Personal Relationships and the Evolution of Trust *

Philipp Schliffke
University of Hamburg
Department of Economics
Von-Melle-Park 5
D-20146 Hamburg
philipp.schliffke@wiso.uni-hamburg.de

March 2009

No. 161

Abstract: In this paper, an indirect evolutionary version of the game of trust is studied. A population consisting of trustworthy and exploitive players is assumed. Players are chosen randomly from the population and are matched with either strangers or players they know in order to play the game of trust. Some fundamental conditions for the evolutionary success of trust and trustworthiness are derived. The results are extended to a situation in which the trustworthy players are not aware of the competitive nature of the game. The main result is that trustworthy players will not get repelled from the population as long as they are aware of the game. Further, it is shown that the vector fields of the dynamic process are discontinuous.

Keywords: trust, trustworthiness, indirect evolution, discontinuous vector fields

*The author thanks Stefan Napel, University of Bayreuth, as well as Manfred Holler, Martin Leroch and Andreas Nohn, University of Hamburg, for their encouragement and help.

1 Introduction

No doubt, trust is important but also delicate. In general, trusting another person requires sequential thinking as trust must be shown before another person decides whether to reward or to exploit it. If exploitation is advantageous and punishment is not feared, trust will hardly develop because it is simply rational to exploit shown trust. However, this kind of rationality is often questioned by actual behavior of individuals. Empirical and experimental studies have repeatedly shown that actual human behavior is driven by other motives than pure individual maximization of outcomes (see e.g. Fehr and Gächter 1998 or Fehr and Fischbacher 2005). Prominent explanations are inequality aversion in the sense of Fehr and Schmidt (1999) or Bolton and Ockenfels (2000), or the willingness to respond kindly to actions that are perceived as kind, a behavior which is usually referred to as reciprocity (see e.g. Falk and Fischbacher 2006). What these explanations have in common is that they all add an individual factor to the objective (in most cases monetary) outcome. If this individual factor is large enough, rational behavior can diverge from what an orientation on objective factors alone would predict: People sometimes do not care that exploiting others' trust will make them better off in pecuniary terms.

A crucial question is whether or not such behavior driven by non-monetary motives can succeed or, more radically, survive. In business life it is almost straight-forward to assume that a non profit maximizing behavior can not survive, at least if there are competitors taking advantage of it. On the individual level, one might ask whether people can be immune against monetary incentives. If people permanently observe that others are more successful in monetary terms, they may eventually switch to the behavior leading to more monetary success. Such an evolutionary perspective with respect to the game of trust and preference distinguished player types is picked up by e.g. Güth and Kliemt (1994 and 1998).[1] A key result is that trustworthy behavior will diminish, given that agents interact under uncertainty concerning other players types and occasionally tremble. Only if players can discriminate perfectly between trustworthy and non-trustworthy types will the trustworthy ones prevail. On the one hand, perfect information about another person is a theoretical myth, so one could argue that there is little hope for a long run success of trustworthy behavior. On the other hand, perfect information, as it is used in these models,

[1]The authors use the so called indirect evolutionary approach. Applications to other, e.g. ultimatum and dictator, games are e.g. Huck and Oechssler (1999) or Güth and Napel (2006).

only concerns knowing how the other agent will behave. The difference in information leads to a difference in strategic behavior. If we know that another person is non-trustworthy, we will never trust him in the first place. If we do not know another person, we might trust or not trust based on our expectation and experience with other people. Knowing the other person in this context means to have some experience that allows to apply a discriminating strategy. This is what is called a personal relationship. A personal relationship need not be private. It is sufficient that there is enough reliable information available that allows to discriminate between different player types.[2] By these assumptions, interaction under personal relationships captures repeated interaction as e.g. with family members, friends, colleagues, or frequent business partners. To some extent it also captures the application of trigger strategies such as the famous tit-for-tat discussed in Axelrod (1984). Clearly, in real life interactions take place both with others to whom there is a personal relationships and with strangers.

The present paper sets up a model capturing all aspects pointed out above. The strategic dilemma of the basic trust game and the existence of multiple player types. The force of evolution driven by material payoffs and the fact that interaction takes place with and without personal relationships. The main question is whether or not and, if yes, under which conditions trust and trustworthiness can prevail. The analysis is carried out using the indirect evolutionary approach. The model description is provided in section 2. The parallel game of interaction with and without the application of a discriminating strategy is introduced in section 3. Section 4 checks the results for robustness against trembles and section 5 introduces trustworthy player types that have a tendency to trust or misinterpret the competitive nature of the game. Section 6 concludes. The main finding is that the fraction of trustworthy players in the population will not vanish in the parallel game. Rather, the evolutionary stable outcomes are either a population state where all players are trustworthy, or an interior solution with both types present. The results are robust against trembles for a wide variety of payoff structures. A tendency to trust by the trustworthy players can, however, lead to an overall decline in their fraction.

[2]For instance, Bolton, Katok, and Ockenfels (2004) and Güth, Mengel, and Ockenfels (2007) have studied the (generally positive) effect of reputation mechanisms in internet trade with anonymous interaction.

2 The model

The basic game studied is the two player sequential game of trust. Player A is the first mover and has to decide whether to trust (denoted by T) or not trust (N) in his co-player.[3] If she does not trust, the game ends and both players receive an identical payoff of s. If the choice is to trust, i.e. play T, the second mover player B decides whether to reward (R) or exploit (E) the trust shown by the first mover. Rewarding trust yields an identical payoff of r to both players while exploitation yields a payoff of 1 to the second mover and 0 to the first mover. The payoffs represent pecuniary or objective payoffs. It is assumed that $1 > r > s > 0$. Thus, the game is a social dilemma game because the only Nash equilibrium is (N, E), yielding the payoff vector (s, s) while cooperative behavior would lead to the pareto superior payoff vector (r, r). The basic game is embedded in an evolutionary process driven by the objective payoffs and played by a possibly infinitely large population. Nature moves at the beginning of each stage and randomly matches the agents in pairs to play the game. Each player is assigned the role of either first or second mover, both with probability 1/2. Time passes continuously from 0 to t with $t \in [0, +\infty)$.

The indirect evolutionary version of the game is studied in Güth and Kliemt (1994), Güth and Kliemt (1998) and Güth, Kliemt, and Napel (2006). Each player is endowed with a preference parameter m_i which affects the subjective evaluation of the pecuniary payoff of exploiting trust. The utility of the second mover associated with the choice E is defined as $u_{i,2}(\cdot, E) = 1 - m_i$. For all other actions utility is simply given as $u(\pi) = \pi$. m_i is a purely private parameter, basically defined on $\mathbb{R}$ and possibly heterogeneously distributed among agents. A preference parameter $m_i > 1 - r$ is denoted as m_R. In this case, the utility of the choice R exceeds the utility of the choice E and players with a preference parameter m_R therefore rationally choose R in second mover position. These players are called trustworthy or rewarding types and are denoted by θ_R. Consequently, m_E labels a preference parameter $m_i < 1 - r$. The influence of m_E is too weak to affect the choice compared to the basic game without such parameters. Players with a preference parameter m_E are called non-trustworthy or exploitive types and are denoted by θ_E.

The introduction of m_i turns the game into a Bayesian game with two types of players.

[3] In experimental studies as e.g. Bolle (1998), player A is endowed with a certain amount of money. A then decides how much to pass to the second mover. The experimenter multiplies the amount and gives it to the second mover who has to decide how much of the enlarged amount to return to the first mover.

A player faces a game of incomplete information if he or she does not know the preference parameter of the co-player. Following Harsanyi (1967), the game can be transformed into a game with complete but imperfect information by introducing additional moves of nature. Assuming that the agents know the exact fractions of types, players will base their decisions on that knowledge. If a player has a personal relationship to his co-player and if this relationship allows him to discriminate between trustworthy and exploitive types, the behavior is the same as in a game of perfect information, i.e. the player behaves as if she knows the type of her co-player. Nature carries out three moves at the beginning of each stage. First, the population is separated into two groups. A λ-fraction ($0 < \lambda < 1$) is chosen to play the game of complete, in fact, perfect information while a fraction of $(1-\lambda)$ plays in the imperfect information setting. Second, nature assigns the type of player A and B according to the distribution of types in the population and third, it assigns first and second mover positions. The game tree in Figure 1 illustrates the situation.[4]

A fundamental assumption of the indirect evolutionary approach is that players behave fully rational with respect to the current stage game and with respect to their utility (not with respect to pecuniary payoffs). As a consequence, second mover behavior is completely determined by the preference parameter m_i. First mover behavior is either determined by the type of the co-player, if known, or by the expected payoffs of the choices T and N calculated on the basis of the fraction of trustworthy players in the population. Rationality with respect to the current stage game means that the evolutionary process, solely driven by material success, is not considered by the agents at all. Although the evolutionary process is not considered by the agents, it may indirectly influence the outcome. Under imperfect information, the strategies are population state dependent because the population state influences the expected first mover payoff and hence first mover behavior. So, evolution might influence behavior and behavior might influence evolution. This is what makes the approach indirect.

The position in the game tree is an element of the relevant information set to each player. In the case of perfect information, the co-players type is added, in the case of imperfect information, the fraction θ_R^t of trustworthy players. The strategy for a trustworthy player

[4]The lower arms of the tree ending with "... " represent those subgames where player B moves first. The upper arms of the tree ending with "..." are those where player A is an exploitive type. Actually, it is assumed that imperfect information is only present with respect to the co-players type. It is assumed that a player always knows his own type. Therefore, the dashed line indicating imperfect information is drawn between two subgames with an identical first mover type.

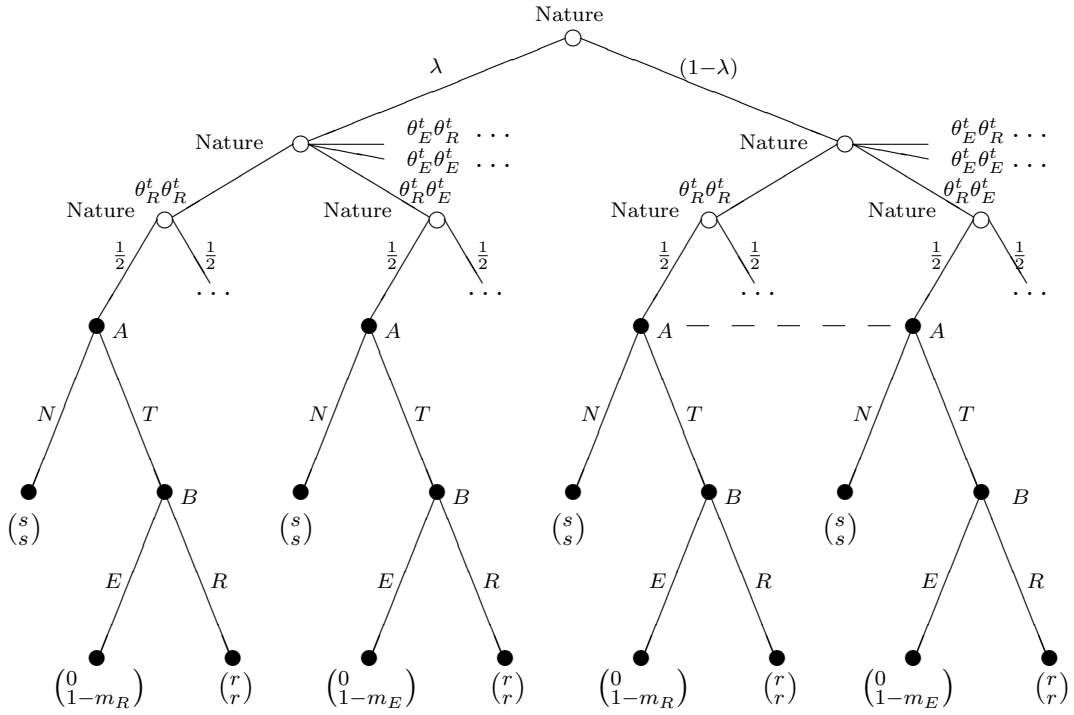

Figure 1: A stage of the parallel game

under perfect information is hence given as:

$$s(\theta_R): \left((1^{st}, \theta_{j,2}) \longmapsto \begin{cases} T & \text{if } \theta_{j,2} = \theta_R \\ N & \text{if } \theta_{j,2} = \theta_E \end{cases}, \; 2^{nd} \longmapsto R\right) \tag{1}$$

A trustworthy player facing incomplete information uses a Bayesian Nash strategy based on the expected payoff he receives from the first mover choices T and N. The threshold value of the current fraction of trustworthy types at which the expected payoffs from playing N and T are equal is denoted by d. Note that θ_R^t is treated as being exogenous, so rationality implies:

$$s(\theta_R): \left((1^{st}, \theta_R^t) \longmapsto \begin{cases} T & \text{if } \theta_R^t > d \\ N & \text{if } \theta_R^t < d \end{cases}, 2^{nd} \longmapsto R\right) \tag{2}$$

First mover behavior is identical for both players while an exploiting type always chooses E in second mover position. The behavior at d is left undefined.[5]

[5]It can be assumed that players choose some mixed action $[0, 1]$ at d but it is pointed out below that this is generally irrelevant.

Evolution is analyzed using replicator dynamics. It is sufficient to study one equation (which will be the one for the trustworthy types) because a gain for one type fraction equals a loss in the other fraction so all changes add up to zero and all solutions lie on the projection $\Delta = [0, 1]$. Each point in Δ corresponds to a vector $\theta^t = (\theta_R^t, \theta_E^t)$ called a population state. It will generally be assumed that both types are present in the population initially so the dynamics start in the interior of Δ, i.e. $\theta^0 \in int(\Delta)$. If the objective payoff of a type is larger than the average payoff in the population, that type spreads. Otherwise it declines. In the case of only two existing types, the general equation can be simplified to:

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^1 = \theta_R^t(1 - \theta_R^t)\left[\pi(\theta_R, \theta^t) - \pi(\theta_E, \theta^t)\right] \tag{3}$$

$\pi(\theta_R, \theta^t),\ \pi(\theta_E, \theta^t)$ are the expected payoffs for each strategy given the current population state. The superscript "1" indicates a game of perfect information and a superscript "0" will be used to label imperfect information. In the parallel game the population is split into two separate groups in each stage. Each group plays the game independently of the other. The joint replicator equation is hence the weighted sum of the equations of both (information) settings. The superscript "λ" indicates the parallel game. The general form is given as follows:

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^\lambda = \lambda\left(\frac{\partial \theta_R^t}{\partial t}\right)^1 + (1 - \lambda)\left(\frac{\partial \theta_R^t}{\partial t}\right)^0 \tag{4}$$

The replicator equations in the parallel game are piecewise defined functions. It turns out that the associated vector field φ_R is discontinuous at the threshold level d stated in the strategies (this is shown in the analysis). At d, the limits from the left and right of φ_R as θ_R^t approaches d do not coincide which implies that there is no well defined phase (i.e. a replicator equation does not exist). This is quite unusual and, to some extend, problematic. Therefore, I will shortly deal with the question of behavior and especially stability at d. First, the two domains of definition of the vector field φ_R are $[0, d)$ and $(d, 1]$. Now, one case that may occur is that all phases have the same sign. Clearly, one would expect that the process will then reach the boundary $bd(\Delta)$ implied by the signs, i.e. either a fraction of trustworthy players equal to 1 or 0. However, as the replicator equation, i.e. the time derivative, at d does not exist, the case that the process halts at d can not be excluded. Let $\xi_R(t, \theta^0)_{t\to\infty}$ be the solution mapping for the trustworthy types. To "solve" the problem, note

Assumption 1 *If all phases in* $\varphi_{R}|_{[0,d)}$ *and* $\varphi_{R}|_{(d,1]}$ *have the same sign, then for all* $\theta_R^0 \in int(\Delta)$, $\xi_{R}(t,\theta^0)_{t\to\infty} \to bd^*(\Delta)$ *with* $bd^*(\Delta)$ *the boundary determined by the signs of* φ_{R}.

So it is just assumed that the process will not halt at d.[6] There are two more possible cases that may arise. The vector fields can have opposing signs or at least one of them can be equal to zero. In the first case, the system will converge to d for any initial value starting in the interior so one naturally thinks of asymptotic stability. In the second case, the system will at least not move further away from d which is typically associated with Lyapunov stability. Stability is generally acompanied by a zero phase but in the present case, there is no phase defined and the limits from the left and right are often unequal to zero. To ensure that the standard definitions of Lyapunuv- and asymptotic stability apply, recall first that Lyapunov stability means that for a neighborhood B around d containing some neighborhood B^0, all forward orbits from B^0 are contained in B, i.e. $\gamma^+(B^0 \cap \Delta) \subset B$ (see Weibull 1996, p. 244). Asymptotic stability additionally requires that the system always converges to $d \in B^0$.

Proposition 1 *If* $\varphi_{R}|_{[0,d)} \geq 0$ *and* $\varphi_{R}|_{(d,1]} \leq 0$, *then* $\xi_{R}(t^*,d) = d$ *for all* $t \geq t^*$ *and* d *is Lyapunov-stable. If both inequalities hold strict,* d *is asymptotically stable.*

Proof. Assume $\xi_{R}(0,\theta) = d$ and $\xi_{R}(t,\theta) \neq d$ for some $t > 0$. Then, from the mean value theorem, it follows that $\varphi_{R}|_{d'} < 0$ for some $d' < d$ or $\varphi_{R}|_{d'} > 0$ for some $d' > d$. Contradiction. Hence, d is stationary and, by assumptions on φ_{R}, Lyapunov or asymptotically stable. ∎

Hence, although the vector fields are not continuous, the standard definitions of Lyapunov- and asymptotic stability are still applicable.

3 The parallel game

Following the setup of the model, the replicator system for the parallel game is derived by combining the replicator equations of the perfect and imperfect information setting. In the perfect information case, all players may be in first or second mover position, each case occurring with probability 1/2. According to their strategy, both types will trust in the

[6] The assumption seems reasonable because every perturbed process would imply that solution.

trustworthy players which have a population share of θ_R^t. Facing a trustworthy player, each agent will receive r. If, however, the first mover faces an exploitive type, he will assure himself a payoff of s, irrespective of his own type. The trustworthy types will always be trusted and they will always reward. Thereby, their second mover expected payoff is r while it is s for the exploitive types. The expected payoffs are given as

$$\pi(\theta_R, \theta^t) = 1/2\left(\theta_R^t r + (1-\theta_R^t)s\right) + 1/2r \tag{5}$$

$$\pi(\theta_E, \theta^t) = 1/2\left(\theta_R^t r + (1-\theta_R^t)s\right) + 1/2s \tag{6}$$

The replicator equation for the trustworthy types is obtained by plugging the expected payoffs into equation (3). For the sake of simplicity, the term $\frac{1}{2}\theta_R^t(1-\theta_R^t)$ is abbreviated by α.

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^1 = \alpha(r-s) \tag{7}$$

This function is clearly positive for all θ_R^t, as, by assumption, $\theta^0 \in int(\Delta)$ and $r > s$. Thus, the fraction of trustworthy types approaches 1 if the players can perfectly discriminate between both player types.[7] In the case of imperfect information actions are dependent on a threshold population share of trustworthy types which yields two different cases. The expected payoff from trusting as a first mover is $\pi_{i,1}(T, \cdot) = \theta_R^t r$, so the choice T is only rational if $\theta_R^t > s/r$. Hence, the threshold level d stated in the strategies is $\theta_R^t = s/r$. For $\theta_R^t < s/r$, no player shows trust and all players receive the same payoff, s. Hence the replicator equation is zero. The case is different for $\theta_R^t > s/r$. The expected payoffs are again the sum of interaction as first and second mover. In second mover position, the rewarding types will receive r and the exploitive types will receive 1. The expected payoffs are given as:

$$\pi(\theta_R, \theta^t) = 1/2\theta_R^t r + 1/2r \tag{8}$$

$$\pi(\theta_E, \theta^t) = 1/2\theta_R^t r + 1/2 \tag{9}$$

The replicator equation for the trustworthy types under imperfect information is the piecewise defined function:

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^0 = \begin{cases} 0 & \text{if } \theta_R^t < s/r \\ \alpha(r-1) & \text{if } \theta_R^t > s/r \end{cases} \tag{10}$$

[7] This result is equivalent to *Theorem 3.1* in Güth and Kliemt (1994).

For all $\theta_R^t > s/r$, this function is strictly negative due to assumptions $\theta^0 \in int(\Delta)$ and $1 > r$. The fraction of trustworthy types will therefore converge toward $\theta_R^t = s/r$.[8] Note that $s/r \neq 0$ or 1 and hence $\alpha > 0$. Note further that $r - 1 \neq 0$. Hence, the limit of φ_R from the left as θ_R^t approaches d is zero, but the limit from the right is not and so the vector field is discontinuous at that point. Besides this discontinuity, the population state $\theta_R^t = s/r$ is Lyapunov stable according to Proposition 1.

The replicator equation for the parallel game is obtained by combining functions (7) and (10) via the weighting parameter $\lambda \in (0, 1)$. The result is, again, a piecewise defined function with a vector field discontinuity at the threshold level of trustworthy types:

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^{\lambda} = \begin{cases} \alpha\lambda(r-s) & \text{if } \theta_R^t < s/r \\ \alpha\left[\lambda(1-s) + r - 1\right] & \text{if } \theta_R^t > s/r \end{cases} \tag{11}$$

The first part of this function is strictly positive as α, λ, and $(r - s)$ are strictly positive. Hence, for any population share of trustworthy types below the threshold level, their fraction approaches $\theta_R^t = s/r$ irrespective of λ. In the second case, $\theta_R^t > s/r$, the dynamics are λ-dependent. The critical λ-value is obtained by setting the second part of the replicator equation equal to zero which yields:

$$\lambda^* = \frac{1-r}{1-s} \tag{12}$$

Clearly, if both parts of equation (11) are positive, then the fraction of trustworthy types approaches 1. Note however, that due to the discontinuity of the vector field at $\theta_R^t = s/r$ $(= d)$ we need Assumption 1 to ensure that the system crosses d if $\theta_R^0 < d$ initially. If the second part of the replicator equation is negative, then the fraction of trustworthy types will converge to the interior solution $\theta_R^t = s/r$.

Theorem 1 *For all $\theta^0 \in int(\Delta)$, the fraction of trustworthy types approaches 1 if $\lambda > \lambda^*$ and $\theta_R^t = s/r$ if $\lambda < \lambda^*$.*

Proof. Let $\lambda = \lambda^* + \eta$ and let $\eta \in \mathbb{R}$ be a sufficiently small real number so that $\lambda \in (0, 1)\ \forall \eta$. The replicator equation for the trustworthy types in the case $\theta_R^t > s/r$ becomes

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^{\lambda}_{\theta_R^t > s/r} = \alpha\eta(1-s) \tag{13}$$

[8] This result is equivalent to *Theorem 3.2* in Güth and Kliemt (1994). The authors also show that the introduction of trembles leads to an overall decline in the fraction of the trustworthy types.

Since $1 > s$, the equation is positive if $\eta > 0$ and negative if $\eta < 0$. ■

Both possible results are asymptotically stable. However, asymptotic stability of the interior solutions is based on Proposition 1. The following bifurcation diagram with $s = .33$ and $r = .66$ summarizes all results of this section.[9]

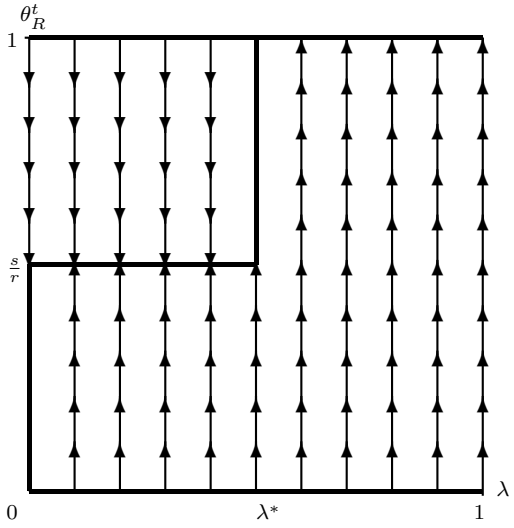

Figure 2: Bifurcation diagramm for equation (11)

For θ_R^t equal to 0 or 1, all states are stationary. In the isolated imperfect information case ($\lambda = 0$), θ_R^t converges to s/r if initially above that value. If $\lambda = \lambda^*$ the fraction of trustworthy types converges to s/r if initially below that value. In all other cases, the system develops as described in Theorem 1. Four results should be pointed out. First, trust and trustworthiness can prevail in a combined version of perfect and imperfect information. Second, even if interaction in personal networks is rarely frequent, the fraction of trustworthy players converges toward an interior solution. Third, λ^* and s/r both increase in s and decrease in r. Hence, there is a kind of trade-off between a more risky environment ("large" λ^* so the probability that $\lambda < \lambda^*$ increases) and a higher security level s/r (players switch to non trusting behavior). Finally, it seems hard to reasonably find out which payoff combinations stand for which kind of situations. But note, whenever the payoff distance $\|1 - r\|$ equals the distance $\|r - s\|$, then $\lambda^* = 1/2$. So it is just the more frequent form of interaction that decides if trust and trustworthiness prevail.

[9]Loosely speaking, a bifurcation occurs whenever the qualitative nature of the solution radically changes when certain parameter values are crossed (see e.g. Hirsch, Smale, and Devaney 2004, chap. 1, 15). The bifurcation diagrams plot several vertical phase lines for different parameter values of λ.

4 The parallel game with trembles

It was argued in the introduction and model description, that personal relationships allow players to discriminate between both player types just as if information were perfect. Of course, players may believe they know the type of their co-player, but, in fact, they do not. Modeling such failures in judgment will amount to the introduction of an error probability ε for the perfect information setting. A misjudgment under perfect information is formally identical to an unintentional choice, i.e. a tremble in the sense of Selten (1975). Trembles under imperfect information are taken into account as well. However, just as in e.g. Güth and Kliemt (1994), only first mover trembles are considered. Second movers are asked to make such a simple choice that trembles seem relatively unreasonable.

The replicator equation for the parallel game with trembles is derived just as in the last section by building the replicator equations for the perfect and imperfect information setting and combining them via λ. Note that in the replicator equations the expected payoffs of both types are subtracted from one another. Since first mover payoffs are identical, it is sufficient to derive the second mover expected payoffs. Under perfect information, a trustworthy player will be trusted with probability $(1-\varepsilon)$ and not trusted with probability ε. An exploitive player receives s with probability $(1-\varepsilon)$ and 1 with probability ε. Second mover expected payoffs under perfect information are:

$$\pi_2(\theta_R, \theta^t) = 1/2\,[(1-\varepsilon)r + \varepsilon s] \tag{14}$$

$$\pi_2(\theta_E, \theta^t) = 1/2\,[(1-\varepsilon)s + \varepsilon] \tag{15}$$

Imperfect information payoffs are derived in a similar way. Again, there are two cases. For $\theta^t_R > s/r$, second mover expected payoffs are given as:

$$\pi_2(\theta_R, \theta^t) = 1/2\,[(1-\varepsilon)r + \varepsilon s] \tag{16}$$

$$\pi_2(\theta_E, \theta^t) = 1/2\,[(1-\varepsilon) + \varepsilon s] \tag{17}$$

and for $\theta^t_R < s/r$:

$$\pi_2(\theta_R, \theta^t) = 1/2\,[(1-\varepsilon)s + \varepsilon r] \tag{18}$$

$$\pi_2(\theta_E, \theta^t) = 1/2\,[(1-\varepsilon)s + \varepsilon] \tag{19}$$

The replicator equation for the perfect information takes the following form:

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^1 = \alpha\left[r - s - \varepsilon(1 + r - 2s)\right] \tag{20}$$

and the replicator equation for the imperfect information setting is the piecewise defined function:

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^0 = \begin{cases} \alpha\varepsilon(r-1) & \text{if } \theta_R^t < s/r \\ \alpha(1-\varepsilon)(r-1) & \text{if } \theta_R^t > s/r \end{cases} \tag{21}$$

Finally, the replicator equation for the parallel game with trembles is derived by combining equations (20) and (21) which yields:

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^\lambda = \begin{cases} \alpha\left[\lambda(1-2\varepsilon)(1-s) - (1-\varepsilon)(1-r)\right] & \text{if } \theta_R^t < s/r \\ \alpha\left[\lambda(1-2\varepsilon)(r-s) - \varepsilon(1-r)\right] & \text{if } \theta_R^t > s/r \end{cases} \tag{22}$$

First, one can observe that both parts of the imperfect information replicator equation (21) are strictly negative, given that $\alpha, \varepsilon > 0$ and $r < 1$. Hence, the overall fraction of trustworthy types must approach zero if the perfect information equation (20) turns negative.

Lemma 1 *The fraction of trustworthy types will approach zero under perfect information if $\varepsilon > \varepsilon'$ with*

$$\varepsilon' = \frac{r - s}{1 + r - 2s} \tag{23}$$

Proof. Let $\eta \in \mathbb{R}$ be a sufficiently small real number and let $\varepsilon = \varepsilon' + \eta$ with $\eta > 0$. The replicator equation for the trustworthy types under perfect information and trembles becomes

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^1 = \alpha(-\eta)\ (1 + r - 2s) \tag{24}$$

This function is strictly negative as, by assumption, $1 > r > s$, $\eta > 0$ and $\theta^0 \in int(\Delta)$. ■

So if $\varepsilon > \varepsilon'$ any combination of the perfect and imperfect information setting must lead the exploitive types to prevail. Now, and in difference to the case with no trembles, there are two critical values of λ leading to bifurcations. Both are obtained by setting the first,

respectively second, part of equation (22) equal to zero. The threshold level for $\theta_R^t < s/r$ is denoted by λ^{**} and given as

$$\lambda^{**} = \frac{\varepsilon(1-r)}{(1-2\varepsilon)(r-s)} \tag{25}$$

The threshold value for $\theta_R^t > s/r$ is denoted λ^{***} with

$$\lambda^{***} = \frac{(1-\varepsilon)(1-r)}{(1-2\varepsilon)(1-s)} \tag{26}$$

By some manipulation of equations, it is possible to show that whenever $\varepsilon < \varepsilon'$, then $\lambda^{**} < \lambda^{***}$, i.e. some possible case dependencies drop out. Therefore, provided that $\varepsilon < \varepsilon'$ and neglecting cases where λ is equal to one of the threshold levels, there are three different cases to consider. The cases depend on the relation of the true value of λ to both threshold levels λ^{**} and λ^{***} identified above.

Theorem 2 *Provided $\varepsilon < \varepsilon'$: For all $\theta^0 \in int(\Delta)$, the fraction of trustworthy types approaches 1 if $\lambda > \lambda^{***}$, s/r if $\lambda^{**} < \lambda < \lambda^{***}$, and 0 if $\lambda < \lambda^{**}$.*

Proof. The proof consists of two parts. First it is shown that θ_R^t converges to either 1 or s/r whenever $\lambda > \lambda^{**}$. Then it is shown that θ_R^t declines toward 0 if $\lambda < \lambda^{**}$.

λ^{***} is a perturbed version of λ^* stated in section 3. The perturbation term $\frac{1-\varepsilon}{1-2\varepsilon}$ is strictly positive because $\varepsilon < 1/2$ if $\varepsilon < \varepsilon'$.[10] Thereby, the proof of Theorem 1 in section 3 will be qualitatively unaffected. So given that $\varepsilon < \varepsilon'$ and hence $\lambda^{**} < \lambda^{***}$, it follows directly that the fraction of trustworthy types will approach 1 whenever $\lambda > \lambda^{***}$ and converges to $\theta_R^t = s/r$ whenever $\lambda^{**} < \lambda < \lambda^{***}$.

If $\lambda < \lambda^{**}$, transitivity implies that $\lambda < \lambda^{***}$. So θ_R^t must approach s/r if initially above that level. If the replicator equation for $\theta_R^t < s/r$ is negative, then Assumption 1 guarantees that s/r is passed and θ_R^t will decline toward zero. Now, let $\lambda = \lambda^{**} + \eta$ and let $\eta \in \mathbb{R}$ be a sufficiently small real number so that $\lambda \in (0,1)\ \forall \eta$. The replicator equation for the trustworthy types in the case $\theta_R^t < s/r$ becomes

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^{\lambda} = \alpha\eta(1-2\varepsilon)(r-s) \tag{27}$$

This equation is, by assumptions, strictly negative if $\eta < 0$ which is required in the theorem. Hence, if $\lambda < \lambda^{**}$, the fraction of trustworthy players diminishes. ∎

[10] Recall that $\varepsilon' = \frac{r-s}{1+r-2s}$. Then I state $\frac{1}{2} > \frac{r-s}{1+r-2s} \Leftrightarrow 1 > r$ which is true.

Each λ implies a unique outcome. Again, the vector field is discontinuous at $\theta_R^t = s/r$ so Assumption 1 may be needed to reach the boundaries of Δ, or, if the system develops toward the interior solution, Proposition 1 is required to state asymptotic stability. Trembles or misjudgments may lead to an overall decline of trustworthy players. To get a better impression of the destabilizing influence of trembles, the following diagrams illustrate the values of the critical tremble probability ε' and the critical frequency of interaction under personal relationships λ^{**} for all possible values of s and r.

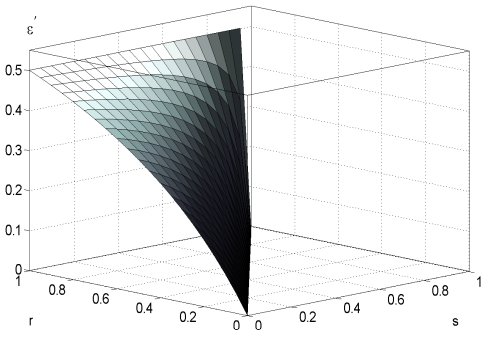

Figure 3: Tremble probability ε'.

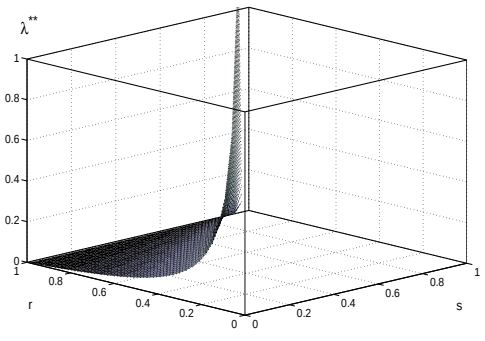

Figure 4: Frequency λ^{**} with $\varepsilon' = 0.05$

Both figures illustrate that trust and trustworthiness are relatively stable against trembles in a combined perfect imperfect information environment. In figure 3, the surface is the critical upper level of the tremble probability ε, so if trembles occur with a smaller probability, then the development of θ_R^t depends on λ. For reasonable tremble probabilities of e.g. 2%, 5%, or maybe even 10%, a sure decline in the fraction of trustworthy types will only occur if the distance $||r-s||$ gets very "small". A similar result is highlighted in figure 4. The surface is the lower boundary for the interaction frequency under personal relationships given a tremble probability of 5%. If the actual value of λ is below the surface, then trustworthiness will diminish. However, one observes that this sure decline generally requires very low interaction frequencies as long as $||r-s||$ is not too "small". Hence, as long as the tremble probability is reasonably low and the payoffs are well specified, the positive support from interaction under perfect information can outweigh the effects of trembles. In diffenerce to what was found in the last section, there is not the kind of trade-off between a more risky environment and a higher security level. If s/r gets closer to 1, due to a smaller distance $||r-s||$, then the system also enters the regions where trembles indeed destabilize

the population of trustworthy players. The following bifurcation diagram summarizes the results with parameters set to $s = .33$, $r = .66$ and $\varepsilon = .05$.

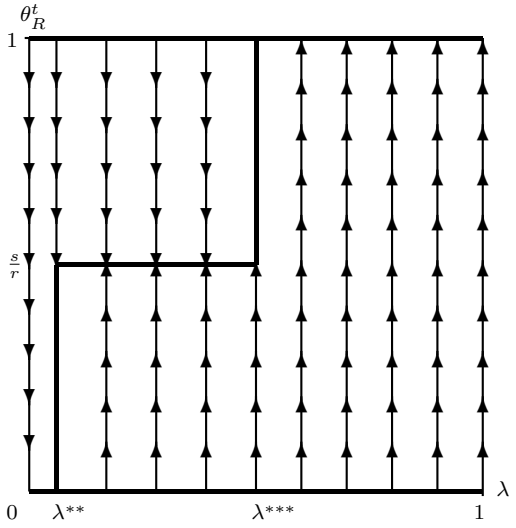

Figure 5: Bifurcation diagramm for equation (22)

5 Unawareness and a tendency to trust

This section will analyze the evolution of types if the trustworthy players are unaware about the competitive nature of the game, or, if they have a tendency to trust. From the perspective of a trustworthy player and in terms of utility, the game is cooperative as rewarding yields a superior outcome compared to exploiting in second mover position. This perspective may lead to two different kinds of failures. One possibility is that the trustworthy players do not recognize that the game is competitive even in utility terms (it is competitive in pecuniary terms anyway, but this is ignored by all players). Of course, if they recognize that there are two distinct player types and that the non-trustworthy players will always exploit in second mover position, then they understand that the game is competitive in utility terms. However, from their perspective it might happen that they judge exploitive second mover behavior as unintentional or at least expect that those players will soon recognize that cooperative play is "superior". A second kind or failure is based on psychological aspects. Players with a preference for the cooperative solution may simply underestimate or even ignore the competitive nature because they "want" the game to be cooperative. They possibly ask themselves why those other people do not understand

that they are all better off if they cooperate. Alternatively, they might refuse to accept that the world is competitive. Both aspects, unawareness or unconsciousness about the competitive nature of the game and possible psychological aspects, lead to a tendency to trust by the trustworthy players. As a consequence, they will trust in first mover position although rationality implies to choose not to trust. This version is a possible setup of the game with bounded rational players in the sense of Gigerenzer and Selten (2002) where players apply simple heuristic rules in complex environments. In this light, the imperfect information environment is complex and the trustworthy agents react to this complexity by carrying out that choice, that appears to be the natural one. It is not assumed that players are irrational. If they know that their co-player is not trustworthy, they will not trust as a first mover. Hence, the strategies for perfect information and the strategy by the exploitive players for imperfect information are left unchanged. In contrast, the strategy of the trustworthy players under imperfect information is now defined as

$$s(\theta_R) : \left((1^{st}) \longmapsto T, (2^{nd}) \longmapsto R\right) \quad . \tag{28}$$

Obviously, for all population states with a fraction of trustworthy players above s/r, nothing changes. Therefore, the second part of equation (11) derived in section 3 is unaffected and the system behaves dependent on λ^* as described in Theorem 1. What is affected are the expected payoffs for $\theta_R^t < s/r$. Now, a trustworthy first mover will be rewarded with probability θ_R^t and exploited with probability $(1 - \theta_R^t)$. In second mover position, he will only get to move if the first mover is a trustworthy type. Otherwise, the game ends immediately. The exploitive types will never show trust in first mover position and hence receive a payoff of s. In second mover position, they can exploit a θ_R^t-fraction of the population because that fraction of players will (always) trust them. They do not get the chance to act and receive s if the first mover is an exploitive type as well. The expected payoffs are given as:

$$\pi(\theta_R, \theta^t) = 1/2 \left[2\theta_R^t r + \left(1 - \theta_R^t\right) s\right] \tag{29}$$

$$\pi(\theta_E, \theta^t) = 1/2 \left[s + \theta_R^t + \left(1 - \theta_R^t\right) s\right] \tag{30}$$

The replicator equation for the trustworthy types under imperfect information becomes:

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^0 = \alpha \left(2\theta_R^t r - \theta_R^t - s\right) \tag{31}$$

Combining equation (31) with equation (7) (perfect information) via λ and adding the equation for the case $\theta_R^t > s/r$ yields the replicator equation for the parallel game with "unaware" trustworthy players:

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^{\lambda} = \begin{cases} \alpha\left[\theta_R^t(1-\lambda)(2r-1) + \lambda r - s\right] & \text{if } \theta_R^t < s/r \\ \alpha\left[\lambda(1-s) + r - 1\right] & \text{if } \theta_R^t > s/r \end{cases} \tag{32}$$

The first part of the equation is in a stationary state whenever $\lambda = \lambda'$ with[11]

$$\lambda' = \frac{s - \theta_R^t(2r-1)}{r - \theta_R^t(2r-1)} \tag{33}$$

In order to avoid too much case dependencies, it is assumed that $r > 1/2$. Note

Lemma 2 *For $\theta_R^t < s/r$, θ_R^t approaches s/r if $\lambda > \lambda'$ and 0 if $\lambda < \lambda'$.*

Proof. Let $\eta \in \mathbb{R}$ be a small real number and let $\lambda = \lambda' + \eta$. The replicator equation for $\theta_R^t < s/r$ becomes

$$\left(\frac{\partial \theta_R^t}{\partial t}\right)^{\lambda} = \alpha(-\eta)\left[\theta_R^t(r-1) + r(\theta_R^t - 1)\right] \tag{34}$$

The inner square bracket term is strictly negative by assumptions $\theta_R^t \in int(\Delta)$ and $r \in (1/2, 1)$ and η enters the relationship with negative sign. Hence, the overall relationship will be positive if $\eta > 0$ and negative if $\eta < 0$ which proves the Lemma. ■

The overall behavior of the system will depend on the position of λ with respect to λ^* and λ'. For example, if $\theta_R^0 > s/r$ and $\lambda < \lambda^*$, then the fraction of trustworthy players will approach $\theta_R^t = s/r$. Now it depends on the relationship of λ to λ' whether the fraction falls to zero or whether the interior state is stable. Note that λ^* can be greater or smaller than λ' and that λ' is a (decreasing) function of θ_R^t. Summarizing all cases yields:

Theorem 3 *For all $\theta^0 \in int(\Delta)$, the fraction of trustworthy players will approach 1 whenever $\lambda > max\{\lambda^*, \lambda'\}$, s/r if $\lambda' < \lambda < \lambda^*$ and 0 if $\lambda < \lambda'$.*

[11] It is of course also possible to solve with respect to a value θ'. However, doing so and plugging θ' into λ' yields $\lambda' = \lambda$. Hence, if the system is in a restpoint because $\theta_R^t = \theta'$, then λ must have the value λ'. It also works the other way around. If $\lambda = \lambda'$, then $\theta_R^t = \theta'$. Therefore, it is not necessary to solve for a value θ' and consider cases where θ_R^t is above or below that value.

Proof. The theorem just combines the results of Theorem 1 and Lemma 2. ∎

The main result of the analysis of the parallel game in section 3 was that, even if players do not interact under personal relationships very often, the fraction of trustworthy players develops at least to an interior solution. Being unaware about the competitive nature of the imperfect information setting, or with trustworthy players having a tendency to trust, this stability becomes dependent on λ and the distribution of types. The following bifurcation diagrams illustrate the situation. In the left hand diagram, $s = .33$ and $r = .66$ so $s + r = 1$. In the right hand diagram $s = .5$ and $r = .75$ so $s + r > 1$. In order to focus on the critical points, the horizontal λ-axis has been restricted to the interval $[.25, .75]$. All phase lines to the left or right of the chosen boundaries are identical to the respective outer ones drawn in the diagrams.

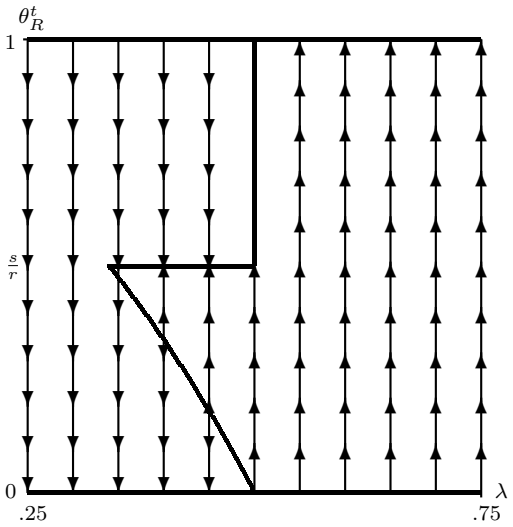

Figure 6: Bifurcation diagram, $s = .33, r = .66$

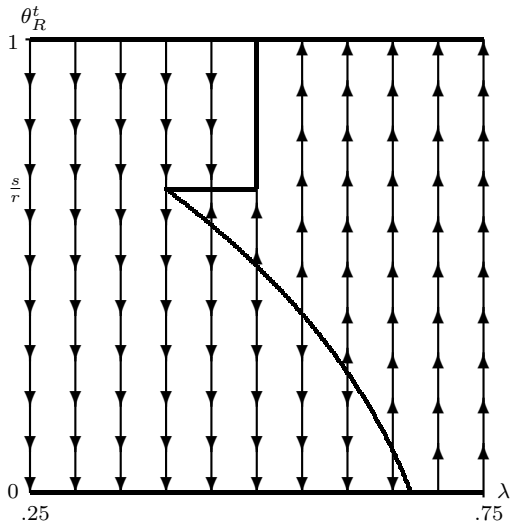

Figure 7: Bifurcation diagram, $s = .5, r = .75$

As the illustrations highlight, a tendency to trust is quite dangerous for the trustworthy types. In sharp contrast to the results of the previous sections, trust and trustworthiness may only prevail if interaction under personal relationships is very frequent and, if it not so frequent, if the initial share of trustworthy player is high.

6 Conclusion

The model presented in this paper is, in a way, a catch-all model with respect to the question whether trust and trustworthiness can survive in a competitive world. Under the label of personal relationships, all aspects of e.g. repeated interaction have been summarized with no further differentiation regarding for example interaction under different group sizes or strategic behavior as the application of trigger strategies. Having made this abstraction and under the assumption that personal relationships allow to discriminate between different player types, the question whether trust and trustworthiness can prevail was analyzed in a world with an infinitely large population and random interaction. The crucial parameter in this model is the frequency with which players interact under personal relationships. In the basic model, it was found that trust and trustworthiness do not diminish. Rather, if the frequency of interaction under personal relationships is high, then a monomorphic population with all players trusting and rewarding evolves. Otherwise, there is a stable interior solution in which trusting and non-trusting, rewarding and exploitive behavior exist. It was shown that the results are relatively stable against trembles for a wide variety of payoff structures. Only if there is little gain from rewarded trust in comparison to showing no trust (small distance between s and r), the trustworthy types will vanish. In contrast to that, trustworthy types with a tendency to trust highly endanger their own survival chances.

Of course, the crucial question is how high the real frequency of interaction under personal relationships is. One may argue that most of our interaction takes place with people we have some experience with, which would induce a high frequency. However, an important aspect in the model is that all decisions are taken to be identical, e.g. all contracts a company signs are of equal worth or all secrets we want to share are equally important to us. This is surely not realistic. A company may have signed multiple contracts of minor importance with little legal underpinning, i.e. by simply relying on their partners. But this need necessarily induce the company to trust the same partners if higher costs or risks are involved. In these cases it may instead rely on written contracts. The probability the firm expects trust to be rewarded will be reflected in the length of the contract. In such a case, the personal relationship exists, but it is irrelevant in the particular trust situation. Hence, the frequency of interaction under personal relationships will be lower and the

interior solution might emerge. After all, an overall success of trust and trustworthiness, as well as their sure decline, seem relatively unreasonable. Additionally, the interior solution allows for all behaviors to be present in the population while still being evolutionary stable. Dealing with issues of human behavior rather than the evolution of genotypes, the interior solution appears to be close to "reality".

Finally, the case of trustworthy players with a tendency to trust yields an interesting insight. One may read a tendency to trust as getting too confident about the presence of trustworthiness. Such an overconfidence is then exploited. But more frequent exploitation may induce people to be more aware. If the unawareness of the trustworthy players is removed, then their share will develop back toward the interior solution. In total, unawareness based on overconfidence could expand the interior solution to a set which is characterized by cycling periods of more and less trustful behavior.

References

Axelrod, R. M. (1984). *The Evolution of Cooperation*. Basic Books.

Bolle, F. (1998). Rewarding Trust: An Experimental Study. *Theory and Decision 45*, pp. 83–98.

Bolton, G. E., E. Katok, and A. Ockenfels (2004). Trust Among Internet Traders: A Behavioral Economics Approach. *Analyse und Kritik 26*, pp. 185–202.

Bolton, G. E. and A. Ockenfels (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review 90*(1), pp. 167–193.

Falk, A. and U. Fischbacher (2006). A Theory of Reciprocity. *Games and Economic Behavior 54*(2), pp. 293–315.

Fehr, E. and U. Fischbacher (2005). The Economics of Strong Reciprocity. In H. Gintis, S. Bowles, R. Boyd, and E. Fehr (Eds.), *Moral Sentiments and Material Interest*. MIT Press.

Fehr, E. and S. Gächter (1998). Reciprocity and Economics: The Economic Implications of Homo Reciprocans. *European Economic Review 42*, pp. 845–859.

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

Gigerenzer, G. and R. Selten (2002). Rethinking Rationality? In G. Gigerenzer and R. Selten (Eds.), *Bounded Rationality - The adaptive toolbox*. MIT Press.

Güth, W. and H. Kliemt (1994). Competition or Co-operation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes. *Metroeconomica 45*(2), pp. 155–187.

Güth, W. and H. Kliemt (1998). The Indirect Evolutionary Approach: Bridging the Gap between Rationality and Adaptation. *Rationality and Society 10*, pp. 377–399.

Güth, W., H. Kliemt, and S. Napel (2006). Population-dependent Cost of Detecting Trustworthiness - An Indirect Evolutionary Analysis. *Max-Planck Working Papers on Strategic Interaction 08-2006*.

Güth, W., F. Mengel, and A. Ockenfels (2007). An Evolutionary Analysis of Buyer Insurance and Seller Reputation in Online Markets. *Theory and Decision 63*(3), pp. 265–282.

Güth, W. and S. Napel (2006). Inequality in a Variety of Games - An Indirect Evolutionary Analysis. *Economic Journal 116*, pp. 1037–1056.

Harsanyi, J. C. (1967). Games with Incomplete Information Played by "Bayesian" Players, i. *Management Science 14*(3), pp. 159–182.

Hirsch, M. W., S. Smale, and R. L. Devaney (2004). *Differential Equations, Dynamical Systems and an Introduction to Chaos* (2 ed.). Elsevier Academic Press.

Huck, S. and J. Oechssler (1999). The Indirect Evolutionary Approach to Explaining Fair Allocations. *Games and Economic Behavior 28*(1), pp. 13–24.

Selten, R. (1975). Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games. *International Journal of Game Theory 4*(1), pp. 25–55.

Weibull, J. W. (1996). *Evolutionary Game Theory*. MIT Press.