum game, but also those of several other studies of behavioral irregularities. Furthermore, it addresses the "repair-shop-critics" by postulating a first-order-condition for utility maximization that is both very basic and - the author believes - resembles the actual cognitive process reasonably closely. Essentially, the model uses only one dynamic parameter: the individual's disposition to share welfare. The dynamics behind this parameter are based on concepts of evolutionary biology and evolutionary psychology.

The paper continues with a brief overview of those concepts in section 2. Section 3 presents the model and section 4 analyzes how it fares with the results of different economic studies. Different aspects of the model, including justification for important assumptions, challenges, and extensions are discussed in section 5. Finally, section 6 offers the summary.

2 Evolutionary Background

2.1 Altruism

What utility is to Economics, fitness is to evolution: the ultimate driving force behind behavior. Natural selection ensures that those specimens most adept at surviving and reproducing eventually displace their less successful counterparts. Consequently, there should be no room for anything but purely selfish behavior. Any truly altruistic act ultimately decreases the fitness of the actor and increases the fitness of the recipient, eventually leading to the extinction of the altruists. Hence, such behavior can not be evolutionary successful. Yet, there is plenty of evidence for seemingly altruistic behavior among many different species, e.g. suicidal hive defense by honey-bees, blood sharing by vampire bats or alarm calls by birds, to name just a few. Darwin himself acknowledged this paradox, declaring that truly altruistic behavior would "annihilate" his theory of natural selection (Darwin 1859).

It was not until Hamilton (1964) that a conclusive explanation for such behavior was found.[1] Because natural selection works on the level of the gene, not on the level of the specimen, altruistic behavior can be evolutionary stable if it satisfies the following inequation, later named Hamilton's rule:

$$\sum r_j b_j \geq c_i \tag{1}$$

The parameter r_j is the genetic relatedness of recipient j to the actor i (i.e. the probability that j shares a certain gene with i), b_j the fitness benefits to j and c_i the fitness costs to i. In other words, a gene for altruistic behavior is favored by natural selection if the total benefits for all other carriers of the gene is higher than the costs to the individual performing the altruistic act. This premise is satisfied, for example, if the carrier of the gene gives up less than one unit of fitness to either transfer two units of fitness on a sibling (with whom he shares one half of his genes[2]), or one unit on two siblings, or eight units on one cousin (with whom he shares one eight of his genes) and so on. This principle - coined "kin selection" by Smith (1964) - can only work if either individuals have the ability to detect their kin or if the natural probability to interact with kin is high (which it often is as relatives are more likely to be geographically close). Hamilton's rule actually explains the vast majority of altruistic behavior among fauna (Stevens et al. 2005).

2.2 Reciprocity and Cooperation

It is evident that humans often behave altruistically even towards non-kin (e.g. sharing of food with non-relatives in need). Viewed in isolation, such

[1] Haldane (1955) had already grasped the idea in principle. He is often quoted to have said that he would not give his life to save a drowning brother, but would do so if it were two brothers or eight cousins, indicating that he was aware that individuals share one half of their genes with their siblings and one eight with their cousins.

[2] Certain hive insects actually share 3/4 of their genes with their siblings, making it all the more perspicuous why most suicidal altruism is found among these species.

behavior obviously could not have evolved by natural selection because the genes of the donor lose fitness. If, however, the recipient will reciprocate the favor with some positive probability in the future, these benefits may outweigh the initial costs, leading to a net-surplus. If the cost/benefit ratio of the altruistic action is small, both the initial giver and the recipient can be better off afterwards. Still, the question remains why the recipient should bother to reciprocate. After all, he and his genes would be better off by behaving selfishly (i.e. not reciprocating).

In his seminal article, Trivers (1971) showed that if altruists are able to condition their future behavior on the observed behavior of the recipients, such conditional altruism can be favored by natural selection. By curtailing altruistic acts towards individuals that do not reciprocate, altruists can restrict the bulk of fitness benefits to fellow altruists. Then, the portion of altruists in the population increases if the following inequation holds:[3]

$$\frac{1}{p^2}(\sum b_k - \sum c_j) > \frac{1}{q^2}\sum b_m, \qquad (2)$$

with b_k the benefit of the kth altruistic act performed towards the altruist, c_j the cost of the jth altruistic act of the altruist, b_m the benefit of the mth altruistic act towards a non-altruist and p and q the frequency of altruists and non-altruists in the population, respectively. If the inequation holds, the average net benefits of altruists exceed the net benefits of non-altruists. This is more likely if there are many opportunities for altruistic behavior during the lifetime, if individuals are exposed symmetrically to them and if there is a small set of individuals with whom to interact repeatedly (for further discussion, see also section 3.5).

With a simulation of iterated prisoner's dilemma games, Axelrod and Hamilton (1981) later demonstrated that cooperation among non-related individuals can be beneficial even if not-cooperating is a dominant strategy (i.e. it yields higher benefits independent of the other's action) in one-shot encounters. When the number of future interactions is uncertain, TIT FOR TAT (cooperating in the first round, then imitating the partner's behavior from the previous round) is an evolutionary stable strategy (ESS), meaning that a population who has adopted it can not be invaded by other strategies.[4] Furthermore, Axelrod (1981) also showed that even a population of consistent defectors can be invaded by TIT FOR TAT-individuals if those are clustered initially and therefore have a high probability to interact with cooperative kin instead of defectors.

[3] Formula taken from Trivers' article.

[4] Technically, it is not a strict ESS because it can be invaded by other friendly (=cooperative) strategies which are indistinguishable in a TIT FOR TAT dominated population.

2.3 Welfare Tradeoff Ratio and Emotions

Evolutionary psychologists view the brain as a processor carrying out computations to solve recurrent adaptive problems. Its algorithms evolved under natural selection to regulate behavior in the direction of improved genetic fitness. Therefore, we should expect to find processes in the human brain reflecting the principles of altruism and cooperation described above. The cognitive variable regulating an individual's disposition to give up his own resources for the benefit of another individual has been called Welfare Tradeoff Ratio (WTR) by Tooby and Cosmides (2008). There are at least two distinct WTR: The intrinsic one $_{int}WTR$ is used when behavior is unobserved. It should generally obey Hamilton's Rule, making relatedness its major, though not only, input parameter. The public one $_{pub}WTR$ is used when behavior is observed by the recipient. It should be guided by the principles of cooperation and reciprocity, although all parameters influencing the intrinsic $_{int}WTR$ apply, too.

Behavior regulating variables like WTR must regularly be adjusted depending on new information available to the individual (e.g. new information about relatedness or recent behavior of the other person). However, since there is a plethora of different variables and parameters continuously tracked and evaluated by the brain, these adjustment processes generally can not work on a deliberate level, not even for human beings. Instead, algorithms are needed that, based on available cues, compel the individual towards behavior that - on average over an evolutionary timespan! - will be fitness promoting. Such algorithms are emotions. Emotions are complex programs that process different cues and stimuli to assign hedonic values to actions which are then weighted in the decision process (Tooby and Cosmides 1990).

Consequently, humans should possess emotions that regulate an individual's WTR in accordance with the principles of kin selection and reciprocity. For one, there should be an emotion (or a set of emotions) that increases the WTR if the recipient is related. Furthermore, another emotion is needed that decreases the WTR, compelling the individual to curtail benefits or possibly execute punishment, when the other person has displayed an inappropriately low WTR himself, i.e. was unwilling to adequately transfer benefits. And finally, emotions are required to induce and maintain positive reciprocity, i.e. emotions that increase the own WTR if failure to do so would make others withdraw benefits or execute punishment in return.

Love and Affection Transferring welfare to loved ones creates a positive hedonic experience, often even if the action is unobserved by the recipient. Hence, love and affection increase an individual's WTR. Kinship is one major factor determining affection (Lieberman et al. 2007), demonstrated for example in parental love. Other cues inducing love and affection are friendship or sexual attraction, for example. Affection and love, however, are typically not relevant factors in laboratory experiments, although they may play a role in prolonged interpersonal economic relationships at some point.

Anger The function of anger is to re-calibrate others' dispositions in favor of the angry individual (Sell et al. 2009). It is triggered when observed behavior of another person signals an inappropriately low WTR. When angry, on the one hand the individual experiences a less positive or possibly even negative sensation when transferring welfare to the other person. On the other hand, punishment can lead to positive sensations (Fischbacher et al. 2004). Hence, observing inappropriately low WTR in others decreases one's own WTR towards them, possibly making it negative.

Guilt and Gratitude Reducing the WTR of others is undesirable, so a mechanism should exist that avoids making others angry and instead induces or keeps up positive reciprocity (Tooby and Cosmides 2008). There are two different emotional programs accomplishing this: guilt and gratitude. On the one hand, guilt induces a negative sensation when choosing or having chosen an inappropriately low transfer oneself. To avoid this, the individual exceeds his WTR. On the other hand, gratitude induces a positive sensation when transferring benefits to a person who has made a higher transfer than expected. Increasing the own WTR in respond to an observed high WTR stabilizes cooperation.

This is of course a markedly simplifying description of a highly complex emotional system that neither considers the entirety of all possible cues (social status, physical strength, life expectancy etc.) and emotional regulators (jealousy, compassion, shame etc.) nor potential conjunctions among them (e.g. repeated cooperation can lead to friendship, which in turn keeps the WTR high without the constant need for reassurance through recent cues). Nevertheless, these three sets of emotions should capture the core principles for the evolution of altruism and reciprocity.

3 The Model

3.1 The central parameter

The model adopts the idea of WTR in its central parameter $\varepsilon_{i,j}$ (ε for *esteem*), player i's willingness to transfer own payoff to player j. If positive, $\varepsilon_{i,j}$ is the ratio to which player i is willing to share a payoff X with player j, i.e. $\varepsilon_{i,j} = \frac{x_j}{x_i}$, $x_i + x_i = X$. If negative, $\varepsilon_{i,j}$, is the negative ratio of own to other payoff that player i is still willing to accept before destroying the whole endowment. Hence, if $\varepsilon_{i,j} = 1$, player i divides the endowment equally among both players. If $\varepsilon_{i,j} = 0$, player i is uninterested in j's outcome and keeps the whole endowment for himself. If $\varepsilon_{i,j} = 1$, player i is willing to destroy the endowment to avoid any allocation where j's payoff exceeds his own. Therefore, the value of $\varepsilon_{i,j}$ is also the relative value player i puts on player j's payoff compared to the value put on his own payoff. It can casually be interpreted as a measure of i's appreciation of j. Theoretically, $\varepsilon_{i,j}$ can be any number from $-\infty$ to $+\infty$, but it usually lies between -1 and $+1$ because values above $+1$ are inefficiently altruistic and values below -1 inefficiently spiteful.

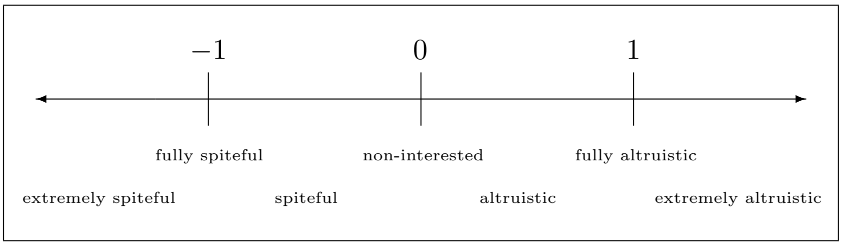

Figure 1: Range and classification of $\varepsilon_{i,j}$

3.2 The utility function

Player i prefers to allocate payoffs according to his effective $\varepsilon_{i,j}$. Therefore, i behaves **as if** he were maximizing the following utility function:

$$U_i = \sqrt{x_i} + sgn(\varepsilon_{i,j}\ x_j) \cdot \sqrt{|\varepsilon_{i,j}\ x_j|}, \tag{3}$$

with x_i and x_j player i's and j's monetary payoff, respectively, and sgn the sign-function ($sng(a) = 1$ if $a > 0$, $sng(a) = -1$ if $a < 0$ and $sng(a) = 0$ if $a = 0$). For positive $\varepsilon_{i,j}$,to maximize the expression, i has to solve the following first order condition:

$$x_i = \frac{1}{\varepsilon_{i,j}} x_j \Leftrightarrow \varepsilon_{i,j}\ x_i = x_j. \tag{4}$$

Solving the maximization problem only requires player i to consider how much he "appreciates" player j, i.e. how much he values player j's payoff compared to his own and then allocate the pie accordingly. The author believes that the cognitive process suggested by the model reflects both the WTR concept and the actual decision process reasonably closely (at least for an economic model).

3.3 Observability and reciprocity

Player i's willingness to transfer welfare is not constant for all situations, but variable. The changes to $\varepsilon_{i,j}$, however, follow certain rules that rest upon the evolutionary principles presented in the preceding section:

1. Player i's willingness to transfer welfare $\varepsilon_{i,j}$ depends on the observability of the action. The intrinsic disposition ${}_{int}\varepsilon_{i,j}$ when behavior is not observed is never higher than the public willingness ${}_{pub}\varepsilon_{i,j}$, i.e. ${}_{int}\varepsilon_{i,j} \leq {}_{pub}\varepsilon_{i,j}$. In accordance with Hamilton's rule, when the action is unobservable, i's intrinsic willingness ${}_{int}\varepsilon_{i,j}$ is usually 0 (or close to 0) unless i and j are related, which increases it. The public willingness ${}_{pub}\varepsilon_{i,j}$ usually lies somewhere between 0 and 1 if there is no prior history between both players.

2. Player i's $\varepsilon_{i,j}$ is also a function of j's observed disposition ${}_{obs}\varepsilon_{j,i}$. If the observed $\varepsilon_{j,i}$ differs from the reference disposition ${}_{app}\varepsilon_{j,i}$ that i considers "appropriate", i will adjust his own disposition accordingly with $\varepsilon'_{i,j}({}_{app}\varepsilon_{j,i} - {}_{obs}\varepsilon_{j,i}) \leq 0$. This corresponds to the emotional programs of anger and gratitude.

3. $\varepsilon_{i,j}$ is also influenced by what constitutes the "appropriate" ${}_{app}\varepsilon_{i,j}$ from j's perspective. If his initial ${}_{ini}\varepsilon_{i,j}$ is below that value, i adjusts it upwards, i.e. $\varepsilon'_{i,j}({}_{app}\varepsilon_{i,j} - {}_{ini}\varepsilon_{i,j}) \geq 0$. This corresponds to the avoidance of the emotion guilt and its effect should be much stronger when i's action is observed.

Which level of ε is considered appropriate depends on the exact circumstances (and the players must not necessarily agree about it), but there are five general guidelines:

1. When both players are cooperating (i.e. contributing equally to the provision of the pie), ${}_{app}\varepsilon = 1$, i.e. the equal split is the norm for cooperative situations.

2. When player i has announced his intention to choose $\varepsilon_{i,j}$ beforehand, it becomes ${}_{app}\varepsilon_{i,j}$ (unless the announced intention is already inappropriately low).

3. When player j has announced his expectation about $\varepsilon_{i,j}$ through his action, it becomes the standard ${}_{app}\varepsilon_{i,j}$ for player i. This expectations must be reasonable, which usually means ${}_{exp}\varepsilon_{i,j} \leq 1$.

4. When player j has revealed an appropriate $\varepsilon_{j,i}$, it becomes the appropriate response ${}_{app}\varepsilon_{i,j}$ for player i.

5. When none of these cases applies, player i considers his beliefs over the population's ${}_{app}\varepsilon$ for the given situation, i.e. the social norm. His initial ${}_{pub}\varepsilon_{i,j}$ already incorporates this belief. However, if i receives new information about the social norm, he will update his ${}_{pub}\varepsilon_{i,j}$ accordingly.

If more than one of those rules apply in concert, they enhance the emotional reaction accordingly. For example, if i signals his expectations $\varepsilon_{j,i} \geq x$ and j announces $\varepsilon_{j,i} \geq x$, but j's action reveal $\varepsilon_{j,i} < x$, then i will be even more angry (i.e. his $\varepsilon_{i,j}$ will decrease more) than when j had not announced his intention. Likewise, j would experience more guilt than had she not made that announcement.

3.4 Noisy signals

When player i has a fixed endowment X to allocate and the whole possible decision space $x_i = \{0, X\}$ available, each allocation directly reveals i's $\varepsilon_{i,j}$ to player j. For example, if $X = 10$ and i chooses $8|2$, player j knows $\varepsilon_{i,j} = 0.25$. However, when the decision space is limited, the signal is more ambiguous, but usually still carries some information value. For example, if player i has to choose between the two allocations $8|2$ and $5|5$ and selects $8|2$, player j only knows $\varepsilon_{i,j} \leq 0.52$ because choosing $8|2$ over $5|5$ reveals $\sqrt{8} + \sqrt{\varepsilon_{i,j}2} \geq \sqrt{5} + \sqrt{\varepsilon_{i,j}5}$. Assuming player j has ex-ante beliefs over the probability distribution of player i's $\varepsilon_{i,j}$, she uses the new information to update her beliefs accordingly. Similarly, if some other kind of noise is introduced to the procedure, e.g. by implementing the possibility that the final allocation was chosen by a random mechanism and not by i, player j will take the increased uncertainty of the signal into account when estimating $\varepsilon_{i,j}$. For player i, the decreased observability would let his chosen $\varepsilon_{i,j}$ move towards his intrinsic ${}_{int}\varepsilon_{i,j}$.

3.5 The Evolution of ε

3.5.1 Altruism and Reciprocity

		player j			player j	
		s	1-s		s	1-s
player	s	R / R	$\frac{\varepsilon_{j,i}R}{1+\varepsilon_{j,i}}$ / $\frac{R}{1+\varepsilon_{j,i}}$	s	$\sqrt{R}$ / $\sqrt{R}$	$\sqrt{\frac{\varepsilon_{j,i}R}{1+\varepsilon_{j,i}}}$ / $\sqrt{\frac{R}{1+\varepsilon_{j,i}}}$
i	1-s	$\frac{R}{1+\varepsilon_{i,j}}$ / $\frac{\varepsilon_{i,j}R}{1+\varepsilon_{i,j}}$	0 / 0	1-s	$\sqrt{\frac{R}{1+\varepsilon_{i,j}}}$ / $\sqrt{\frac{\varepsilon_{i,j}R}{1+\varepsilon_{i,j}}}$	0 / 0
		Payoffs			Fitness	

Figure 2: Example of social interaction

The evolution of ε and its regulatory programs is highly complex. It depends on the type and frequency of interactive situations individuals face, the exact payoffs for each situation, distribution of parameters and strategies in the population, the probability for repeated interaction, intertemporal discount factors, etc. To illustrate this, consider the following example, illustrated in figure (2):

Pairs of individuals i and j perform a task (e.g. hunting) for an indefinite number of periods. The probability that the game ends after a certain period is $1-\delta$ (Alternatively, δ is the discount factor between periods or a combination of both). Each individual is successful with probability s, in which case he receives an amount R of a resource (e.g. food), which he can convert into fitness with $F(R) = \sqrt{R}$, otherwise he receives nothing. If both individuals are successful, both convert R into fitness. If only player i is successful, he can transfer some of his resource R to the unsuccessful player j - revealing $\varepsilon_{i,j}$ - and vice versa.

However, such altruistic behavior is obviously fitness decreasing for each individual, unless the partner is related with some positive probability. If the partner shares the gene for unconditional transfers with probability p, unconditionally transferring resources is fitness increasing for the altruistic gene when $\varepsilon < \frac{4p^2}{(1-p^2)^2}$, with the optimal $\varepsilon = p^2$ (see A.3). Yet, even though unconditional transfers increase the fitness of the altruistic gene, the fitness of selfish genes increase even more because they profit from unconditional transfers, too, but sustain no losses. As a result, the share of altruistic genes in the population falls, decreasing in turn the probability p to encounter another carrier. Since the upper limit for fitness increasing ε decreases in p, the unconditional altruists will eventually fade out of the population (unless they are geographically clustered, see Axelrod 1981).

Obviously, the altruistic gene could do better by limiting transfers only to other carriers of the gene. Provided the individuals possess the cognitive ability to discern the behavior of their partner, altruistic individuals could adapt to cease transfers towards free-riders, who evidently do not possess the same altruistic gene. Such conditional behavior would spread among altruists because it constitutes a clear improvement over unconditional transfers. Eventually, all remaining altruists will be equipped with this ability. At this point, altruists may be better off than free riders on an individual level if their willingness to transfer is not too big ($\varepsilon < 4\left(\frac{x}{1-x^2}\right)^2$ with $x = \frac{\delta rs(1-s)}{(1-\delta)+\delta rs(1-s)}$, see A.4). If that is the case, it becomes negligible whether i and j actually are related or not.[5] Either way, altruists now will steadily increase their share in the population (theoretically up to 100%). However, this state is very vulnerable to shocks.

If a second type of altruists with lower willingness to transfer appears either as a mutation of the original altruists or of the free riders, these individuals will have an advantage over the competition. These new altruists do better than remaining free riders because they receive transfers from other altruists and they do better than the original altruists because their transfers are lower. Now, the optimal transfer an individual can make is the lowest possible transfer that is not considered free riding and answered with the termination of transfers. As a result, such "minimalistic altruists" will eventually dominate the population.

Evidently, the current reciprocity mechanism is too simplistic to keep cooperation on a high level. An altruistic individual needs to react not only to free riders, but to everybody who makes lower transfers than himself. There are several feasible possibilities what such a reaction may look like. For example, an individual could completely cease transfers when the partner makes a lower transfer or he could just lower his own transfers to the level of the partner's. In comparison, on the one hand, the latter strategy provides the higher fitness payoff because the individual continues to collect some transfers from his partner. On the other hand, the former strategy puts higher adaptive pressure on weak altruists. In any case, depending on the current distribution of types and parameters in the population, both strategies are potentially fitness increasing.

One way or the other, once strong altruists curtail their transfers to weak altruists, weak altruists in turn can improve if they keep transfers from strong altruist partners high. Assume partner j has already displayed a high willingness to transfer and player i is now in the situation that he has to reveal

[5] Relatedness is good explanation how altruism initially got started, but it is not necessary for its continued maintenance.

his own $\varepsilon_{i,j}$. Depending on the type of his partner j, any $\varepsilon_{i,j} < \varepsilon_{j,i}$ results either in j ceasing all future transfers ($\varepsilon_{j,i} = 0$) or in j replicating i's lower transfer ($\varepsilon_{j,i} = \varepsilon_{i,j}$) in all future rounds. In the first case, player i optimally copies the observed $\varepsilon_{j,i}$ if $\varepsilon_{j,i} \leq 4\left(\frac{ab}{b^2-a2}\right)^2$ with $a = \delta s(1-s)$ and $b = 1-\delta(1-s+s^2)$ (see A.5) and free rides otherwise. In the second case, he optimally chooses $\varepsilon_{i,j} = min[\varepsilon_{j,i}; \left(\frac{\delta s(1-s)}{1-\delta(1-s+s^2)}\right)^2]$ (see A.6), i.e. he replicates the observed transfer unless it is unduly high.

Eventually, replication of first movers' transfers will spread across the population and displace other types because it produces a higher expected fitness than any other behavior of second movers. Then, when faced with a population of replicators, a first mover i optimally reveals $\varepsilon_{i,j} = \left(\frac{1-\delta}{1-\delta(1-s+s^2)}\right)^2$ (see A.7), which is always copied by both possible types of replicators (see A.7.1). At this point, the population will eventually drift towards an equilibrium in which each player makes transfers according to $\varepsilon_{i,j}*$, irrespective of his role. Both strategies "*transfer $\varepsilon_{i,j}*$ unless the partner has revealed a lower $\varepsilon_{j,i}$, in which case* a) *cease all transfer* b) *replicate the partner's transfer*" form an evolutionary stable state (ESS) because neither can be successfully invaded by individuals with a different strategy.

The example has illustrated how a population of (in an evolutionary sense purely selfish) fitness maximizers develops altruism and reciprocity with the initial spark of kin selection and later help of an anger system (decreasing own transfer when observing low transfers) and a guilt & gratitude system (replicating observed high transfers). The actual human evolution of altruism and reciprocity is certainly, that goes without saying, indefinitely more complex. Mixed equilibria may very well exist and of course there is no reason to assume that we did actually reach an equilibrium at all - even back in prehistorical times. We could very well still remain in some kind of transition state, to this day adapting to the ever changing conditions of our environment. This question, however, is far beyond the scope of this paper. Nevertheless, the result that altruism increases with the share of altruists in the population (proof 3 - proof 4), the value put on the future and the frequency of opportunities for reciprocal behavior (proof 4 - proof 6) should apply generally (see also Trivers 1971).[6]

[6] The theoretical foundations of evolutionary stable reciprocity are relatively simple, yet intraspecies reciprocity is conspicuously rare in nature. Stevens et al. (2005) ascribe that to the substantial cognitive requirements. These include individual recognition, memory of former interactions and patience for delayed payoffs. While the first two are obviously mandatory for reciprocity, many animals are able of both. Intertemporal discounting, however, is probably crucial. In the example, as δ approaches 0, transfers break down. While humans devalue monetary payoffs on the order of months, most animals devalue food

3.5.2 Cooperation

When individuals cooperate, they work together to increase the size of the pie that is distributed between them. There are some cooperative interactions that are unproblematic because they yield stable and immediate benefits for all participants. This is called mutualism. Other interactions require at least one individual to forgo potential benefits. Consider the example above: Although no individual can do better on its own in the ESS, the ESS is not the most efficient solution possible. If the two players i and j could agree up front to always transfer half of the resource R when applicable, they could both increase their expected fitness (see A.8). However, unlike the ESS, which is the natural point of convergence, there is no reason that cooperation must necessarily evolve. With the agreement to share equally, the eventual first mover i is actually worse off than the first mover in the ESS. Therefore, first mover i would be tempted to make a smaller transfer than agreed upon, revealing $\varepsilon^*_{i,j}$ instead. If the second mover j "accepts" and replicates $\varepsilon^*_{i,j}$ - her fitness maximizing reaction! - no arrangement will ever be uphold.

For such cooperative arrangements to work, it is mandatory that the second mover reacts to the broken agreement with anger and reduces her $\varepsilon_{j,i}$ sharply in return (e.g. to $\varepsilon_{j,i} = 0$) so that i would be better off honoring the agreement (see A.9), even if this is not the fitness maximizing response. In that case, either cooperative agreements will eventually vanish again because they are broken too often and therefore offer no advantages over the ESS or first movers will adapt to the harsh reactions of second movers by increasing their willingness to uphold the agreement. There are several possible reasons why and how this could happen. For example, if many distinct opportunities for cooperation occur during the lifetime, individuals may go through an initial inefficient learning period in which agreements are broken more frequently and still ultimately receive net benefits once cooperation becomes more stable later.[7] Also, if first and second mover roles are distributed relatively even among all individuals, it is more likely that losses in the first mover role are compensated with gains in the second mover role. Finally, reputation building and social education (e.g. the social norm to keep one's

on the order of seconds (Stevens and Hauser 2004). In that case, even if opportunities for reciprocal interaction just lie a few minutes apart, the effective δ already is 0 and resources are never shared. Also note that the discount rate is irrelevant for kin selection.

[7] As a side note, since each individual can potentially be a first mover or a second mover, it is conceivable that first movers can anticipate the second mover's reaction by considering their own reaction in that role without having to actually experience it. Therefore, not many first movers would choose to break the agreement and the inefficiency would be small.

promises) are also helpful in establishing cooperation. However, the exact evolutionary process of non-mutualistic cooperation is not pivotal. The point is that IF cooperative arrangements exist, those arrangements must be the benchmark on which behavior is evaluated by the emotional system.

3.5.3 The Equal Split

Both mutualistic and non-mutualistic cooperation increase overall efficiency, but there is no a-priori rule for how the generated surplus is allocated between the parties. Theoretically, each amount that leaves an individual better off than the outside option makes cooperation the preferred choice, even if the partner receives a higher share. However, there are two reasons - besides its intuitive appeal - why we should expect the equal split to evolve as the norm for cooperation of individuals on equal terms. First, if an individual is engaged in multiple cooperative interactions, keeping track of several different agreed ratios and former interactions is cognitively taxing. The normative rule to split equally reduces the cognitive task of choosing the correct ratio and allows to classify and memorize each partner's behavior as a binary variable "*did split equally*" and "*did not split equally*". Second, in order to maximize the probability for successful cooperation, the individual probability for defection should be minimized over all participants (e.g. maximizing $min[P(\text{defect})_i, P(\text{defect})_j]$). Assuming this defection probability depends directly on the size of the individual cooperative gain compared to the outside option, then - given equal outside options - the equal split is the optimal allocation.

The equal split also has desirable properties from a group perspective.[8] Although group selection has largely been rejected as a (strong) force in evolution (Smith 1964), it is conceivable that as memes - units of social information (Dawkins 1976) - social norms of sharing resources equally even in non-cooperative situations are favored in intergroup competition. If resources have diminishing marginal returns, sharing them equally is efficiency optimizing and groups adhering to such a norm would ceter paribus be stronger than those that do not. Therefore, social norms of sharing and cooperation tend to be particularly philanthropic and charitable towards group members, even if those assertion and reality will often diverge.

However, despite its generally desirable properties, the equal split is not always feasible, in particular when the interacting individuals are different in

[8] Maxims and morals promoting such behavior can be found in many societies, for example "*Regard your neighbor's gain as your own gain, and your neighbor's loss as your own loss*", Laozi (Suzuki and Carus 2008) or "*Love your neighbor as you love yourself*", Luke 10:27.

one aspect or another. For example, in non-mutualistic cooperations, more able individuals usually are less willing to share equally because they are more likely to be in a position where they have to forgo benefits. Returning to the example above, if player i's probability for success is s_i and player j's probability for success is s_j with $s_i > s_j$, then player i may actually be better off not interacting with player j at all instead of agreeing to the equal split (see A.10). Similarly, if both players value future payoffs differently, the equal split is not equally attractive to both and one player may not be patient enough to honor any agreement to share equally while the other would be (see A.9). Furthermore, at times, a player may prefer the equal split to not cooperating, but nevertheless feel entitled to a bigger share, for example when his outside option is better. Similarly, a player who can inflict higher costs on his parter or who has invested more cooperative effort may feel entitled to more than the equal share. In these situations, the advantaged player may try to enforce an allocation rule that favors him, e.g. splitting the cooperative surplus instead of the total amount of resources when outside options differ (which would minimize the probability for defection) or allocating the surplus according to the ratio of the outside options. However, there is no universally valid solution for these situation and self-serving bias is to be expected on both sides. Nevertheless, regardless of those exceptions, among equals, the equal split should in general be the standard for cooperative interactions.

4 Applications

In this section, the WTR-model is applied to a number of common economic games and studies.

4.1 Dictator Game

In the dictator game, player i divides an endowment X between himself and player j, thereby directly revealing $\varepsilon_{i,j}$. Maximizing

$$U_i = \sqrt{x_i} + sgn(\varepsilon_{i,j})\sqrt{|\varepsilon_{i,j}(X - x_i)|}, \tag{5}$$

yields

$$x_i = \frac{1}{1+\varepsilon_{i,j}} X \Leftrightarrow x_j = \frac{\varepsilon_{i,j}}{1+\varepsilon_{i,j}} X \Leftrightarrow \varepsilon_{i,j} = \frac{X - x_i}{x_i} \tag{6}$$

4.2 Ultimatum Game

In the ultimatum game, player i proposes an allocation of an endowment X between himself and player j. If j agrees to the proposed allocation, it is paid out. Otherwise, both players receive nothing. Since both players' consent is needed for the implementation of the allocation, the ultimatum game can be considered a cooperative situation, so the appropriate offer from player i is $\frac{X}{2}$, i.e. ${}_{app}\varepsilon_{i,j} = 1$ (at the very least from player j's perspective).

Except for the very unlikely case that player j is deeply spiteful ($\varepsilon_{j,i} \leq -1$), she will always accept the appropriate offer of $\frac{X}{2}$ ($\varepsilon_{i,j} = 1$). Any lower offer, however, decreases $\varepsilon_{j,i}$ because $\varepsilon'_{j,i}({}_{app}\varepsilon_{i,j} -_{obs} \varepsilon_{i,j}) \leq 0$. This leads to the rejection of the offer if $\varepsilon_{j,i}$ decreases to such a degree that player j prefers the payoff of zero for both players to the proposed allocation, i.e.

$$0 > \sqrt{X - x_i} + sgn(\varepsilon_{j,i})\sqrt{|\varepsilon_{j,i}\ x_i|} \tag{7}$$

or

$$\varepsilon_{j,i}\ < 1 - \frac{X}{x_i} \tag{8}$$

Player i never offers less than $\frac{\varepsilon_{i,j}}{1+\varepsilon_{i,j}}X$, but may offer more if he beliefs that player j will reject the offer with some positive probability.

4.3 Falk et al. 2003

Falk et al. (2003) showed that identical offers in ultimatum games can lead to different rejection rates depending on the choices available to the proposer. In their study, proposers could choose between two different allocations, one of which was always 8 points for the proposer and 2 points for the recipient (8|2). The recipients in turn were gradually less likely to accept the 8|2 proposal when the alternative was 10|0, 8|2, 2|8 and 5|5, respectively.

As discussed in section 3.4, in this setting, proposer i's choice of 8|2 reveals less about $\varepsilon_{i,j}$ than it would in a regular ultimatum game. Choosing 8|2 over 10|0 signals $\varepsilon_{i,j} \geq 0.056$), while selecting 8|2 when the alternative is also 8|2 obviously reveals nothing about $\varepsilon_{i,j}$. Furthermore, the choice of 8|2 over 2|8 reveals $\varepsilon_{i,j} \leq 1$ and over 5|5, it reveals $\varepsilon_{i,j} \leq 0.52$. If recipient j initially had beliefs over the distribution of $\varepsilon_{i,j}$ that assigned a positive probability to at least one value below 0.056, between 0.056 and 0.52, between 0.52 and 1 and above 1, then updating those beliefs with the information received through i's choice for 8|2 yields expected values of $\varepsilon_{i,j}$ that are ordered accordingly to the observed rejection rates.

4.4 Engelmann & Strobel 2004 and comments

Engelmann and Strobel (2004) let subjects choose among three different allocations of money for three players. In all decisions, the decider had very little or no own payoff at stake. Engelmann and Strobel found that efficiency- and maximin-preferences as proposed in Charness and Rabin (2002) describe their results better than the models of inequity-aversion by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). The WTR-model, however, also predicts efficiency- and maximin-like preferences when the decider has no self-interest in the outcome.

For example, assume player i has to choose between allocation $A = a_i|a_j|a_k$ and allocation $B = b_i|b_j|b_k$, has no own payoff at stake ($a_i = b_i$) and has equal positive disposition towards both players j and k ($\varepsilon_{i,j} = \varepsilon_{i,k} > 0$). When $a_j \geq b_j$ and $a_k \geq b_k$ (and at least one true inequality), then player i always strictly prefers the more efficient allocation A. Furthermore, when the efficiency of both allocations is equal and $a_j = b_j + c$, $a_k + c = b_k$, $c > 0$, then player i will display Rawlsian preferences and choose allocation A because the utility function is concave in the other players' payoffs.

In their comment on the paper, Bolton and Ockenfels (2006) remark that despite the results of Engelmann and Strobel, people display a much lower willingness to pay for efficiency than for equality. According to the WTR-Model, if player i can choose between allocation $a|b$ and allocation $a - x|b + 1$ ($a \geq 1$, $b > 0$), there always exists a positive value of x that makes player i prefer the second allocation if $\varepsilon_{i,j} > 0$. The maximum acceptable value of x strictly decreases in b and if $b \geq a$, then $x < 1$ unless $\varepsilon_{i,j} > 1$ (see appendix A.1 for proofs). Therefore, for any given a, player i is always willing to pay more to increase the payoff of a player who has less than him (i.e. to increase equality) than to increase the payoff of a player who has more (i.e. to increase efficiency).

For good measure, Bolton and Ockenfels also present a pair of payoff allocations for six players in which the decider always receives 8 and the other five players receive either 8|8|8|15|1 or 2|2|2|33|2. The first allocation is picked by 94% of all subjects even though it violates both efficiency- and maximin-preferences (but decreases inequality as measured by Fehr and Schmidt as well as Bolton and Ockenfels). According to the WTR-model, assuming the decider's ε is the same towards each of the five other players, the first allocation is strictly prefered if $\varepsilon > 0$.

4.5 Dana et. al 2007

The study of Dana et al. (2007) has a between-subject design in which participants made decisions in three different two-player and one three-player dictator-game-like situations. In the *baseline* treatment, the dictator simply had to decide between a 6|1 and a 5|5 allocation. In the *hidden payoff* treatment, the dictator was given the choice between two partly hidden allocations, 6|? and a 5|?. He was informed that a coin flip had determined the receiver's payoff for each allocation and that the possible two pairs were either 6|1, 5|5 or 6|5, 5|1. If desired, the dictator could view the result of the coin toss without costs by pressing a button. Furthermore, the dictator knew that the receiver was not informed about whether the button was actually pressed or not. Finally, the *plausible deniability* treatment was similar to the *baseline* treatment with the following exception: After the dictator had one minute to contemplate, the decision between 6|1 and 5|5 had to be made in a ten second time window during which the dictator could possibly be cut off by the experimental software. If that happened, one of the two allocations was chosen randomly. The cut-off, while random, was set up in such a way that it allowed the dictator plenty of time to make the decision if so desired.

The results showed that while 74% (14 of 19) of the dictators chose 5|5 in the *baseline* treatment, only 56% (18 of 32) chose to reveal the concealed outcomes in the *hidden payoff* treatment. Those 44% leaving the outcomes hidden predominantly choose 6|? (12 of 14). In the *plausible deniability* treatment, 75% of the dicators made their choice before being cut off, but only 45% of them picked the fair allocation 5|5. Dana et al. conclude that while people feel the compulsion to give, they actually do not inherently value the other's payoff that much. Instead, they use the "moral wiggle room" provided in the experiment to act more selfishly. The WTR-model generally agrees with Dana and his co-authors that the intrinsic willingness to transfer resources is lower than the one displayed publicly ($_{int}\varepsilon < _{pub}\varepsilon$). It also allows a more detailed analysis of their results:

By choosing 6|1 over 5|5, player i signals $\varepsilon_{i,j} \leq 0.03$ ($\sqrt{5} + \sqrt{\varepsilon_{i,j}5} \leq \sqrt{6} + \sqrt{\varepsilon_{i,j}1}$), which is quite low and close to 0, therefore not particularly kind. In fact, the choice allows for the possibility that i is acting not only out of selfishness, but out of spite ($\varepsilon_{i,j} < 0$). Overall, the signal is so negative that only a minority of dictators is willing to make this choice in the baseline treatment. In contrast, the signal of the concealed allocation 6|? in the hidden payoff treatment - although not "nicer" in the sense that it signals a higher $\varepsilon_{i,j}$ - is at least less negative. Selecting 6|? implies not only preferring 6|1 over 5|5, but also 6|5 over 5/1. The latter signals $\varepsilon_{i,j} \geq -0.03$, i.e. the dictator is basically disinterested in the other's payoff, but is not spiteful either. Since the receiver is not informed whether the dictator has revealed the actual allocations or not, the signal is strictly for the dictator himself. Apparently, the guilt system reacts differently to the signals $\varepsilon_{i,j} \leq 0.03$ and $-0.03 \leq_{pub} \varepsilon_{i,j} \leq 0.03$. Contemplating the former induces enough guilt to increase $\varepsilon_{i,j}$ to such an extend that 5|5 is chosen in the baseline treatment, while contemplating the latter does not.

In the *plausible deniability* treatment, receivers did not know if the payoff relevant allocation was chosen by the dictator or by chance. To the receiver, this decreased the signaling effect of the outcome to some extend. On the one hand, if the final payoff is 6|1, it is still possible the dictator has $\varepsilon_{i,j} \geq 0.03$ (if he chose 5|5, but was cut off). On the other hand, if the payoff is 5|5, the dictator may still have $\varepsilon_{i,j} \leq 0.03$ (if he chose 6|1, but was cut off), too. This makes it more likely that the dictator chooses a lower transfer than he usually would under full observation. Ultimately, if he knew the receiver (wrongfully) believed that all payoffs were decided by pure chance, the dictator would choose according to his intrinsic ${}_{int}\varepsilon_{i,j}$.

4.6 Dana et. al 2006

Dana et al. (2006) gave subjects the option to silently exit a \$10-dictator game and take \$9 instead. About one third of their subjects took that option, even though they could have secured a higher payoff by choosing a transfer of \$0 in the dictator game. This model offers an explanation for this result using the difference between ${}_{int}\varepsilon_{i,j}$ and ${}_{pub}\varepsilon_{i,j}$. When i deliberates the choice between the exit option and the dictator game, he is unobserved by j and hence uses his intrinsic willingness to transfer payoff in his deliberations, aware, however, that ${}_{int}\varepsilon_{i,j}$ will change to ${}_{pub}\varepsilon_{i,j}$ once the public dictator game is entered. Therefore, i will choose to exit if the anticipated choice in the dictator game leaves his intrinsically motivated self worse off than taking the \$9, i.e.

$$\sqrt{\frac{1}{1 +_{pub} \varepsilon_{i,j}} \$10} + \sqrt{{}_{int}\varepsilon_{i,j} \frac{{}_{pub}\varepsilon_{i,j}}{1 + \ {}_{pub}\varepsilon_{i,j}} \$10} < \sqrt{\$9}\,. \tag{9}$$

If, for example, ${}_{int}\varepsilon_{i,j} = 0$, then i will strictly prefer the exit option if ${}_{pub}\varepsilon_{i,j} \geq \frac{1}{9}$.

4.7 Trust Game

Instead of the version of the game initially developed by Berg et al. (1995), in order to focus on the aspects unique to the trust game, this analysis looks at a variant of the game in which both players i and j start out with the same endowment X and assume for the same reason that no action of the first mover i is motivated by altruism. Player i can now transfer any amount x up to X to player j. If he does so, the amount is multiplied by $m > 1$ (typically $m = 2$ or $m = 3$) and added to j's endowment. Then, if i chooses to transfer a positive amount x, he does so because he expects j to make a back-transfer y that leaves i with more than his original endowment X, i.e. $y > x$. Which amount player i optimally send depends on the parameters X and m and his expectations about player j's $\varepsilon_{j,i}$.

When j receives xm from i, she will make a back transfer

$$y = max[\frac{X(\varepsilon_{j,i} - 1) + \varepsilon_{j,i}\ x(m+1)}{1 + \varepsilon_{j,i}}; 0],$$

which, if positive, increases with the initial transfer x and of course $\varepsilon_{j,i}$, but decreases in the initial endowment X (see A.2). However, player j will only make a positive back transfer if her willingness to transfer exceeds a certain threshold $\varepsilon_{j,i} > \frac{X}{X - x(m+1)}$, which increases in X and decreases in x and m (if $x > 0$). Furthermore, looking at the ratio of back transfer to received amount $g = \frac{y}{mx}$, notably, this ratio increases in the initially transfered amount x for a given willingness to transfer $\varepsilon_{j,i}$ unless $\varepsilon_{j,i} = 1$, in which case player j always sends back 50% of what she received from player i. This is interesting because it offers an explaination for the increase of relative back transfers with increasing initial transfers that is often found in trust games (e.g. Sapienza et al. 2007) that does not rely on perceived kindness and reciprocity. After all, it is not a conclusively settled question whether a higher initial transfer by player i is in fact "nicer" or just more risky when the sender's intention is to get a sufficiently large back transfer instead of increasing the recipient's final payoff.

Solving for player i's optimal transfer is rather complex even if i is risk neutral. Just using player i's expectation about the average $\varepsilon_{j,i}$ in the population of recipients is not sufficient because back transfers are not linear in $\varepsilon_{j,i}$. Instead, the expectations about the distribution of $\varepsilon_{j,i}$ would have to be considered. To simplify the analysis, consider only a binary-choice version of the trust game in which player i can either send his whole endowment X or nothing at all. In response, player j can either make a back transfer so that both players end up with the same payoff ($y = \frac{X(m+1)}{2}$) or keep everything for herself.

Player j makes the back transfer $y = \frac{X(m+1)}{2}$ if she prefers the allocation $\frac{X(m+1)}{2}/\frac{X(m+1)}{2}$ over $X(m+1)/0$. This is the case if $\varepsilon_{j,i} \geq 3 - \sqrt{2^3} \approx 0.172$. Therefore, if player i is only concerned with his own payoff ($\varepsilon_{j,i} = 0$), he should make the initial transfer if he beliefs that the percentage p of recipients with $\varepsilon_{j,i} \geq 3 - \sqrt{2^3}$ is at least $\sqrt{\frac{2}{m+1}}$.[9]

4.8 Betrayal Aversion

Betrayal aversion was first discovered by Bohnet and Zeckhauser (2004), who compared subjects' choices in a binary-choice trust game with a risky dictator game in which the payoffs of both players were determined by chance. First mover i could either opt out of the game and secure a payoff of 10|10 for both players, or he could enter the game, in which case depending on the treatment either second mover j or a random mechanism chose between 8|22 or 15|15. For both treatments, they elicited the minimum acceptable probability for the outcome 15|15 so that first movers would choose to enter the game. On average, first movers were willing to accept a lower probability for 15|15 in the risky dictator game than in the trust game. Bohnet and Zeckhauser concluded that first movers suffer betrayal costs when second movers choose 8|22, but that no such costs exist if the result occurred by chance.

The WTR-model agrees with this explanation. The first mover i enters the risky dictator game if the probability $P(15|15) \geq \frac{\sqrt{10}+\sqrt{10\varepsilon_{i,j}}-\sqrt{8}-\sqrt{22\varepsilon_{i,j}}}{\sqrt{15}+\sqrt{15\varepsilon_{i,j}}-\sqrt{8}-\sqrt{22\varepsilon_{i,j}}}$. In the trust game, however, he chooses differently because of possible anger. The second mover j will prefer 15|15 over 8|22 if $\varepsilon_{j,i} > 0.61$, so if the trust game were equivalent to the risky decision, player i would enter it if $P(\varepsilon_{j,i} > 0.61) \geq \frac{\sqrt{10}+\sqrt{10\varepsilon_{i,j}}-\sqrt{8}-\sqrt{22\varepsilon_{i,j}}}{\sqrt{15}+\sqrt{15\varepsilon_{i,j}}-\sqrt{8}-\sqrt{22\varepsilon_{i,j}}}$. However, player i's $\varepsilon_{i,j}$ changes depending on j's choice. By opting into the game, player i signals player j that he expects $\varepsilon_{j,i} > 0.61$. Since this is a cooperative situation, this expectation is appropriate. Therefore, player j can do little to positively surprise player i, which would raise i's $\varepsilon_{i,j}$ through gratitude. An increased $\varepsilon_{i,j}$ would have made the allocation 15|15 even more attractive to i and made him accept a lower probability $P(\varepsilon_{j,i} > 0.61)$ to enter the trust game. However, if player j chooses 8|22, revealing $\varepsilon_{j,i} < 0.61$, player i will become angry, i.e. his original $\varepsilon_{i,j}$ decreases to ${}_{ang}\varepsilon_{i,j}$. If player i anticipates this reaction, he will only accept a probability $P(\varepsilon_{j,i} > 0.61) \geq \frac{\sqrt{10}+\sqrt{10\varepsilon_{i,j}}-\sqrt{8}-\sqrt{22\ {}_{ang}\varepsilon_{i,j}}}{\sqrt{15}+\sqrt{15\varepsilon_{i,j}}-\sqrt{8}-\sqrt{22\ {}_{ang}\varepsilon_{i,j}}}$, which is higher than the probability required to enter the risky dictator game.

[9] Using the utility function of the model in the context of risk is a bit ominous since it always implies risk aversion. Since the focus is not on risk, however, I simply accept this and do not dig deeper into the issue for now, although I can not help but find it at least pleasingly coherent that (from my subjective impression) risk aversion seems to be the more "natural" state of mind while risk neutrality is usually learned.

4.9 Guilt and Disappointment Aversion

Charness and Dufwenberg (2006) let subjects play a kind of trust game in which the first mover i had to decide between the option *out* and *in*. If *out* was chosen, both players received a payoff of 5. Otherwise, player j had to decide between rolling a dice or not rolling it. When she chose not to roll, j received a payoff of 14 and i a payoff of 0. Otherwise, j received a payoff of 10 for sure and i received 0 with probability $\frac{1}{6}$ and 12 with probability $\frac{5}{6}$. Furthermore, in the treatment condition, the second mover j could send a non-binding message to i before he made his decision to opt in or out. Most second movers used this message to promise the first mover to roll the dice if given the chance. The results showed that player i was more likely to choose the *in* option after receiving such a promise. Likewise, j was more likely to indeed roll the dice when she had previously promised to. Additionally, the message increased both i's beliefs about the probability of j choosing to roll and j's beliefs about i's beliefs.

These results are in line with the WTR-model. Choosing to roll the dice over not rolling it requires just $\varepsilon_{j,i} \geq 0.04$. By opting into the game, first mover i signals his expectation $\varepsilon_{j,i} \geq 0.04$. If second mover j announces his willingness to roll the dice, it further increases first mover i's expectations that j will indeed display $\varepsilon_{j,i} \geq 0.04$. Then, if player j instead reveals $\varepsilon_{j,i} < 0.04$ by not rolling the dice, player i will experience more anger if j has made a promise than when she has not (according to the model). Unfortunately, Charness and Dufwenberg did not elicit anger or mood, but it seems reasonable to presume that i's increased beliefs about the probability of j rolling the dice do in fact translate into increased anger if this expectation is not met. The key issue, however, is that second movers j, as expected, react with stronger guilt, i.e. contemplating not rolling the dice increases more guilt after a promise, which leads to a stronger increase in $\varepsilon_{j,i}$, which makes j more likely to actually roll the dice when given the chance.

Although second movers apparently are aware that a promise changes the first movers' expectations, it is not clear whether they are purposely trying to avoid anger when they increase their willingness to roll the dice. That question was approached by Vanberg (2008), who slightly modified the above game by randomly re-assigning half of the pairs of first and second movers after the message from the second mover was sent. Only the second mover j was informed if her partner was switched or not. If her first-mover was switched, the second mover was also informed about the message the new partner had received previously. Overall, second movers were more likely to choose rolling the dice when they had sent a message containing a promise to do so irrespective of whether their partner was switched or not. On the other hand, the message a re-assigned first mover received before the switch had very little influence on the second mover's disposition to choose roll. Vanberg concluded that people have an inherent preference for keeping their promises, but not for avoiding to let others down.

At first sight, this conclusion seems to contradict the assumption that individuals try to avoid anger in others. However, while the experimental design by Vanberg was certainly clever, such situations are virtually impossible in real life. For basically all conceivable circumstances, keeping one's promise is tantamount to avoiding disappointment while an inclination to keep one's promise requires less cognitive resources and is less error prone than considering the possible expectation of another individual. Therefore, Vanberg's results do not contradict the evolutionary approach of the WTR-model (on the contrary).

5 Discussion

5.1 Challenges and Extensions

5.1.1 Jealousy

Both outcome- and intention-based models offer explanations for why players reject unequal offers in the ultimatum game. At the same time, both types of models run into different problems with modifications of the game. For example, outcome-based models are at a loss when rejection rates change depending on alternative options not chosen by the proposer, while intention-based models cannot explain any rejections when offers are not made by the other player but some kind of non-interested mechanism. Combining both outcomes and intentions, the WTR-model can explain both, but while the former modification is easily dealt with (see 4.3), the latter is a little bit more problematic. For example, receiver j rejects an offer of $2|8$ if $\varepsilon_{j,i} < -0.25$. If such an offer does not come from player i but, for instance, a random mechanism, player j's $\varepsilon_{j,i}$ has not shifted from its "natural" state. By implication, player j should also prefer $2|0$ over $10|8$, which seems very implausible.

The problem could be solved by postulating different $\varepsilon_{i,j}$ when player i's payoff is higher than player j's ($\varepsilon^{>}_{i,j}$) and when player i's is lower ($\varepsilon^{<}_{i,j}$). Assuming $\varepsilon^{>}_{i,j} \geq \varepsilon^{<}_{i,j}$, then player i may be willing to reduce player j's payoff when behind, but not when ahead. This adjustment would partly resemble the model of Fehr and Schmidt (1999), but would not be so strict to impose $\varepsilon^{<}_{i,j} < 0$ for all players. However, while there probably is some truth to the general idea, it adds another parameter to the model, so there is a trade off between model fit and degrees of freedom.[10]

5.1.2 Procedural Fairness

In the study of Blount (1995), subjects were more likely to reject unfair ultimatum game offers when those were made by a disinterested third party than when made by a random mechanism. Similarly, Bolton and Ockenfels (2005) found lower rejection rates when the random process was fair (i.e. equal probabilities for good and bad outcome) than when it was unfair (i.e. higher probability for bad outcome). Obviously, these results can not be explained by any of the aforementioned models, including the basic WTR-model. However, a simple extension of the the WTR-model that is in line with its evolutionary explanation may integrate procedural fairness into the framework.

The first question that needs to be answered is why individuals would care about the fairness of a disinterested third party or procedure? Or more precisely, why do some individuals prefer to destroy payoffs when either the third party or the procedure appears to be unfair? After all, it only hurts themselves and another non-responsible person. Such behavior only makes sense in an evolutionary scenario. In nature, truly disinterested third parties that nevertheless distribute resources do not exist. When individuals receive benefits from others, the distributors always have some kind of interest in the receiver. Either they expect a reciprocal response or - more applicable here - they are to some degree intrinsically concerned about the well-being of the recipient. They may, however, have varying degrees of intrinsic interest in different individuals. For example, parents may prefer to give more food to offspring with higher probability of survival or with certain traits.

If an individual observes the distributor displaying an inappropriately low disposition towards itself, the individual becomes angry in order to evoke a favorable adjustment of the distributor's disposition. Yet, for some reason, it cannot directly withdraw benefits from the distributor.[11] It can, however,

[10] It would also be interesting to find out to what degree a framing effect is responsible for the relatively large rejection rates of unfair offers - around 20% or higher in Falk and Fehr (2003); Bolton and Ockenfels (2005); Blount (1995). It is conceivable that if individuals were asked to choose between the two allocations $0|0$ and $2|8$ instead of accepting or rejecting $2|8$, more players would choose $2|8$ because the psychological effect of choosing $2|8$ over $0|0$ (being kind to the other player) may be different from not rejecting $2|8$ (accepting heteronomous unfairness).

[11] There may be no direct stream of resources from the individual towards the distributor, the individual may lack the physical ability necessary for such an act or the act may simply be to risky.

impair the distributor indirectly by destroying (some of) the allocated resource. Since the distributor is intrinsically interested in the welfare of the angry individual, he may prefer changing the allocation to the destruction of the resource. This effect is even stronger if the angry individual also destroys the other recipient's resources, especially if by choosing an allocating favoring him, the distributor has revealed a higher intrinsic interest in that recipient.

To illustrate, assume that X is the amount set aside by distributor i to be distributed among receivers j and k and that distributor i prefers to bestow x_j to receiver j and $x_k = X - x_j$ to receiver k. Then we denote the ratio $\frac{x_j}{x_k}$ as ${}^{i}\varepsilon_{j,k}$ (and ${}^{i}\varepsilon_{k,j} = {}^{i}\varepsilon_{j,k}^{-1}$). ${}^{i}\varepsilon_{j,k} = 1$ implies that distributor i values the welfare of both receivers equally, while ${}^{i}\varepsilon_{j,k} > 1$ and ${}^{i}\varepsilon_{j,k} < 1$ imply favoritism of receiver i and receiver j, respectively. Generally, the appropriate disposition for distributor i is ${}^{i}\varepsilon_{j,k} = 1$, although variations may exist similar to those described in 3.3 and 3.5.3. Deviations from the appropriate ${}^{i}_{app}\varepsilon_{j,k}$ alter recipient j's ${}^{j}\varepsilon_{j,k}$ analogously to deviations from a direct interaction partner, although the effect should tend to be weaker than direct interaction, i.e. $\varepsilon'_{j,k}({}^{i}_{app}\varepsilon_{j,k} - {}^{i}_{obs}\varepsilon_{j,k}) \leq 0$ and $|\varepsilon'_{j,k}({}^{i}_{app}\varepsilon_{j,k} - {}^{i}_{obs}\varepsilon_{j,k})| \leq |\varepsilon'_{j,k}({}^{k}_{app}\varepsilon_{j,k} - {}^{k}_{obs}\varepsilon_{j,k})|$.[12] As a result, recipient j is more likely to reject an allocation of $2|8$ when it is made by a "real" decision maker than by a fair random mechanism and also more likely to reject such an allocation made by an unfair mechanism than by a fair mechanism.

Usually, distributor i's ${}^{i}\varepsilon_{j,k}$ can be observed directly from the chosen allocation. If the distributor chooses a random mechanism, ${}^{i}\varepsilon_{j,k}$ can be calculated as follows. If the chosen lottery leads to allocation $x_{j,1}|x_{k,1}$ with probability p and to $x_{j,2}|x_{k,2}$ with probability $(1-p)$, then ${}^{i}\varepsilon_{j,k} = \frac{px_{j,1}+(1-p)x_{j,2}}{px_{k,1}+(1-p)x_{k,2}}$. Therefore, a procedure is "appropriate" if it yields the same expected outcome to each player. Since utility is concave in payoffs, players consider a lottery that yields each player a payoff of x in expectations equally "appropriate" (or "fair" or "kind") as a secure allocation of $x|x$, but they would prefer the secure allocation over the lottery.

5.1.3 Multiplayer Situations

The model is primarily designed to work in two-player-scenarios, but when decision makers have no own payoff at stake like in 4.4, it can be applied to multiplayer-scenarios without problems. When decisions do affect their own payoff as well as that of multiple agents, however, decision makers may

[12] Of course, this also decreases recipient j's disposition towards distributor i, but j has no means to express that directly.

either consider their own payoff in relation to the combined group payoff or in relation to each individual. Accordingly, there are two possible utility functions:

$$U_i = \sqrt{x_i} + \sum_{j=1}^{n} \frac{1}{n} sgn(\varepsilon_{i,j}\ x_j)\sqrt{|\varepsilon_{i,j}\ x_j|} \tag{10}$$

$$U_i = \sqrt{x_i} + \sum_{j=1}^{n} sgn(\varepsilon_{i,j}\ x_j)\sqrt{|\varepsilon_{i,j}\ x_j|} \tag{11}$$

For example, given these two utility functions, an individual i with $\varepsilon_{i,j} = \varepsilon_{i,k} = 1$ would choose the allocations 6|3|3 and 4|4|4, respectively, when dividing an endowment of 12. The former suggests that i considers the combined payoff of the other individuals while the latter suggests that i values payoff of each individual as much as his own and it also corresponds to the idea that $\varepsilon = 1$ represents the peak of other regarding concerns.

6 Summary

The WTR-model adds a new approach to the economic literature on social preferences. By employing a single dynamic parameter that is sensitive to observed behavior of others, the model is able to combine outcome- and intention-based aspects. The single parameter allows for comparative numerical analysis of interaction between players without relying on first- and second-order beliefs or psychological utility which are often difficult to pinpoint explicitly.

To the author's knowledge, the WTR-model is the first economic model that considers the underlying evolutionary and psychological mechanisms of altruism, reciprocity, and cooperation. The WTR-model assumes that individuals are inherently selfish and not intrinsically interested in non-related others unless they find themselves in a social interaction. Then, most individuals feel compelled to help those who are worse off than themselves and to honor (implicit) agreements for cooperation. Unlike models of inequality aversion, the WTR-model does not claim that individuals suffer from disutility when payoffs are unequal, but instead suggests that they receive positive utility when transferring recourses to others.

Section 4 demonstrated how the results of different games and studies can be explained using the WTR-model. This suggests that the model is consistent with a wide variety of behavioral phenomena. The next step would be to test the WTR-model on more studies to further assess its range and also its shortcomings. Ultimately, we need to develop experiments that explicitly test the predictions of the model against competing explanations.

The model also crucially relies on the idea of a normative reference behavior that is used to assess own and others' behavior. Although some preliminary concepts about the emergence of such norms are introduced in the paper, this area definitely warrants further research, too.

References

Axelrod, R. "The Emergence of Cooperation among Egoists The Emergence of Cooperation among Egoists." *The American Political Science Review*, 1981, 75(2): 306–318.

Axelrod, R., and Hamilton, W. D. "The evolution of cooperation." *Science*, 1981, 211(4489): 1390–6.

Berg, J., Dickhaut, J., and McCabe, K. "Trust, reciprocity, and social history." *Games and Economic Behavior*, 1995, 10(1): 122–142.

Blount, S. "When Social Outcomes Arent Fair: The Effect of Causal Attributions on Preferences." *Organizational Behavior and Human Decision Processes*, 1995, 63(2): 131–144.

Bohnet, I., and Zeckhauser, R. "Trust, risk and betrayal." *Journal of Economic Behavior & Organization*, 2004, 55(4): 467–484.

Bolton, G. E., and Ockenfels, A. "ERC - A theory of equity, reciprocity and competition." *American Economic Review*, 2000, 100(1): 166–93.

Bolton, G. E., and Ockenfels, A. "A stress test of fairness measures in models of social utility." *Economic Theory*, 2005, 25(4).

Bolton, G. E., and Ockenfels, A. "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment." *American Economic Review*, 2006, 96(5): 1906–1911.

Charness, G., and Dufwenberg, M. "Promises and partnership." *Econometrica*, 2006, 74(6): 1579–1601.

Charness, G., and Rabin, M. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics*, 2002, 117(3): 817–869.

Dana, J., Cain, D. M., and Dawes, R. M. "What you don't know won't hurt me: Costly (but quiet) exit in dictator games." *Organizational Behavior and Human Decision Processes*, 2006, 100(2): 193–201.

Dana, J., Weber, R., and Kuang, J. X. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory*, 2007, 33(1): 67–80.

Darwin, C. *The Origin of Species.* . First Edit ed., London:Penguin, 1859.

Dawkins, R. *The Selfish Gene.* . 30th Anniv ed., Oxford, 1976.

Dufwenberg, M. "A theory of sequential reciprocity." *Games and Economic Behavior*, 2004, 47(2): 268–298.

Engelmann, D., and Strobel, M. "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments." *American Economic Review*, 2004, 94(4): 857–869.

Falk, A., and Fehr, E. "Why labour market experiments?" *Labour Economics*, 2003, 10(4): 399–406.

Falk, a., and Fischbacher, U. "A theory of reciprocity." *Games and Economic Behavior*, 2006, 54(2): 293–315.

Falk, A., Fehr, E., and Fischbacher, U. "On the Nature of Fair Behavior." *Economic Inquiry*, 2003, 41(1): 20–26.

Fehr, E., and Schmidt, K. M. "A theory of fairness, competition, and cooperation." *Quarterly journal of Economics*, 1999, 114(3): 817–868.

Fischbacher, U., De Quervain, D., Treyer, V., Schellhammer, M., Schnyder, U., Fehr, E., and Buck, A. "The neural basis of altruistic punishment." *Science*, 2004, 305(5688): 1254.

Güth, W. "On ultimatum bargaining experiments A personal review." *Journal of Economic Behavior & Organization*, 1995, 27(3): 329–344.

Haldane, J. "Population Genetics." *New Biology*, 1955, 18: 34–51.

Hamilton, W. D. "The genetical evolution of social behaviour. II." *Journal of theoretical biology*, 1964, 7(1): 17–52.

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

Lieberman, D., Tooby, J., and Cosmides, L. "The architecture of human kin detection." *Nature*, 2007, 445(7129): 727–31.

Sapienza, P., Toldra, A., and Zingales, L. 2007. "Understanding trust."

Sell, A., Tooby, J., and Cosmides, L. "Formidability and the logic of human anger." *Proceedings of the National Academy of Sciences of the United States of America*, 2009, 106(35): 15073–8.

Smith, J. "Group selection and kin selection." *Nature*, 1964, 201(4924): 1145–1147.

Stevens, J. R., and Hauser, M. D. "Why be nice? Psychological constraints on the evolution of cooperation." *Trends in cognitive sciences*, 2004, 8(2): 60–5.

Stevens, J. R., Cushman, F. a., and Hauser, M. D. "Evolving the Psychological Mechanisms for Cooperation." *Annual Review of Ecology, Evolution, and Systematics*, 2005, 36(1): 499–518.

Suzuki, D. T., and Carus, P. *T'Ai-Shang Kan-Ying P'Ien: Treatise of the Exalted One on Response and Retribution.* Whitefish, Montana:Kessinger Pub Co, 2008.

Tooby, J., and Cosmides, L. "The past explains the present Emotional adaptations and the structure of ancestral environments." *Ethology and Sociobiology*, 1990, 11(4-5): 375–424.

Tooby, J., and Cosmides, L. 2008. "The evolutionary psychology of the emotions and their relationship to internal regulatory variables." In: *Handbook of Emotions.* 114–137.

Trivers, R. "The evolution of reciprocal altruism." *The Quarterly review of biology*, 1971, 46(1): 35–57.

Vanberg, C. "Why Do People Keep Their Promises? An Experimental Test of Two Explanations." *Econometrica*, 2008, 76(6): 1467–1480.

A Proofs

A.1 Proof I

Given $a \geq 1$, $b \geq 0$ and $x > 0$, player i never prefers $a - x/b + 1$ over a/b when $\varepsilon_{i,j} \leq 0$. Now, if $0 < \varepsilon_{i,j} \leq 1$,[13] i prefers $a - x/b + 1$ over a/b when

$$\begin{aligned} &\sqrt{a-x} + \sqrt{\varepsilon_{i,j}\ (b+1)} \geq \sqrt{a} + \sqrt{\varepsilon_{i,j}\ b} \\ \Leftrightarrow &\sqrt{a-x} \geq \sqrt{a} + \sqrt{\varepsilon_{i,j}\ b} - \sqrt{\varepsilon_{i,j}\ (b+1)} \\ \Leftrightarrow &\sqrt{a-x} \geq \sqrt{a} - \sqrt{\varepsilon_{i,j}}\ \underbrace{\left(\sqrt{b+1} - \sqrt{b}\right)}_{\leq 1} \end{aligned} \tag{12}$$

The right-hand term is always non-negative, therefore i will accept any x that satisfies the following condition:

$$\begin{aligned} &\sqrt{a-x} \geq \sqrt{a} - \sqrt{\varepsilon_{i,j}}\left(\sqrt{b+1} - \sqrt{b}\right) \\ \Leftrightarrow &a - x \geq a - 2\sqrt{\varepsilon_{i,j}\ a}\left(\sqrt{b+1} - \sqrt{b}\right) + \varepsilon_{i,j}\left(\sqrt{b+1} - \sqrt{b}\right)^2 \\ \Leftrightarrow &x \leq 2\sqrt{\varepsilon_{i,j}\ a}\left(\sqrt{b+1} - \sqrt{b}\right) - \varepsilon_{i,j}\left(\sqrt{b+1} - \sqrt{b}\right)^2 \end{aligned} \tag{13}$$

Since $\sqrt{\varepsilon_{i,j}\ a} \geq \varepsilon_{i,j}$ and $\sqrt{b+1} - \sqrt{b} > (\sqrt{b+1} - \sqrt{b})^2$, there always exists a positive x player i would accept and the maximum x given by the right hand expression. If $b > a$,

The first derivative of this expression with respect to b yields

$$\frac{\sqrt{\varepsilon_{i,j}\ ab} - \sqrt{\varepsilon_{i,j}\ a(b+1)} + 2b\varepsilon_{i,j} - 2\varepsilon_{i,j}\sqrt{b(b+1)} + \varepsilon_{i,j}}{\sqrt{b(b+1)}} \tag{14}$$

which becomes negative if

$$\begin{aligned} &\sqrt{\varepsilon_{i,j}\ ab} - \sqrt{\varepsilon_{i,j}\ a(b+1)} + 2b\varepsilon_{i,j} - 2\varepsilon_{i,j}\sqrt{b(b+1)} + \varepsilon_{i,j} < 0 \\ \Leftrightarrow &\sqrt{\varepsilon_{i,j}}(2b + 1 - \sqrt{b(b+1)}) < \sqrt{a}(\sqrt{b+1} - \sqrt{b}) \\ \Leftrightarrow &\sqrt{\varepsilon_{i,j}} < \frac{\sqrt{a}(\sqrt{b+1} - \sqrt{b})}{b + 1 - \sqrt{b(b+1)} + b} \\ \Leftrightarrow &\varepsilon_{i,j} < \frac{a}{(\sqrt{b+1} - \sqrt{b})^2} \end{aligned} \tag{15}$$

[13] The general results would remain the same if $a < 1$ and $\varepsilon_{i,j} > 1$ were allowed. However, the restrictions assure that $x < a$, i.e. i does not end up with a negative payoff.

which is always true under the given assumptions. Hence, x decreases as b increases.

Furthermore, if $a = b$, the expression for x becomes

$$\begin{aligned} x &\leq 2\sqrt{\varepsilon_{i,j}\, b}\left(\sqrt{b+1} - \sqrt{b}\right) - \varepsilon_{i,j}\left(\sqrt{b+1} - \sqrt{b}\right)^2 \\ \Leftrightarrow x &\leq (2\sqrt{\varepsilon_{i,j}} + 2\varepsilon_{i,j})\underbrace{(\sqrt{b(b+1)} - b)}_{<0.5} - \varepsilon_{i,j} \end{aligned} \tag{16}$$

Under the restriction $0 < \varepsilon_{i,j} \leq 1$, the expression reaches its maximum when $\varepsilon_{i,j} = 1$, which results in $x_{a=b} < 1$. Since x decreases as b increases, $x < 1 \forall b \geq a$.

A.2 Proof II

Player j chooses her back transfer y so that the first order condition (4) is satisfied. However, y is restricted to non-negative values.

$$\begin{aligned} \varepsilon_{j,i}(X + mx - y) &= X - x + y \\ \Leftrightarrow y &= \frac{X(\varepsilon_{j,i} - 1) + \varepsilon_{j,i}\, x(m+1)}{1 + \varepsilon_{j,i}} \\ y \geq 0 \Rightarrow y &= max[\frac{X(\varepsilon_{j,i} - 1) + \varepsilon_{j,i}\, x(m+1)}{1 + \varepsilon_{j,i}}; 0] \end{aligned} \tag{17}$$

If $y > 0$, the back transfer y decreases in X, but increases in x, m(if $x > 0$) and of course $\varepsilon_{j,i}$.

$$\frac{\partial \frac{(\varepsilon_{j,i} - 1) + \varepsilon_{j,i}\, x(m+1)}{1 + \varepsilon_{j,i}}}{\partial X} = \frac{\varepsilon_{j,i} - 1}{\varepsilon_{j,i} + 1} < 0 \tag{18}$$

$$\frac{\partial \frac{(\varepsilon_{j,i} - 1) + \varepsilon_{j,i}\, x(m+1)}{1 + \varepsilon_{j,i}}}{\partial x} = \frac{\varepsilon_{j,i}(m+1)}{1 + \varepsilon_{j,i}} > 0 \tag{19}$$

$$\frac{\partial \frac{(\varepsilon_{j,i} - 1) + \varepsilon_{j,i}\, x(m+1)}{1 + \varepsilon_{j,i}}}{\partial m} = \frac{\varepsilon_{j,i} x}{1 + \varepsilon_{j,i}} > 0 \tag{20}$$

$$\frac{\partial \frac{(\varepsilon_{j,i} - 1) + \varepsilon_{j,i}\, x(m+1)}{1 + \varepsilon_{j,i}}}{\partial \varepsilon_{j,i}} = \frac{x(m+1) + 2X}{(1 + \varepsilon_{j,i})^2} > 0 \tag{21}$$

Player j chooses $y > 0$ if her $\varepsilon_{j,i}$ exceeds a certain threshold.

$$\begin{aligned}
&\frac{X(\varepsilon_{j,i}-1)+\varepsilon_{j,i}\; x(m+1)}{1+\varepsilon_{j,i}} > 0 \\
\Leftrightarrow &\varepsilon_{j,i}\,(X - x(m+1)) > X \qquad (22)\\
\Leftrightarrow &\varepsilon_{j,i} > \frac{X}{X-x(m+1)}
\end{aligned}$$

The minimum $\varepsilon_{j,i}$ necessary for a positive back transfer y increases in X, but decreases in x and m (if $x > 0$) .

$$\frac{\partial \frac{X}{X-x(m+1)}}{\partial X} = \frac{x(m+1)}{(X+x(m+1))^2} > 0 \qquad (23)$$

$$\frac{\partial \frac{X}{X-x(m+1)}}{\partial x} = -\frac{X(m+1)}{(X+x(m+1))^2} < 0 \qquad (24)$$

$$\frac{\partial \frac{X}{X-x(m+1)}}{\partial m} = -\frac{xX}{(X+x(m+1))^2} < 0 \qquad (25)$$

A.3 Proof III

Assume proportion p of the population shares the altruistic gene that compels its carrier i to unconditionally make positive transfers (i.e. $\varepsilon > 0$) to his partner j each time only i is successful, i's expected fitness from a single interaction is given by:

$$EF = \underbrace{s^2\sqrt{R}}_{\text{both successful}} + \underbrace{s(1-s)\sqrt{\frac{R}{1+\varepsilon}}}_{\substack{i \text{ successful,}\\ j \text{ unsuccessful}}} + \underbrace{p\; s(1-s)\sqrt{\frac{\varepsilon R}{1+\varepsilon}}}_{\substack{i \text{ unsuccessful,}\\ j \text{ successful}\\ \text{and also a carrier}}}. \qquad (26)$$

Unconditionally transferring is fitness improving if its expected fitness is larger than the expected fitness of never transferring, which is $s\sqrt{R}$. partner j each time only i is successful, i's expected fitness from a single interaction is given by:

$$\begin{aligned}
& s\sqrt{R} < && s^2\sqrt{R} + s(1-s)\sqrt{\frac{R}{1+\varepsilon}} + p\ s(1-s)\sqrt{\frac{\varepsilon R}{1+\varepsilon}} \\
\Leftrightarrow\quad & 1 < && s + (1-s)\sqrt{\frac{1}{1+\varepsilon}} + p\ (1-s)\sqrt{\frac{\varepsilon}{1+\varepsilon}} \\
\Leftrightarrow\quad & 1 < && \sqrt{\frac{1}{1+\varepsilon}} + p\sqrt{\frac{\varepsilon}{1+\varepsilon}} \\
\Leftrightarrow\quad & \sqrt{1+\varepsilon} < && 1 + r\sqrt{\varepsilon} \\
\Leftrightarrow\quad & \varepsilon < && 2p\sqrt{\varepsilon} + p^2\varepsilon \\
\Leftrightarrow\quad & \varepsilon < && \frac{4p^2}{(1-p^2)^2}
\end{aligned} \tag{27}$$

The first derivative of (26) is

$$EF^{'} = \frac{s(s-1)\left(\sqrt{\frac{\varepsilon R}{1+\varepsilon}} - p\sqrt{\frac{\varepsilon R}{1+\varepsilon}}\right)}{2\sqrt{\varepsilon}(\varepsilon+1)}, \tag{28}$$

which becomes 0 if

$$\begin{aligned}
& \sqrt{\frac{\varepsilon R}{1+\varepsilon}} - p\sqrt{\frac{R}{1+\varepsilon}} && = 0 \\
\Leftrightarrow\ & \varepsilon && = p^2
\end{aligned} \tag{29}$$

The second derivative of (26) is

$$EF^{''} = \frac{(1-s)s\left(-4p\varepsilon\sqrt{\frac{\varepsilon R}{1+\varepsilon}} - p\sqrt{\frac{\varepsilon R}{1+\varepsilon}} + 3\varepsilon^2\sqrt{\frac{R}{1+\varepsilon}}\right)}{4\varepsilon^2(1+\varepsilon)^2}, \tag{30}$$

which becomes negative if

$$\begin{aligned}
& -4p\varepsilon\sqrt{\frac{\varepsilon R}{1+\varepsilon}} - p\sqrt{\frac{\varepsilon R}{1+\varepsilon}} + 3\varepsilon^2\sqrt{\frac{R}{1+\varepsilon}} < 0 \\
\Leftrightarrow\ & 3\varepsilon^{\frac{3}{2}} < 4p^3 + p.
\end{aligned} \tag{31}$$

If $\varepsilon = p^2$, then (31) becomes $3p^3 < 4p^3 + p$, which is always true for positive p. Therefore $\varepsilon^* = p^2$ is the optimal willingness to transfer and $\frac{\varepsilon^* R}{1+\varepsilon^*}$ the optimal unconditonal transfer.

A.4 Proof IV

Assume that proportion p of the population consists of (related) altruists and $1-p$ of free riders. All altruists will stop their transfers once they have observed that their partner is a free rider. All individuals discount each round with δ.

Free Riders: Each round, free rider i_{fr} is successful and receives R with probability s. With probability $s(1-s)$, i_{fr} is unsuccessful, but his partner j is successful. Then, with probability p, his partner j is an altruist who transfers $\frac{\varepsilon_{j,i}R}{1+\varepsilon_{j,i}}$ if i_{fr} has not yet revealed himself as a free rider. Each round, i_{fr}'s free rider type is revealed with probability s(1-s). A free rider's expected fitness therefore is given by

$$EF_{i_{fr}} = \frac{s\sqrt{R}}{1-\delta} + \frac{ps(1-s)\sqrt{\frac{\varepsilon_{j,i}R}{1+\varepsilon_{j,i}}}}{1-\delta(1-s+s^2)} \tag{32}$$

Conditional Altruists: Conditional altruist i_{ca} always receives R when both he and j are successful. This happens with probability s^2 each round. With probability p, his partner j is also an altruist ($\varepsilon_{j,i} = \varepsilon_{i,j}$). Then, when j was successful but i_{ca} was not, i_{ca} always receives a transfer of $\frac{\varepsilon_{i,j}R}{1+\varepsilon_{i,j}}$ from j vice versa. Each case occurs with probability $s(1-s)$. With probability $(1-p)$, j is a free rider. Then, when i_{ca} is successful and j is not, i_{ca} makes a transfer only if he has not yet learned j's type. Once j has revealed herself to be a free rider, i_{ca} ceases all transfer. The conditional altruist's expected fitness is therefore given by

$$\begin{aligned} EF_{ca} =& \frac{s^2\sqrt{R}}{1-\delta} + \frac{ps(1-s)\left(\sqrt{\frac{R}{1+\varepsilon_{i,j}}} + \sqrt{\frac{\varepsilon_{j,i}R}{1+\varepsilon_{j,i}}}\right)}{1-\delta} + \\ & (1-p)s(1-s)\left(\frac{\sqrt{\frac{R}{1+\varepsilon_{i,j}}}}{1-\delta(1-s+s^2)} + \delta\left(\frac{\sqrt{R}}{1-\delta} - \frac{(1-s+s^2)\sqrt{R}}{1-\delta(1-s+s^2)}\right)\right) \end{aligned} \tag{33}$$

Conditional altruists are doing better than free riders if

$$\begin{aligned} & EF_{fr} < EF_{ca} \\ \Leftrightarrow & \frac{s\sqrt{R}}{1-\delta} + \frac{ps(1-s)\sqrt{\frac{\varepsilon R}{1+\varepsilon}}}{1-\delta(1-s+s^2)} < \end{aligned}$$

$$
\begin{aligned}
&\frac{s^2\sqrt{R}}{1-\delta} + \frac{ps(1-s)\sqrt{\frac{R}{1+\varepsilon}} + \sqrt{\frac{\varepsilon R}{1+\varepsilon}}}{1-\delta} + \\
&(1-p)s(1-s)\left(\frac{\sqrt{\frac{R}{1+\varepsilon}}}{1-\delta(1-s+s^2)} + \delta\left(\frac{\sqrt{R}}{1-\delta} - \frac{(1-s+s^2)\sqrt{R}}{1-\delta(1-s+s^2)}\right)\right) \\
\Leftrightarrow&\frac{s(1-s)}{1-\delta} + \frac{ps(1-s)\sqrt{\frac{\varepsilon}{1+\varepsilon}}}{1-\delta(1-s+s^2)} < ps(1-s)\frac{\sqrt{\frac{1}{1+\varepsilon}} + \sqrt{\frac{\varepsilon}{1+\varepsilon}}}{1-\delta} + \\
&(1-p)s(1-s)\left(\frac{\sqrt{\frac{1}{1+\varepsilon}}}{1-\delta(1-s+s^2)} + \delta\left(\frac{1}{1-\delta} - \frac{(1-s+s^2)}{1-\delta(1-s+s^2)}\right)\right) \\
\Leftrightarrow&\frac{1}{1-\delta} + p\frac{\sqrt{\frac{\varepsilon}{1+\varepsilon}}}{1-\delta(1-s+s^2)} < p\frac{\sqrt{\frac{1}{1+\varepsilon}} + \sqrt{\frac{\varepsilon}{1+\varepsilon}}}{1-\delta} + \\
&(1-p)\left(\frac{\sqrt{\frac{1}{1+\varepsilon}}}{1-\delta(1-s+s^2)} + \delta\left(\frac{1}{1-\delta} - \frac{(1-s+s^2)}{1-\delta(1-s+s^2)}\right)\right) \\
\Leftrightarrow&1 + \frac{\delta p}{1-\delta} + p\frac{\sqrt{\frac{\varepsilon}{1+\varepsilon}}}{1-\delta(1-s+s^2)} < p\frac{\sqrt{\frac{1}{1+\varepsilon}} + \sqrt{\frac{\varepsilon}{1+\varepsilon}}}{1-\delta} + \\
&(1-p)\left(\frac{\sqrt{\frac{1}{1+\varepsilon}}}{1-\delta(1-s+s^2)} - \delta\frac{(1-s+s^2)}{1-\delta(1-s+s^2)}\right) \\
\Leftrightarrow&(1-\delta)(1-\delta(1-s+s^2)) + \delta p(1-\delta(1-s+s^2)) + p(1-\delta)\sqrt{\frac{\varepsilon}{1+\varepsilon}} < \\
&p(1-\delta(1-s+s^2)(\sqrt{\frac{1}{1+\varepsilon}} + \sqrt{\frac{\varepsilon}{1+\varepsilon}}) + \\
&(1-p)\left((1-\delta)\sqrt{\frac{1}{1+\varepsilon}} - \delta(1-\delta)(1-s+s^2)\right) \\
\Leftrightarrow&\delta ps(1-s) + (1-\delta) < \\
&\delta ps(1-s)\sqrt{\frac{\varepsilon}{1+\varepsilon}} + (\delta ps(1-s) + (1-\delta))\sqrt{\frac{1}{1+\varepsilon}} \\
\Leftrightarrow&1 < \sqrt{\frac{\varepsilon}{1+\varepsilon}}\frac{\delta ps(1-s)}{\delta ps(1-s) + (1-\delta)} + \sqrt{\frac{1}{1+\varepsilon}} \\
\Leftrightarrow&\sqrt{1+\varepsilon} < \sqrt{\varepsilon}\frac{\delta ps(1-s)}{\delta ps(1-s) + (1-\delta)} + 1
\end{aligned}
$$

Denoting $\frac{\delta ps(1-s)}{\delta ps(1-s)+(1-\delta)}$ as x, we get

$$\begin{aligned}
\Leftrightarrow 1+\varepsilon &< \varepsilon x^2 + 2\sqrt{\varepsilon}x + 1 \\
\Leftrightarrow \varepsilon(1-x^2)^2 &< 4x^2 \\
\Leftrightarrow \varepsilon &< \frac{4x^2}{(1-x^2)^2}
\end{aligned} \tag{34}$$

The first derivative of that expression with respect to x is

$$\frac{\partial \frac{4x^2}{(1-x^2)^2}}{\partial x} = -\frac{8(x^3+x)}{(x^2-1)^3}. \tag{35}$$

Since $0 < x < 1$, the expression is always > 0. The first derivatives of x with respect to δ, p and s are

$$\frac{\partial \frac{\delta ps(1-s)}{\delta ps(1-s)+(1-\delta)}}{\partial \delta} = \frac{ps(1-s)}{\left(\delta\delta ps(1-s)+(1-\delta)\right)^2} > 0 \tag{36}$$

$$\frac{\partial \frac{\delta ps(1-s)}{\delta ps(1-s)+(1-\delta)}}{\partial p} = -\frac{\delta(1-\delta)s(1-s)}{\left(\delta\delta ps(1-s)+(1-\delta)\right)^2} > 0 \tag{37}$$

$$\frac{\partial \frac{\delta ps(1-s)}{\delta ps(1-s)+(1-\delta)}}{\partial s} = -\frac{8(x^3+x)}{\left(\delta\delta ps(1-s)+(1-\delta)\right)^2} \begin{cases} > 0 \text{ if } s < 0.5 \\ = 0 \text{ if } s = 0.5 \\ < 0 \text{ if } s > 0.5 \end{cases} \tag{38}$$

Therefore, the maximum ε with which conditional altruists have a higher expected fitness than free riders increases in δ and p and increases in s until it reaches its maximum at $s = 0.5$ after which is decreases in s. Furthermore, this means that it increases strictly in the probability that a transfer actually occurs $2s(1-s)$ since

$$\frac{\partial 2s(1-s)}{\partial s} = 2 - 4s \begin{cases} > 0 \text{ if } s < 0.5 \\ = 0 \text{ if } s = 0.5 \\ < 0 \text{ if } s > 0.5 \end{cases} \tag{39}$$

The derivative of the conditional altruists expected fitness given in (33) with respect to ε is given by

$$\frac{\partial EF_{ca}}{\partial \varepsilon} = \frac{s(1-s)\sqrt{R}}{2(1+\varepsilon)^{3/2}} \left(\frac{p\left(\sqrt{\frac{1}{\varepsilon}}-1\right)}{(1-\delta)} - \frac{1-p}{1-\delta(1-s+s^2)} \right) \tag{40}$$

which becomes zero if

$$\begin{aligned} &\frac{p\left(\sqrt{\frac{1}{\varepsilon}}-1\right)}{(1-\delta)}-\frac{1-p}{1-\delta(1-s+s^2)}=0 \\ \Leftrightarrow &p(1-\sqrt{\varepsilon})(1-\delta(1-s+s^2))-\sqrt{\varepsilon}(1-p)(1-\delta)=0 \\ \Leftrightarrow &\sqrt{\varepsilon}\left(p(1-\delta(1-s+s^2))+(1-p)(1-\delta)\right)=p(1-\delta(1-s+s^2)) \\ \Leftrightarrow &\varepsilon^*=\left(\frac{p(1-\delta(1-s+s^2))}{p(1-\delta(1-s+s^2))+(1-p)(1-\delta)}\right)^2 \end{aligned} \tag{41}$$

The second derivative of (33) with respect to ε is given by

$$\begin{aligned} \frac{\partial^2 EF_{ca}}{\partial\varepsilon^2}=&\frac{\sqrt{s(1-s)R}}{4(1+\varepsilon)^{5/2}}\times \\ &\left[-\frac{p(1+\varepsilon)}{(1-\delta)\varepsilon^{3/2}}-3\left(\frac{p\left(\sqrt{\frac{1}{\varepsilon}}-1\right)}{(1-\delta)}-\frac{1-p}{1-\delta(1-s+s^2)}\right)\right]. \end{aligned} \tag{42}$$

This expression becomes negative if the term in square brackets is negative. Since the term in normal brackets is 0 for ε^*, the whole expression is negative. Therefore, the fitness reaches its maximum at ε^*.

The first derivatives of ε^* with respect to δ, p and s are

$$\frac{\partial\varepsilon^*}{\partial\delta}=\frac{2p^2(1-p)s(1-s)(1-\delta(1-s+s^2))}{\left(1-\delta(1-ps(1-s))\right)^3}>0 \tag{43}$$

$$\frac{\partial\varepsilon^*}{\partial p}=\frac{2p(1-\delta)(1-\delta(1-s+s^2))^2}{\left(1-\delta(1-ps(1-s))\right)^3}>0 \tag{44}$$

$$\frac{\partial\varepsilon^*}{\partial s}=\frac{2\delta p^2(1-p)(2s-1)[(1-\delta)^2+\delta(1-\delta)s(1-s)]}{\left(1-\delta(1-ps(1-s))\right)^3}\begin{cases}>0 \text{ if } s<0.5\\ =0 \text{ if } s=0.5\\ <0 \text{ if } s>0.5\end{cases} \tag{45}$$

Therefore, the optimal ε^* increases in δ and p, increases in s until it reaches its maximum at $s=0.5$ after which is decreases in s, and increases in the probability for transfers $2s(1-s)$.

A.5 Proof V

Assume that in at least one previous round, player j has made a positive transfer and revealed $\varepsilon_{j,i}$. The current round is the first round in which i

was successful, but j was not. If i makes a transfer that reveals $\varepsilon_{i,j} < \varepsilon_{j,i}$, j will cease all future transfers. Therefore, the only feasible options are $\varepsilon_{i,j} = 0$ (free riding) and $\varepsilon_{i,j} = \varepsilon_{j,i}$ (replication), all other $\varepsilon_{i,j}$ necessarily lead to outcomes inferior to at least one of those two. If i reveals $\varepsilon_{i,j} = 0$, his expected fitness is given by

$$EF_{fr} = \sqrt{R} + \delta s \frac{\sqrt{R}}{1-\delta}. \tag{46}$$

If i reveals $\varepsilon_{i,j} = \varepsilon_{j,i}$ instead, i's expected fitness is given by

$$EF_c = \sqrt{\frac{R}{1+\varepsilon_{j,i}}} + \delta s^2 \frac{\sqrt{R}}{1-\delta} + \delta s(1-s) \frac{\sqrt{\frac{R}{1+\varepsilon_{j,i}}} + \sqrt{\frac{\varepsilon_{j,i} R}{1+\varepsilon_{j,i}}}}{1-\delta}. \tag{47}$$

Replicating j's transfer is superior to free riding if

$$\begin{aligned}
& EF_{fr} \leq EF_c \\
\Leftrightarrow & \sqrt{R} + \delta s \frac{\sqrt{R}}{1-\delta} \leq \sqrt{\frac{R}{1+\varepsilon_{j,i}}} + \delta s^2 \frac{\sqrt{R}}{1-\delta} + \delta s(1-s) \frac{\sqrt{\frac{R}{1+\varepsilon_{j,i}}} + \sqrt{\frac{\varepsilon_{j,i} R}{1+\varepsilon_{j,i}}}}{1-\delta} \\
\Leftrightarrow & 1 + \frac{\delta s(1-s)}{1-\delta} \leq \sqrt{\frac{1}{1+\varepsilon_{j,i}}} + \delta s(1-s) \frac{\sqrt{\frac{1}{1+\varepsilon_{j,i}}} + \sqrt{\frac{\varepsilon_{j,i}}{1+\varepsilon_{j,i}}}}{1-\delta} \\
\Leftrightarrow & \sqrt{1+\varepsilon_{j,i}} \left(1 + \frac{\delta s(1-s)}{1-\delta}\right) \leq \left(1 + \frac{\delta s(1-s)}{1-\delta}\right) + \delta s(1-s) \frac{\sqrt{\varepsilon_{j,i}}}{1-\delta} \\
\Leftrightarrow & \sqrt{1+\varepsilon_{j,i}} \leq 1 + \sqrt{\varepsilon_{j,i}} \frac{\delta s(1-s)}{1-\delta(1-s+s^2)} \\
\Leftrightarrow & 1 + \varepsilon_{j,i} \leq 1 + 2\sqrt{\varepsilon_{j,i}} \frac{\delta s(1-s)}{1-\delta(1-s+s^2)} + \varepsilon_{j,i} \left(\frac{\delta s(1-s)}{1-\delta(1-s+s^2)}\right)^2 \\
\Leftrightarrow & \sqrt{\varepsilon_{j,i}} \left(1 - \left(\frac{\delta s(1-s)}{1-\delta(1-s+s^2)}\right)^2\right) \leq 2 \frac{\delta s(1-s)}{1-\delta(1-s+s^2)} \\
\Leftrightarrow & \varepsilon_{j,i} \leq 4 \left(\frac{\delta s(1-s)[1-\delta(1-s+s^2)]}{[1-\delta(1-s+s^2)]^2 - [\delta s(1-s)]^2}\right)^2
\end{aligned} \tag{48}$$

Denoting $\delta s(1-s)$ as a and $1-\delta(1-s+s^2)$ as b, (48) becomes

$$\varepsilon_{j,i} \leq 4 \left(\frac{ab}{b^2 - a2}\right)^2 \tag{49}$$

The first derivatives of (48) with respect to δ and s are given by

$$\begin{aligned}&\frac{\partial(48)}{\partial p} = 8\delta s^2(1-s)^2\times\\&\frac{(1-\delta)^3+\delta s(1-s)[3(1-\delta)^2+2\delta s(2(1-\delta)(1-s)+\delta s(1-s)^2)]}{((1-\delta)^2+2\delta(1-\delta)s(1-s))^3} > 0\end{aligned} \tag{50}$$

$$\begin{aligned}&\frac{\partial(48)}{\partial s} = 8\delta^2 s(1-s)(1-2s)\times\\&\frac{(1-\delta)^3+\delta s(1-s)[3(1-\delta)^2+2\delta s(2(1-\delta)(1-s)+\delta s(1-s)^2)]}{(1-\delta)^2[(1-\delta)+2\delta s(1-s)]^3}\\&>0 \text{ if } s<0.5, =0 \text{ if } s=0.5, <0 \text{ if } s>0.5\end{aligned} \tag{51}$$

Therefore, the maximum $\varepsilon_{j,i}$ that player i should replicate increases in δ, increases in s until it reaches its maximum at $s = 0.5$ after which is decreases in s, and increases in the probability for transfers $2s(1-s)$.

A.6 Proof VI

Assume that in at least one previous round, player j has made a positive transfer and revealed ${}_{rev}\varepsilon_{j,i}$. The current round is the first round in which i was successful, but j was not. If i makes a transfer that reveals $\varepsilon_{i,j} \leq_{rev} \varepsilon_{j,i}$, j will replicate this transfer in all future rounds. Any larger transfer revealing $\varepsilon_{i,j} >_{rev} \varepsilon_{j,i}$ will not change j's ${}_{rev}\varepsilon_{j,i}$ Now, in the current round, i keeps $\frac{R}{1+\varepsilon_{i,j}}$. In all future rounds, when both player are successful (probability s^2), i will receive R, when only i is successful (probability $s(1-s)$), he will keep $\frac{R}{1+\varepsilon_{i,j}}$ and when only j is successful (probability $s(1-s)$), j will copy i's transfer and choose $\varepsilon_{j,i} = \varepsilon i, j$, so that i will receive $\frac{\varepsilon_{i,j}R}{1+\varepsilon_{i,j}}$. Therefore, i's expected fitness is given by

$$EF_i = \sqrt{\frac{R}{1+\varepsilon_{i,j}}} + \delta s^2\frac{\sqrt{R}}{1-\delta} + \delta s(1-s)\frac{\sqrt{\frac{R}{1+\varepsilon_{i,j}}}+\sqrt{\frac{\varepsilon_{i,j}R}{1+\varepsilon_{i,j}}}}{1-\delta}. \tag{52}$$

The first derivative with respect to $\varepsilon_{i,j}$ is

$$\frac{\partial EF_i}{\partial \varepsilon_{i,j}} = -\frac{\sqrt{\varepsilon_{i,j}R}-\delta(1-s+s^2)\sqrt{\varepsilon_{i,j}R}+s(1-s)\sqrt{R}}{2(1-\delta)\sqrt{\varepsilon_{i,j}}(1+\varepsilon_{i,j})^{3/2}}, \tag{53}$$

which becomes 0 if

$$
\begin{aligned}
&\sqrt{\varepsilon_{i,j}R} - \delta(1-s+s^2)\sqrt{\varepsilon_{i,j}R} - \delta s(1-s)\sqrt{R} = 0 \\
\Leftrightarrow&\sqrt{\varepsilon_{i,j}}(1-\delta(1-s+s^2) == \delta s(1-s) \\
\Leftrightarrow&\varepsilon^*_{i,j} = \left(\frac{\delta s(1-s)}{1-\delta(1-s+s^2)}\right)^2 = \left(1 - \frac{1-\delta}{1-\delta(1-s+s^2)}\right)^2
\end{aligned}
\tag{54}
$$

The second derivative is

$$
\frac{\partial^2 EF_i}{\partial \varepsilon^2_{i,j}} = \frac{\sqrt{R}\left(3\varepsilon^{3/2}_{i,j}(1-\delta(1-s+s^2)) - 4\varepsilon_{i,j}(\delta s(1-s)) - \delta s(1-s)\right)}{4(1-\delta)\varepsilon^{3/2}_{i,j}(1+\varepsilon_{i,j})^{5/2}}.
\tag{55}
$$

The denominator is always positive and inserting $\varepsilon^*_{i,j}$ into the enumerator yields

$$
-\frac{\frac{(\delta s(1-s))^3}{1-\delta(1-s+s^2)} + \delta s(1-s)}{\delta s(1-s)},
\tag{56}
$$

which is always negative. Therefore, i optimally chooses $\varepsilon_{i,j} = min[\varepsilon^*_{i,j}, \varepsilon_{j,i}]$. If player j would replicate not only lower transfers, but higher transfers as well, player i would always choose $\varepsilon^*_{i,j}$.

The first derivatives of $\varepsilon^*_{i,j}$ with respect to δ and s are

$$
\frac{\partial \varepsilon^*_{i,j}}{\partial \delta} = \frac{2\delta s^2(1-s)^2}{(1-\delta(1-s+s^2))^3} > 0
\tag{57}
$$

$$
\frac{\partial \varepsilon^*_{i,j}}{\partial s} = \frac{2\delta^2(1-\delta)s(1-s)(1-2s)}{(1-\delta(1-s+s^2))^3}
\begin{cases} > 0 \text{ if } s < 0.5 \\ = 0 \text{ if } s = 0.5 \\ < 0 \text{ if } s > 0.5 \end{cases}
\tag{58}
$$

Therefore, the optimal $\varepsilon^2_{i,j}$ increases in δ, increases in s until it reaches its maximum at $s = 0.5$ after which is decreases in s, and increases in the probability for transfers $2s(1-s)$.

A.7 Proof VII

A first mover i in a population consisting entirely of replicators faces the same problem that the second type of conditional altruist faced as the first mover's expected fitness is also given by equation (52). Therefore, first mover i always reveals $\varepsilon^*_{i,j} = \left(\frac{\delta s(1-s)}{1-\delta(1-s+s^2)}\right)^2 = \left(1 - \frac{1-\delta}{1-\delta(1-s+s^2)}\right)^2$.

A.7.1 Proof VIII

A second mover j of the second type always replicates $\varepsilon^*_{i,j}$ because it is the optimal reaction. A second mover j of the first type replicates $\varepsilon^*_{i,j}$ if

$$\begin{aligned}
&\left(\frac{\delta s(1-s)}{1-\delta(1-s+s^2)}\right)^2 \leq 4\left(\frac{\delta s(1-s)[1-\delta(1-s+s^2)]}{[1-\delta(1-s+s^2)]^2-[\delta s(1-s)]^2}\right)^2 \\
\Leftrightarrow &\frac{\delta s(1-s)}{1-\delta(1-s+s^2)} \leq 2\frac{\delta s(1-s)[1-\delta(1-s+s^2)]}{[1-\delta(1-s+s^2)]^2-[\delta s(1-s)]^2} \\
\Leftrightarrow &\frac{1}{2} \leq \frac{[1-\delta(1-s+s^2)]^2}{[1-\delta(1-s+s^2)]^2-[\delta s(1-s)]^2}
\end{aligned} \tag{59}$$

Denoting $\delta s(1-s)$ as a and $1-\delta(1-s+s^2)$ as b, this expression becomes

$$\frac{1}{2} < \frac{b^2}{b^2-a^2} \tag{60}$$

which is always true. Therefore, a second mover of the first type will also always replicate $\varepsilon^*_{i,j}$.

A.8 Proof IX

If $\varepsilon_{i,j} = \varepsilon_{j,i} = 1$, each player's expected fitness each round is given by

$$EF_{\varepsilon=1} = s^2\sqrt{R} + s(1-s)\left(\sqrt{\frac{R}{2}} + \sqrt{\frac{R}{2}}\right). \tag{61}$$

If $\varepsilon_{i,j} = \varepsilon_{j,i} = \varepsilon^*$, each player's expected fitness each round is given by

$$EF_{\varepsilon^*} = s^2\sqrt{R} + s(1-s)\left(\sqrt{\frac{R}{1+\varepsilon^*}} + \sqrt{\frac{\varepsilon^* R}{1+\varepsilon^*}}\right). \tag{62}$$

The expected fitness of both players making transfers revealing $\varepsilon = 1$ is larger than both player making transfers revealing $\varepsilon = \varepsilon^*$ if

$$\begin{aligned}
&EF_{\varepsilon=1} > EF_{\varepsilon^*} \\
\Leftrightarrow &s^2\sqrt{R} + s(1-s)\left(\sqrt{\frac{R}{2}} + \sqrt{\frac{R}{2}}\right) > s^2\sqrt{R} + s(1-s)\left(\sqrt{\frac{R}{1+\varepsilon^*}} + \sqrt{\frac{\varepsilon^* R}{1+\varepsilon^*}}\right) \\
\Leftrightarrow &\sqrt{\frac{1}{2}} + \sqrt{\frac{1}{2}} > \sqrt{\frac{1}{1+\varepsilon^*}} + \sqrt{\frac{\varepsilon^*}{1+\varepsilon^*}} \\
\Leftrightarrow &2 > \frac{1 + sqrt\varepsilon^* + \varepsilon^*}{1+\varepsilon^*} \\
\Leftrightarrow &1 + \varepsilon^* > sqrt\varepsilon^*
\end{aligned} \tag{63}$$

which is always true.

A.9 Proof X

Assume player i and j have agreed to always transfer $0.5R$ (i.e. $\varepsilon_{i,j} = \varepsilon_{j,i} = 1$) when applicable. Now player i finds himself in the role of the first mover. Player i knows that player j will reduce her $\varepsilon_{j,i}$ to 0 if i makes a transfer lower than $0.5R$. If i transfers $0.5R$, j will also transfer $0.5R$ each round when applicable. The only two valid options for i are therefore honoring the agreement or not making any transfer at all. Honoring the agreement yields an expected fitness given by

$$EF_{\varepsilon_{i,j}=1} = \sqrt{0.5R} + \delta s^2 \frac{\sqrt{R}}{1-\delta} + 2\delta s(1-s)\frac{\sqrt{0.5R}}{1-\delta} \tag{64}$$

Not honoring the agreement yields an expected fitness given by

$$EF_{\varepsilon_{i,j}<1} = \sqrt{R} + \delta s \frac{\sqrt{R}}{1-\delta} \tag{65}$$

Honoring the agreement is preferable if

$$\begin{aligned}
& EF_{\varepsilon_{i,j}=1} \geq EF_{\varepsilon_{i,j}<1} \\
\Leftrightarrow & \sqrt{0.5R} + \delta s^2 \frac{\sqrt{R}}{1-\delta} + 2\delta s(1-s)\frac{\sqrt{0.5R}}{1-\delta} \geq \sqrt{R} + \delta s \frac{\sqrt{R}}{1-\delta} \\
\Leftrightarrow & \sqrt{0.5} + \frac{\delta s^2}{1-\delta} + \frac{\sqrt{2}\delta s(1-s)}{1-\delta} \geq 1 + \frac{\delta s}{1-\delta} \\
\Leftrightarrow & (1-\delta)(\sqrt{0.5}-1) + \delta\left(s^2 + \sqrt{2}s(1-s) - s\right) \geq 0 \\
\Leftrightarrow & \delta(1-\sqrt{0.5} + s(1-s)(\sqrt{2}-1)) \geq 1-\sqrt{0.5} \\
\Leftrightarrow & \delta \geq \frac{1-\sqrt{0.5}}{1-\sqrt{0.5}+s(1-s)(\sqrt{2}-1)}
\end{aligned} \tag{66}$$

If player i values future payoffs strongly enough, he will honor the agreement and transfer half of the resource. The first derivative with respect to s is given by

$$\frac{\partial(66)}{\partial s} = \frac{(2s-1)(1-\sqrt{0.5})(\sqrt{2}-1)}{\left(1-\sqrt{0.5}+s(1-s)(\sqrt{2}-1)\right)^2} \begin{cases} > 0 \text{ if } s < 0.5 \\ = 0 \text{ if } s = 0.5 \\ < 0 \text{ if } s > 0.5 \end{cases} \tag{67}$$

Therefore, the δ necessary to honor the agreement increases in s until it reaches its maximum at $s = 0.5$ after which is decreases in s, and increases in the probability for transfers $2s(1-s)$. If player i and player j have different factors with $\delta_i > \frac{1-\sqrt{0.5}}{1-\sqrt{0.5}+s(1-s)(\sqrt{2}-1)} > \delta_j$, player i will honor the agreement while player j will not (she may, however, still agree to the arrangement and defect later).

A.10 Proof XI

Player i has a higher probability for success than player j, i.e. $s_i > s_j$. If player i does not interact with player j, he receives $\sqrt{R}$ with probability s_i each round. Sharing resources equally with player j each round yields an expected fitness of

$$EF_{es} = s_i s_j \sqrt{R} + s_i(1-s_j)\sqrt{\frac{R}{2}} + s_j(1-s_i)\sqrt{\frac{R}{2}}. \tag{68}$$

Player i is better off not interacting with player j if

$$\begin{aligned} & s_i\sqrt{R} > s_i s_j \sqrt{R} + s_i(1-s_j)\sqrt{\frac{R}{2}} + s_j(1-s_i)\sqrt{\frac{R}{2}} \\ \Leftrightarrow & s_i > s_i s_j + \frac{s_i}{\sqrt{2}} - \frac{s_i s_j}{\sqrt{2}} + \frac{s_j}{\sqrt{2}} - \frac{s_i s_j}{\sqrt{2}} \\ \Leftrightarrow & s_i(\sqrt{2}-1+s_j(2-\sqrt{2})) > s_j \\ \Leftrightarrow & s_i > \frac{s_j}{\sqrt{2}-1+s_j(2-\sqrt{2})} \end{aligned} \tag{69}$$

For each s_j that satisfies $0 \leq s_j < 1$, the right hand side is always smaller than 1.

$$\begin{aligned} & 1 > \frac{s_j}{\sqrt{2}-1+s_j(2-\sqrt{2})} \\ \Leftrightarrow & \sqrt{2}-1+s_j(2-\sqrt{2}) > s_j \\ \Leftrightarrow & s_j(1-\sqrt{2}) > 1-\sqrt{2} \\ \Leftrightarrow & s_j < 1 \end{aligned} \tag{70}$$

Therefore, for each s_j that satisfies $0 \leq s_j < 1$, there exists a success probability s_i for which player i's fitness is higher when not interacting with player j than when agreeing to always transfer the equal split.