The Psychological Game of Trust

Martin Dufwenberg[¤] and Werner Güth[y]

March 2004

Abstract

Two traditional assumptions in neo-classical economics have been material self-interest and (commonly known) decision rationality. Since there is ample contradictory empirical evidence, many recent attempts have been made to remodel the situation so that rational behavior is more in line with actual results (game …tting). Here we concentrate on intrinsic let-down aversion whose strength can depend on the relative frequency of such concerns, i.e. on a sociological aspect, and examine how these ideas apply to a game of trust. We discuss whether the ‡exibility of the approach is a virtue or a vice.

[¤]Department of Economics, University of Arizona, Tucson, AZ 85721-0108, USA; email: martind@eller.arizona.edu

[y]Max Planck Institute for Research into Economic Systems, Strategic Interaction Group, Kahlaische Str. 10, 07745 Jena, Germany; email: gueth@mpiew-jena.mpg.de

1 Introduction

Instead of discussing in abstract terms consider the basic game of trust (Figure 1) whose intuitive interpretation will be provided below.

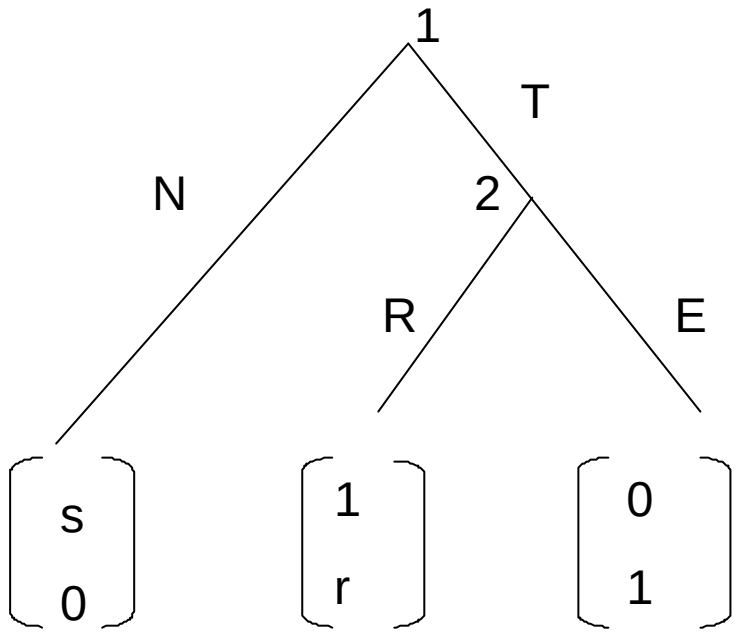

Figure 1: The interpretation of moves is T(rust), N(o trust), R(eward) and E(xploit) (0 $<$ r, s $<$ 1)

All payo¤s are material, e.g. monetary earnings. We assume $0 < r, s < 1$. If player 2 is rational and only interested in his own material payo¤, he will choose E. If player 1 is aware of this and is similarly rational in pursuing his own material interests, he will rely on N. The unambiguous solution is therefore the strategy vector (N, E) although this outcome is payo¤ dominated by (T, R).

If such a game is experimentally explored, it is rather likely that many player pairs rely on (T, R) rather than (N, E). A survey of experimental results supporting such a conjecture is Roth (1995). Actual trust experiments on this

kind of game are Güth, Ockenfels, and Wendel (1997) and Snijders (1996); two experiments on related games are Berg, Dickhaut, and McCabe (1995) and Dufwenberg and Gneezy (2000). A common reaction to such findings, to which we refer as game fitting[1], is

- not to question (commonly known) decision rationality,
- but to remodel the situation, e.g. by including non-material incentives (an early example is Bolton (1991); see Fehr and Schmidt (2001) and Sobel (2001) for surveys of more recent work) or other changes of the rules.

Here we will first discuss intrinsic motivation and whether one can expect such motivation to survive in the long run. Our main focus is, however, the quite sophisticated (although possibly inappropriately named) concept of psychological game theory (Geanakoplos, Pearce, and Stacchetti, 1989). We focus on how this toolbox can be used to capture players' aversion to letting others down, which we will call "let-down aversion".[2]

In our analysis, all games are embedded in a population context whose evolution over time is analyzed. This will allow us to link let-down aversion to sociological aspects, e.g. how frequent such feelings are in the population. The interrelation of sociological and psychological aspects allows for a variety of psychological games which are more or less reasonable. As we will see, the partly extremely opposite specifications imply very different results for the evolutionarily stable population share of people that are affected by let-down aversion.

In the light of such exercises it is natural to discuss how one should react to empirical, mostly experimental findings contradicting rational material

[1] Similar to the notion of curve fitting in statistical regression analysis.

[2] The term "let-down aversion" was introduced by Dufwenberg and Gneezy (2000). Related models appear in Huang and Wu (1994), Dufwenberg (2002), Bacharach, Guerra and Zizzo (2001), Charness and Dufwenberg (2003).

self-interest, i.e. materially opportunistic behavior. Should one maintain the assumption of (commonly known) decision rationality and adjust the "rules of the game" only, e.g. by applying psychological game theory and importing sociological aspects? Or should one give up the rational choice-approach and rely on (social and cognitive) psychology and subscribe to bounded rationality? We will discuss these methodological issues in some depth.

In section 2, we review earlier work which can be compared to our analysis of let-down aversion in section 3 and 4. Section 5 concludes.

2 Intrinsic let-down aversion

Intrinsic motivation refers to positive or negative evaluations of certain types of behavior (see Frey, 1997, for a general discussion). Here it will be distinguished from other secondary concerns by assuming that intrinsic motivation

(i) depends on no other consideration than the choice itself, and

(ii) does not depend on how prevalent such preferences are in a society.

In the next section, we will give up assumption (i) by making the evaluation of behavior 'belief-dependent' using psychological game theory to model let-down aversion. Thereafter also assumption (ii) will be given up by linking sensitivity to let-down aversion to sociological aspects. In this section our basic paradigm, previously analyzed by Güth and Kliemt (1993, 1994, 2000), will be introduced such that both assumptions, (i) and (ii), are met.

The basic game of trust, already introduced above, is the (sub)game after the choice by chance which occurs with probability $1 - p$ in Figure 2 where we assume, for the moment, perfect information (all information sets are singletons).

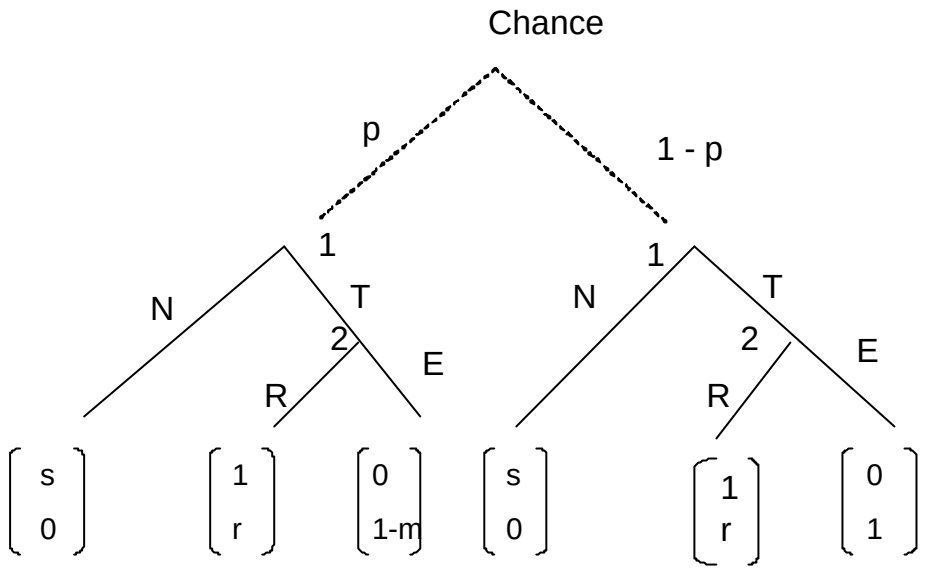

Figure 2: $(0 \cdot p \cdot 1; m > 1 \text{ ¡ } r; 0 < r, s < 1)$

The subgame after the alternative chance move with probability p di¤ers from the one, just described, only by the payo¤ parameter m of player 2 in case of the choices T and E. Whereas all other payo¤ components are representing material success like monetary earnings, parameter m stands for a purely non-material intrinsic evaluation. More speci…cally, a positive m is interpreted as the guilt caused by the choice of E. Clearly, $m > 1 \text{ ¡ } r$ renders such guilt prohibitive, i.e. would induce the choice of R by player 2 whereas any $m < 1 \text{ ¡ } r$, e.g. m = 0, does not prevent the dilemma-type outcome.

What the initial chance move in Figure 2 captures is a potentially (in case of $0 < p < 1$) bimorphic population of players 2 with a population share p ($1 \text{ ¡ } p$ not) a¤ected by (non)prohibitive let-down aversion. In the tradition of indirect evolutionary analysis (see Güth and Yaari (1992), Güth (1995)) we derive the evolutionarily stable population composition $p^{¤}$ 2 $[0, 1]$ after solving the one parameter (p) class of games for each value of p.

Given p, clearly, the solution[3] is

[3] Resulting frome once repeated elimination of (weakly) dominated strategies or from the concept of (subgame) perfect equilibria.

$$s^{\alpha} = (s_1^{\alpha}, s_2^{\alpha}) = \begin{cases} (N, E) \text{ for } m = 0 \\ (T, R) \text{ for } m > 1 \text{ ¡ } r \end{cases}$$

Assuming that evolutionary success is purely materially determined[4], an m-type of player 2 with $m > 1$ ¡ r earns r whereas the m = 0-type of player 2 only gets 0. Thus sooner or later the population share p of the m-types with $m > 1$ ¡ r should converge to 1, i.e. only the monomorphic population with p^{α} = 1 is evolutionarily stable (see Güth and Kliemt, 1993). The crucial assumption of this derivation is, of course, that player 1 can perfectly discriminate between the $m > 1$ ¡ r ¡ player 2 and the m = 0-type of player 2 (see Güth and Kliemt, 2000, for a much more general analysis which allows 1 to receive limited information about 2's type).

3 "Psychological" let-down aversion

Let us now substitute intrinsic guilt by belief-dependent feelings, as a means to model let-down aversion. We draw on the toolbox of "psychological game theory" introduced by Geanakoplos, Pearce, and Stacchetti (1989). The basic idea is that a decision maker does not like to disappoint others.

Regarding the game form in Figure 3, let ρ with $0 \cdot \rho \cdot 1$ denote player 2's probability of using his move R after the chance move with probability p. Denoting by E_i f¢g player i's expectation operator, player 1's expectation of ρ is $\rho^{0} = E_1$ fρg and player 2's expectation of ρ^{0} the probability $\rho^{"} = E_2$ fρ^{0}g $= E_2$ fE_1 fρgg. The substitution of intrinsic guilt $m(> 1$ ¡ $r)$ in Figure 2 is simply done by substituting ρ" for m in Figure 2 to get Figure 3. In equilibrium both expectations must be consistent with actual behavior, i.e. $\rho = \rho^{0} = \rho^{"}$. For the game in Figure 3, we proceed as in the section above for the game in Figure 2.

[4] Intrinsic motivation (here guilt) can thus in‡uence success only indirectly by its e¤ects on choice behavior.

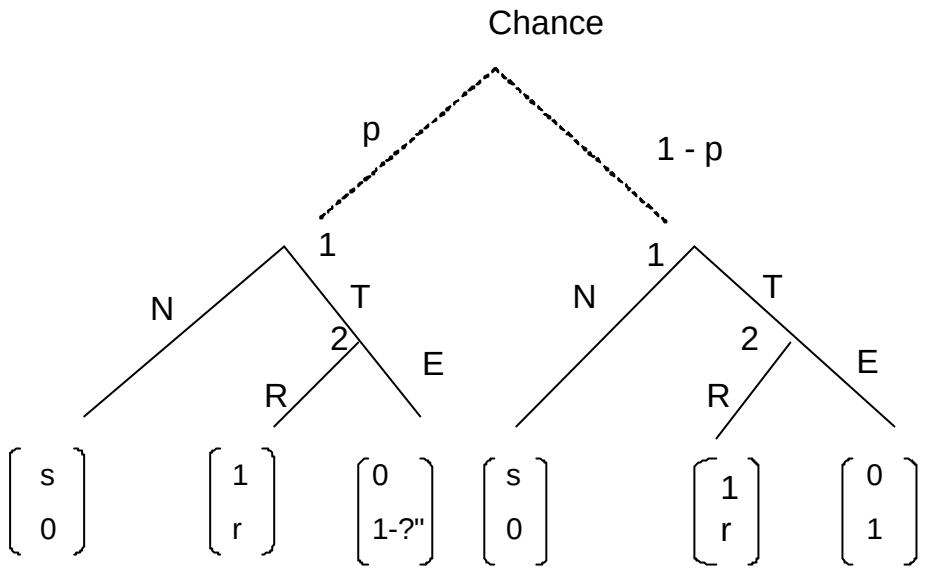

Figure 3: ($\rho^{"}$ = E_2 fE_1 fρgg with 0 · ρ · 1)

Let us refer to player 2 with payo¤ 1-$\rho^{"}$after T and E as the p-type of player 2. We …rst consider the case of complete type information, discussed in section 2 when studying the evolution of m, when player 1 knows whether the p-type has been chosen or not when deciding between N and T. As we will see, there may be multiple equilibria.[5]

The p-type prefers R over E if $r > 1$ ¡ $\rho^{"}$. An equilibrium where this holds implies ρ = 1 and via consistency of beliefs also $\rho^{"}$ = 1. The condition then obviously holds. The game in Figure 3 thus has the equilibrium solution $s^{¤}$ = (T, R) in case of the p-type and $s^{¤}$ = (N, E) otherwise.

Assume now on the contrary that the p-type of player 2 prefers E over R what requires 1 ¡ $\rho^{"} > r$. Now such an equilibrium implies ρ = 0 and via consistency also $\rho^{"}$ = 0 what shows that the game in Figure 3 has another solution, namely $s^{¤}$ = (N, E) for both types.

In case of incomplete information when player 1 only knows the a priori probability p of the p-type but not which type he actually confronts (formally:

[5] The solution concept we apply is Geanakoplos et al's notion of "psychological subgame perfect equilibrium". We will focus on the pure strategy equilibria, although there are also mixed equilibria (in which 2 is indi¤erent between R and E).

both decision nodes of player 1 are in the same information set), the true posterior beliefs $\rho'' = \rho^0 = \rho$ are

$$\rho'' = \rho^0 = \rho = p \cdot 1 + (1 - p) \cdot 0 = p$$

in case of $r > 1 - \rho''$.[6] The condition for prohibitive let-down aversion of the p-type is thus $r > 1 - p$. Hence, there exists an equilibrium where 1 chooses T and only the p-type of 2 chooses R, which we write as $(T, (R/p - \text{type}, E/1 - p - \text{type}))$ or simply $(T, (R, E))$, if the two conditions

$$r > 1 - p \text{ and } p > s$$

hold. This is equivalent to

$$p > \max\{1 - r, s\}.$$

hold. Of course, also the other equilibrium exists since the choice of $\rho = 0$ by the p-type of player 2 leads to the consistent expectation $\rho = 1 \cdot 0 + (1 - p) \cdot 0 = 0 = \rho^0 = \rho''$ which implies the equilibrium $(N, (E/p - \text{type}, E/1 - p - \text{type}))$ or $(N, (E, E))$. One can try to deal with this troublesome ambiguity of the solution of Figure 3 by applying some concept of equilibrium selection, e.g. the theory of Harsanyi and Selten (1988): Clearly, if types are commonly known (complete information), the equilibrium $s^* = (T, (R, E))$ payoff dominates $s^* = (N, (E, E))$ for $p > \max\{1 - r, s\}$ regardless whether one relies on material success only ($p > s, r > 0, 1 > 0$) or on (phenotypic) utility ($p > s, r > 0, 1 > 0$) due to $\rho'' = 0$ where the three inequalities refer to the three players involved (player 1 and the two types of player 2).

When checking risk dominance based on psychological game theory note first of all that the $1 - p$-type of player 2 chooses E in both equilibria, $(T, (R, E))$ and $(N, (E, E))$. In the comparison game (see Harsanyi and Selten, 1988) the only active players are thus player 1 and the $p - 1$ type of player 2. This comparison game is presented in normal form by Table 1 where s_2 stands only for the p-type's choice (the $1 - p$-type chooses E in all four cases of

[6] Our formulation here presumes that what matters to 2's motivation is 2's expectation of 1's expectation that 2 will chooses T, not 2's expectation of 1's expectation that 2 will choose T conditional on 2 being a p-type.

Table 1). One encounters a di¢culty, namely that a certain payo¤ (here of the p-type player 2 after T and E) is not uniquely de…ned (see Table 1) if one relies on utility rather than material success. For the case at hand the di¢culty is, however, less serious since (N, E) in Table 1 is non-strict since the p-type of player 2 does not su¤er from unilaterally deviating from (N, E) in Table 1. In the comparison game (T, R) is thus the only solution candidate so that $(T, (R, E))$ risk dominates $(N, (E, E))$.

s_1/s_2	R	E
N	$s, 0$	$s, 0$
T	$1, r$	$\frac{1}{2}$ 0, 1 ¡ p in view of (T, R) 1 in view of (N, E)

Table 1 (the comparison game for the p-type of player 2 only)

As in section 2, let success be measured purely by material payo¤s and assume again complete type information. The same arguments as in section 2 then imply that $p^{¤}$ = 1 is the only evolutionarily stable population composition. In case of private type information, the 1-p-type earns materially more than the p-type of player 2 if $(T, (R, E))$ is played. Thus if the initial population composition p_0 satis…es $p_0 >$ max f1 ¡ r, sg, the population share p will decline and …nally fall below max f1 ¡ r, sg where the solution becomes $(N, (E, E))$. Unless when refering to trembles (see Güth and Kliemt, 2000) the evolutionary drive then vanishes.

4 Importing sociology into psychological game theory

What we want to capture here is how "let-down aversion", as modelled in Figure 3, can depend on the relative frequency of the p-type player 2, as measured by its probability p. In Figure 4 the psychological e¤ect of Figure 3 is weighted by $f(p)$ so that p can in‡uence how strongly "psychological

guilt" is felt. For the sake of speci…city, we distinguish only two extreme speci…cations of $f(p)$, namely

(a) $f(p) = p$ for all $p \in [0,1]$ and

(b) $f(p) = 1 - p$ for all $p \in [0,1]$.

Case (a) can be justi…ed by the assumption that let-down aversion should be felt less strongly when such feelings are less common. Although case (a) appears formally similar to the case of incomplete information, its justi…cation is different. How strongly aversion is felt depends on how frequent it is even if the aversion-type of player 2 would be commonly known.

The opposite speci…cation (b) could also be seen as reasonable: Here "let-down aversion" is felt strongest when such feelings are extremely rare. If there is just one "Mother Theresa", wouldn't she feel most badly when knowing that even she does not care? A rare moralist might feel more strongly than when the moral behavior under consideration is usual. In the following, we will proceed for the two variants (a) and (b) of Figure 4 as for the previously analyzed games where we restrict attention to the case of complete type-information.

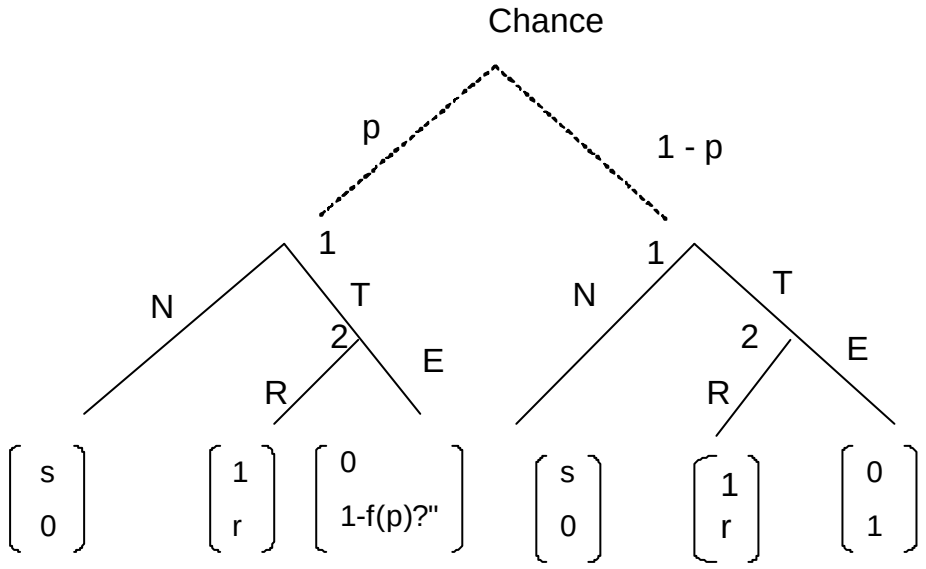

Figure 4: Frequency dependent let-down aversion

Case (a): The p-type of player 2 prefers R over E if $r > 1 - p\rho^{"}$. Via consistency ρ = 1 this implies $r > 1 - p$ or $p > 1 - r$. The reversed preference $1 - p\rho^{"} > r$ implies ρ = 0 and via consistency $1 > r$. Thus in case of the p-type player 2, the solution is (N, E) for all $p < 1 - r$ whereas for $p > 1 - r$ there exist two equilibria, namely (T, R) and (N, E), for this subgame.

As before, there is reason, namely payo¤ dominance, to 'select' (T, R) as the unambiguous solution if it coexists with (N, E). Thus, p will increase in the range $p > 1 - r$. In the range $p < 1 - r$ the solution (behavior) is the same, regardless whether the p-type or the opportunistic (in the sense of no let-down aversion) type of player 2 is present. Thus all $p < 1 - r$ are weakly (in the sense of no drift towards as well as away from such rest points) stable (see Figure 5) whereas only $p^{¤}$ = 1 is strictly stable.

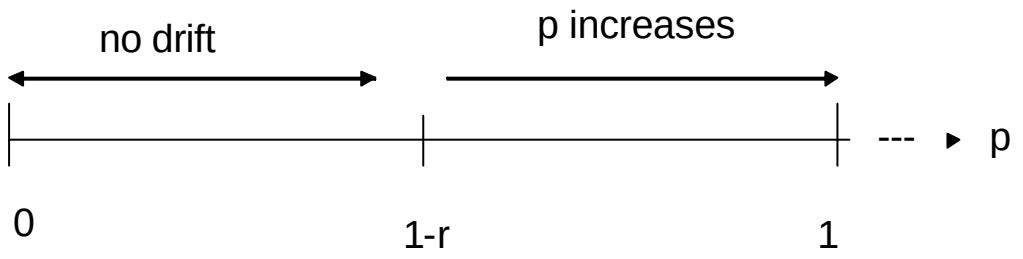

Figure 5

Case (b): The p-type of player 2 prefers R over E if $r > 1 - (1 - p)\rho^{"}$ what together with the consistency requirement $\rho^{"} = \rho^{0} = \rho$ = 1 implies $r > p$. The reversed preference, similarly, implies $r < 1$. Thus for $p < r$ there coexist two equilibria, namely (T, R) and (N, E), for this subgame whereas the only equilibrium for $p > r$ is (N, E). Since for the p-type of player 2 and complete information (T, R) is selected in the range $p < r$, the population share p increases in that range. The weakly evolutionarily stable population compositions p ¸ r are illustrated by Figure 6. There exists no strictly stable population composition p, but a continuum $(p$ ¸ $r)$ of weakly stable ones.

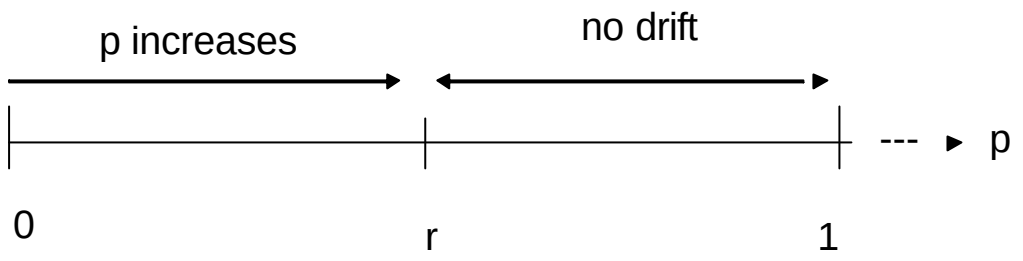

Figure 6

What the two extreme speci…cations for importing sociological ideas into psychological game theory illustrate is

- that quite opposite e¤ects can be viewed as intuitive and
- that the evolutionarily stable population composition depends crucially on which of the more or less intuitive speci…cations one applies.

Such ‡exibility of sociologically enriched, psychological games can be a blessing since it allows for more equilibria, for other solutions, as well as for other stable population compositions. The curse, however, may be that the concept as such is rather uninformative. In order to get clear predictions, one must obtain some rather detailed understanding regarding how sociological and psychological e¤ects interact. This may appear as di¢cult as hypothesizing directly about the likely behavior in trust games. No, or little, structural information is gained a priori by just translating all motivational and emotional aspects into the terminology of psychological game theory and applying more or less sophisticated concepts of commonly known rationality. However, careful experimental tests, involving the elicitation of the decision makers' beliefs and beliefs about beliefs, can help to shed light on which model formulation is best. Dufwenberg and Gneezy (2000), Bacharach, Guerra and Zizzo (2002), and Charness and Dufwenberg (2003) take some steps in this direction.

Of course, psychological game theory imposes some discipline when analyzing the interaction of sociological and psychological concerns. But this does not come cheaply. Most applications impose rather serious restrictions like

- that all players know exactly others' expectations and
- that all these expectations coincide with actual behavior.

Moreover, most applications of psychological game theory assume that idiosyncratic concerns, like sensitivity to let-down aversion, are commonly known. This sort of assumption is not new in economic theory. For example, idiosyncratic degrees of risk aversion are often assumed to be commonly known in spite of the unrealism of such common observability. Allowing for incomplete information about such idiosyncratic concerns by imposing some Bayesian setting (of prior beliefs over all possibilities) like in section 3 above does not help much although specifying prior beliefs by the true population shares (Güth, 1995) somewhat limits the arbitrariness of such exercises.

Future research on psychological game theory may do well to consider relaxing the assumptions that behavior, beliefs, and various parameters re‡ecting psychological propensities are common knowledge.

5 Final remarks

Basically, we have tried

- to illustrate with the help of the trust game how psychological game theory can account for psychological and sociological aspects and how one can thereby discuss the evolution of psychological motives, here of let-down aversion, and
- to demonstrate that psychological game theory allows for quite a variety of game …tting exercises which can be seen as being more or less intuitively convincing.

Without strong evidence of what motivates decision makers in addition to material incentives hardly anything speci…c can be concluded. Collecting such empirical evidence on "psychological motives", as suggested by psychological game theory, may seem di¢cult. However, it may be possible to make progress in this direction by creating experimental designs that allow beliefs, about choices and beliefs, to be elicited.

It should be noted that we have restricted ourselves rather seriously when exploring the possibilities of psychological game theory for the basic trust game. Denote by τ player 1's probability for using T rather than N and let $\tau^{0} = E_2 f\tau g$ and $\tau'' = E_1 f\tau^{0} g = E_1 f E_2 f\tau g g$ be the 1st, respectively 2nd order expectation concerning τ. The game in Figure 7 illustrates how also player 1 can su¤er from "psychological motives", e.g. from not wanting to let down a rewarding player 2, and how this (in case of $\rho(¢)$ being non-constant) can also be linked to sociological aspects.

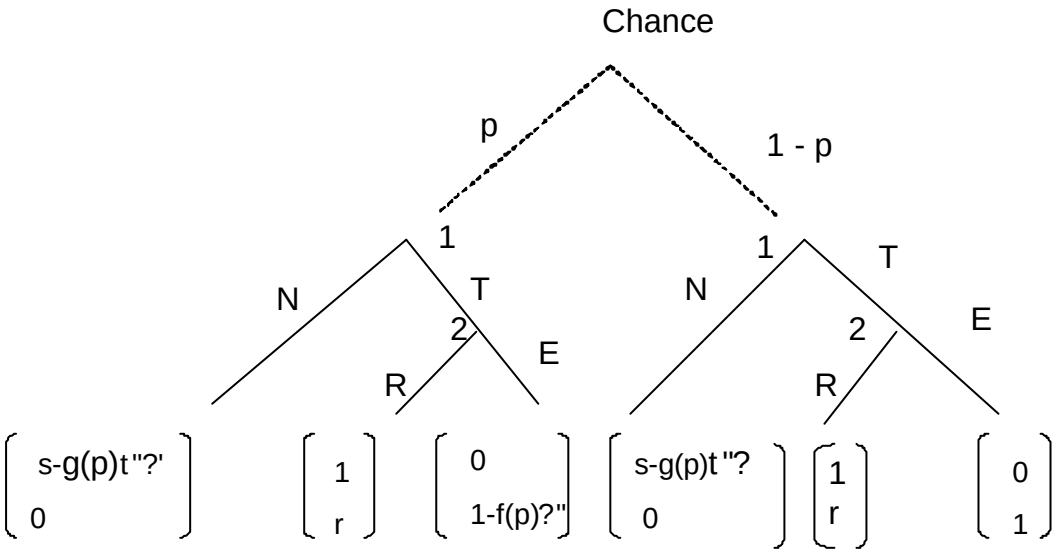

Figure 7: ($\tau'' = E_1 f\tau^{0} g = E_1 f E_2 f\tau g g$ with $\tau = \text{Prob} f T g$)

Let us just consider the speci…cation

$$f(p) = p \text{ and } g(p) = 1 - p,$$

i.e. player 1 su¤ers more when letting down a rare rather than a frequent p-type player 2 who would reward trust. This apparently does not question the solution (N, E) for the subgame (we rely again on complete type information) since this implies via consistency τ" = 0 = ρ^0 so that the additional e¤ect plays no role. Furthermore, $s_2^{\text{¤}} = R$ requires $p > 1$ ¡ r (see case (a) in section 4). Player 1 prefers T over N when recognizing a p-type if $1 > s$ ¡ $(1$ ¡ $p)\tau$"ρ^0. Consistency with (T, R) implies τ" = 1 = ρ^0 so that the condition becomes $1 > s$ ¡ $(1$ ¡ $p)$ which holds for all p 2 $[0, 1]$. The results of case (a) in section 4 are thus not questioned by the richer psychological motives of Figure 7 as compared to Figure 4.

The box of game …tting via "psychological games" may appear to be bottomless. Is it like Pandora's box? Our exercises illustrate how one may maintain common knowledge of rationality and gain ‡exibility, but possibly lose structural information and empirical content. Ultimately it makes sense to seek empirical guidance as to which modeling track is more reasonable. The development of behavioral economics using psychological game theory should therefore probably go hand in hand with careful experimentation. There is hope in the box if psychological motives are parsimoniously used only after having proved empirically their relevance.

References

[1] Bacharach, M., Guerra, G. and Zizzo, D. (2001): Is trust self-ful…lling? An experimental study. Mimeo.

[2] Berg, J., Dickhaut, J. and McCabe, K. (1995): Trust, reciprocity, and social history, Games and Economic Behavior 10, 122-142.

[3] Bolton, G. (1991): A comparative model of bargaining: theory and evidence. American Economic Review 81, 1096-1136.

[4] Charness, G. and Dufwenberg, M. (2003): Promises & partnership. Working Paper 03-07, Department of Economics, University of Arizona.

[5] Dufwenberg, M. (2002): Marital investment, time consistency, and emotions. Journal of Economic Behavior and Organization 48, 57-69.

[6] Dufwenberg, M. and Gneezy, U. (2000): Measuring beliefs in an experimental lost wallet game, Games and Economic Behavior 30, 163-82.

[7] Fehr, E. and Schmidt, K. (2001): Theories of fairness and reciprocity - Evidence and Economic Applications, forthcoming in: M. Dewatripont, L. Hansen and St. Turnovsky (Eds.), Advances in Economics and Econometrics - 8th World Congress, Econometric Society Monographs .

[8] Frey, S. B. (1997): Not Just For the Money. An Economic Theory of Personal Motivation. Edward Elgar Publishing Limited, Cheltenham, 156 pages. (Paperback edition reprinted in 2000.).

[9] Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989): Psychological games and sequential rationality. Games and Economic Behavior 8(1), 56-90.

[10] Güth, W. (1995): An evolutionary approach to explaining cooperative behavior by reciprocal incentives. International Journal of Game Theory 24, 323-344.

[11] Güth, W. and Kliemt, H. (1993): Menschliche Kooperation basierend auf Vorleistungen und Vertrauen - Eine evolutionstheoretische Betrachtung. Jahrbuch für neuere politische Ökonomie 12, 252-277.

[12] Güth, W. and Kliemt, H. (2000): Evolutionarily stable co-operative commitments. Theory and Decision 49, 197-221.

[13] Güth, W., Ockenfels, P. and Wendel, M. (1997): Cooperation based on trust. An experimental investigation. Journal of Economic Psychology 18, 15-43.

[14] Güth, W. and Yaari, M. (1992), "Explaining reciprocal behavior in simple strategic games: An evolutionary approach", Ch. 2 in In: U. Witt (Ed.) Explaining process and change: Approaches to evolutionary economics, University of Michigan Press, Ann Arbor, 23-34.

[15] Harsanyi, J. C. and Selten, R. (1988): A general theory of equilibrium selection in games, Cambridge Mass: M.I.T. Press.

[16] Huang, P. and Wu, H.-M. (1994): More order without more law - A theory of social norms and organizational cultures, Journal of Law, Economics and Organization 10, 390-406.

[17] Roth (1995b): Bargaining experiments. Handbook of Experimental Economics (eds. J.H. Kagel and A.E. Roth), Princeton, NJ: Princeton University Press, 253-348.

[18] Sobel, J. (1999): Interdependent preferences and reciprocity. Mimeo.

[19] Snijders, Ch. (1996): Trust and commitments, Utrecht University: dissertation.