A Theory of Sequential Reciprocity

Dufwenberg, M.; Kirchsteiger, G.

Publication date:
1998

A THEORY OF SEQUENTIAL RECIPROCITY*

Martin Dufwenberg** & Georg Kirchsteiger***

March 1998

Abstract: Many experimental studies indicate that people are motivated by reciprocity. Rabin (1993) develops techniques for incorporating such concerns into game theory and economics. His model, however, does not fare well when applied to situations with an interesting dynamic structure (like many experimental games), because it is developed for normal form games in which information about the sequential structure of a strategic situation is suppressed.

In this paper we develop a theory of reciprocity for extensive games in which the sequential structure of a strategic situation is made explicit. We propose a new solution concept—sequential reciprocity equilibrium—which is applicable to extensive games, and we prove a general equilibrium existence result. The model is applied in several examples, including some well known experimental games like the Ultimatum game and the Sequential Prisoners' Dilemma.

JEL codes: A13, C70, D63

Keywords: Reciprocity, extensive games

* We are grateful to seminar participants in Stockholm and Zürich for helpful comments.

** Department of Economics, Uppsala University; martin.dufwenberg@nek.uu.se

*** CentER, Tilburg University; g.kirchsteiger@kub.nl

1.INTRODUCTION

"Reciprocity is the key to every relationship"

(Danny DeVito in LA Confidential)

Almost all of economic theory is built on the assumption that people act selfishly and do not care about the well-being of other human beings. Lots of recent evidence, however, contradicts pure selfishness. For example, Kahneman, Knetsch & Thaler (1986) show in a seminal paper that consumers' opinions about price increases depend crucially on the costs of the firm, but not on the market conditions—a price increase due to cost increases is regarded as justified, while a demand shock is not a valid justification. Whereas Kahneman et al's study deals with the *fairness perception* of consumers, experimental evidence suggest that also *actual behavior* is shaped by factors inconsistent with pure selfishness. For example, in ultimatum bargaining experiments people often reject allocations in which they receive a much smaller monetary payoff then their partners in favor of an allocation where neither player receives anything (see Roth (1995) for an overview). In gift exchange games, where two persons in turn determine how large gifts to give to one another, people behave reciprocally. A large gift by the first mover prompts a generous response (see Berg, Dickhaut & McCabe 1995, Falk, Gächter & Kovacs 1997 and Fehr, Gächter & Kirchsteiger 1997). If the size of the gift of the first player is determined on a double auction market, these gift exchange forces are even strong enough to prevent the market from clearing (see Fehr, Kirchsteiger & Riedl 1993, 1998).

These deviations from selfishness may have important economic consequences. As Fehr et al. (1997) show experimentally the set of enforceable contracts increases considerably due to non-selfish behavior. These effects are of particular importance for understanding labor markets. In a series of theoretical papers Akerlof (1982) and Akerlof & Yellen (1988, 1990) show that fairness is a possible explanation why wages may be above the market clearing level so that involuntary unemployment occurs. Fehr & Kirchsteiger (1994) use this approach to explain why two-tier systems are rarely observed in reality. Bewley (1995) finds strong empirical evidence for the validity of these theories. When asked for the reason why wages remain above the market clearing level in recessions, managers and other labor market

participants say that wage declines may destroy "working morale"—workers would decrease their working effort after a decline in wages which therefore cannot be enforced.

All this evidence suggests that people are not motivated solely by material self-interest. Also considerations of altruism, emotions, fairness, et cetera play a role. Models designed to capture some of these phenomena can be roughly divided into two classes: those that focus on distributional concerns, and those that focus on a concern for reciprocity. The distributional approach permits decision makers to be motivated not only by their own material gain, but rather by the final distribution of the material payoff.[8] In Fehr & Schmidt (1997), for example, it is assumed that for a given own material payoff a person's utility is decreasing in the difference between the own payoff and that of the partner. They show that a lot of experimental evidence can be explained by their theory which, furthermore, has the advantage of being very close to standard models. There is, however, a certain cost. The assumption that individuals care only about final distributions implies that they must be indifferent concerning *how* distributions come about. This is problematic if in fact individuals regard information about their co-players' specific choices or intentions as important to their decision making.[9]

Rabin (1993) convincingly argues that intentions play a crucial role when individuals are motivated by reciprocity considerations. When a person wants to be kind to someone who was kind to her, and unkind to unkind persons, she has to assess the kindness (or unkindness) of her own action as well as that of others.[10] To do this she may have to look at the intentions that accompany an action. Take as an example the game Γ_1 in Figure 1 (with monetary payoffs).

---

[8] Examples of this approach are the models of Akerlof & Yellen (1980, 1990), Bolton & Ockenfelsfels (1997), Fehr & Kirchsteiger (1994), Fehr et al (1998), Fehr & Schmidt (1997), Kirchsteiger (1994), and Levine (1997), where in addition to distributional concerns persons are also motivated by the degree of altruism of the partner.

[9] A related problem discussed in social choice theory concerns whether welfare assessments can be made with reference to final distributions only. See Sen (1979) for a critical discussion.

[10] A word of caution about terminology is in order. Some authors (for example Bolton & Ockenfels 1997) distinguish between direct and indirect reciprocity, the former being a principle like the one we describe here (and simply call "reciprocity") , whereas the latter is a pure concern for distributive justice.

(Insert Figure 1)

Is *F* an unkind action? Clearly, this depends on what player 1 believes that player 2 will do. Suppose player 1 believes that player 2 will choose *d*. By choosing *F* player 1 then intends to give a payoff of 2 to player 2, whereas player 2 would get a payoff of only 1 if player 1 chose *D*. Hence, one may conclude that player 1 acts kindly if he chooses *F*. By an analogous argument, however, one must conclude that 1 is unkind if he chooses *F* while believing that 2 will choose *f*. This example shows not only that intentions are crucial in order to model reciprocity; it also makes clear that intentions depend on the *beliefs* of the players. Furthermore, the kindness of a player also depends on the *possibilities* he has. Change the game of Figure 1 such that player 1's strategy set consists only of *F*—she has to "choose" *F*. In such a game a "choice" of *F* is of course neither kind nor unkind—it is simply the only thing that 1 can do. Hence, in order to model the impact of intentions one has to take into explicit account both the possibilities and the beliefs of the players.

This is what Rabin (1993) does. Using the framework of Psychological Game Theory (Geanakoplos, Pearce & Stacchetti, 1989), he assumes that the players in two-player normal form games experience psychological payoffs in addition to the underlying material payoffs. The former payoffs depend on the players' kindness. Given the belief of player *i* about the strategy choice of the other player *j*, *i* regards himself as kind to the extent that he gives the other player a high payoff. In a similar way the kindness *i* expects from *j* depends on *i*'s belief about *j*'s strategy and on *i*'s belief about what *j* believes about *i*'s strategy choice, i.e. on a second-order belief of *i*. Given these kindness assessments, Rabin models the psychological payoff such that *i* wants to be kind to *j* if he believes *j* to be kind (as long as the material payoff does not become too important for *i*). Notice that the kindness of *i* depends on his belief about *j*'s strategy. Hence, *i*'s kindness depends on the payoff he intends to "give" to *j*, compared to the payoffs he is able give him—intentions and possibilities define the kindness of action. Therefore, the approach can model reciprocity. Furthermore, Rabin shows that a redefiniton of the payoff functions does in general not lead to the same results as his concept—since intentions matter, models of reciprocal behavior have to lead to different results than an approach where beliefs are not allowed to affect payoffs directly.

However, Rabin's model has a serious drawback that may restrict the potential for applications of his approach. Since it is a normal form concept it does not take into account the sequential structure of a strategic situation. Since a "Rabin-equilibrium" is calculated using the normal form, it may be that non-optimizing behavior is prescribed at information sets that are not reached (like in usual game theory where in Nash equilibrium players do not necessarily optimize off the equilibrium path). This lack of sequential calls for a modification of Rabin's concept.[11] It turns out that there is then another problem which makes this more complicated than in usual game theory. In sequential games players' may revise their beliefs as play unravels, and, since kindness depends on beliefs, the nature of reciprocity concerns may have to be revised accordingly. Consider the "Sequential Prisoners Dilemma" game as shown in Figure 2, which is is a stylized version of the experiments conducted by Fehr et al. (1993, 1998).

(Insert Figure 2)

It can be easily shown that cooperation by player 1 (the choice *C*) and unconditional cooperation by player 2 (i.e., the choice *c* at each node controlled by 2) form a "Rabin equilibrium" (defined in the normal form of Γ_2), as long as the the concern for material payoffs does not overcome the concern for reciprocity. [12] Unconditional cooperation of player 2, however, is very unplausible. Why should player 2 be kind after 1 took the unkind action?[13]

The problem arises because optimization is not mandated at 2's second node (which in the equilibrium (C,cc) is not reached). However, solving this problem is not just a matter of looking at the extensive game and mandating optimization at all nodes. After all, for 2 to

[11] Rabin (1993, p 1296) himself notes that "[e]xtending the model to sequential games is also essential for applied research".

[12] The problem we describe does not depend on the rather special kindness- and payoff-functions Rabin (1993) uses. The same problems arise with the more generalized kindness functions he discusses in his Appendix A.

[13] In the experiments by Fehr et al. (1993, 1998) such behavior was nearly never observed.

choose c at her rightmost node may be in her interest *if she conceives of 1 as kind*. However, it seems clear that even if 2 initially believes that 1 is kind, she should not maintain such a belief after 1 chooses D. Rather she should then regard 1 as unkind, which would motivate her to take revenge by choosing d.

The general upshot of this example is that a sensible model of reciprocity in sequential games must with care handle how beliefs change and how this affects reciprocity considerations. Incorporating such a "sequential reciprocity principle" is important in many potential applications which have a non-trivial dynamic structure. For example, the game shown by Figure 2, is a very stylized version of the Fair Wage Effort models of Akerlof and Yellen, with the firm (player 1) making a generous or greedy wage offer and the worker (player 2) deciding about providing a high or low working effort. Other examples are the ultimatum bargaining games and the gift exchange games discussed above. Given the sequential structure of these games and other potential applications, it is crucial to derive a concept of sequential reciprocity. This is the main objective of our paper.

In order to highlight and isolate the consequences of sequential reciprocity, and in order to facilitate a clearcut comparison with Rabin's (1993) model, we focus exclusively on incorporating a concern for reciprocity. We disregard distributional concerns. This is not to say that such concerns are unimportant. In reality both motives seem to play a role.[14] We hence regard the two approaches complementary.

In the next section we present our model and define the concept of a *sequential reciprocity equilibrium* (*SRE*). In Section 3 we apply this concept to some well known games and we show how the experimental results may be explained by our approach. In Section 4 we prove an existence theorem concerning the concept of SRE. In Section 5 our approach is compared in detail to Rabin's (1993) approach. Final comments conclude.

---

[14] The experimental evidence on the importance of reciprocity vis-à-vis the importance of distributional concerns is somewhat mixed. Whereas Bolton et al. (1996) find that only the final distribution matters, Charness' (1996) results suggest that reciprocity as well as distributional concerns play a role. Outside economics social psychologists have found strong experimental evidence of the importance of reciprocity, stressing the important role played by intentions (see e.g. Goranson and Berkowitz 1966 or Greenberg and Frisch 1972). Also anthroplogists and sociologists regard reciprocity as a main factor of human behavior, crucial for the functioning of human societies. For an overview of this literature, see Komter (1996).

2. THE MODEL

In the Introduction we argued that whether a person is kind or unkind depends not only on what he does but also on what he *believes* will be the consequence of his decision, as compared to what he believes would be the consequences of other decisions. Said differently, a person's kindness depends on his intentions. When another person wants to reciprocate kindness with kindness, she must assess the first persons intentions. Hence in taking decisions she will be motivated by her beliefs about the first persons intentions. Since intentions depend on beliefs, it follows that reciprocal motivation depends on *beliefs about beliefs.*

To come to grips with such issues, we work within the framework of *psychological game theory*. Psychological games differ from standard games in that a player's payoff depends not only on what strategy profile is played, but possibly also on what is the player's beliefs about other players' strategic choices or beliefs. The approach we use is inspired by Rabin (1993). We start off with a standard game, which is viewed as a description of a strategic situation which specifies only the material payoffs. We then derive a *psychological game* in which the payoff functions are redefined so as to reflect also reciprocity considerations. The main difference between our model and Rabin's is that he works with normal form representations of strategic situations, while we work with extensive forms and impose a requirement of sequential rationality.[15]

When this is done, a subtle issue arises: If a subgame is reached, perhaps unexpectedly, this may force a player to change his beliefs about the strategy profile being played. Since kindness relates to beliefs, assessments about kindness may therefore change and affect the ways in which a player is motivated by reciprocity concerns. It becomes necessary to somehow distinguish between a player's initial and subsequent beliefs. We handle this by keeping track of how the players' beliefs change as any new subgame is reached, and by assuming that whenever a player makes a choice he is motivated according to the beliefs he holds at that stage. These assumptions are central to our model. We argued already in

[15] There are also certain other less important differences between our model and Rabin's, but we postpone a discussion of these until Section 5.

connection to Γ_2 in the Introduction that if reciprocity is important one may get unreasonable conclusions unless players are assumed to update their assessments of how kind their co-player are as play unravels, and then reciprocate accordingly. However, this also means that the psychological games we consider do not belong to the class of psychological games that receives most attention in Geanakoplos et al (1989), as they confine attention to psychological games where only *initial beliefs* have a direct bearing on players payoff perception (although they suggest (p78) that other assumptions may be important).

We deal with extensive games without nature and with perfect recall. Any such game Γ is a quintuple which specifies respectively the game tree, the player partition, the information partition, the choice partition, and the assignment of payoffs to endnodes. We refer to standard texts (for example van Damme (1991, Chapter 6)) for the general formalism of extensive games and state only those basic and derived concepts we shall need. Given an extensive game Γ, let R be the set of nodes that are roots of subgames, and Γ^r the subgame of Γ which has $r \in R$ as its root. Define the *depth* of a subgame as the number of its proper subgames. Since we need to keep track of how the players' beliefs change as new subgames are reached, it is convenient to introduce the following new concept: Define the *r-part of* Γ^r as the set of vertices in Γ^r that do not appear in some proper subgame of Γ^r.

Let $N=\{1,...,n\}$ be the set of players where $n \geq 2$. Let A_i be the non-empty set of (behavior) strategies of $i \in N$ – each strategy assigns to each of i's information sets a probability distribution on the set of possible choices at that information set (and if i owns no information set, A_i is defined to be a singleton). Define $A=_{i \in N} A_i$. For any a_i, $a_i' \in A_i$ and $\mu \in [0,1]$, let $\mu \cdot a_i + (1-\mu) \cdot a_i'$ be the strategy that at any given information set of player i with probability μ prescribes the same choice behavior as a_i, and with probability $1-\mu$ the same choice behavior as a_i'. For any $a_i \in A_i$ and $r \in R$, let $a_i(r)$ be the strategy that prescribes the same choices as a_i, except on the path to r where choices are made in accordance with that path. By the definition of a subgame (see van Damme (1991)), $a_i(r)$ is uniquely defined. For any $a=(a_i)_{i \in N} \in A$ and $r \in R$, let $A_i(r,a) \subseteq A_i$ be the set of strategies that prescribe the same choices as the strategy a_i at all information sets outside the r-part of Γ^r. That is, $A_i(r,a)$ is the set of strategies i may use if he behaves according to a_i at information sets outside the r-part of Γ^r, but is free to make any choices in the r-part of Γ^r. Note that $A_i(r,a) \neq \varnothing$ since $a_i(r) \in A_i(r,a)$. Note also that if no

information sets of *i*'s appears in the *r*-part of Γ^r, then $A_i(r,a_i)=\{a_i(r)\}$. Define $A(r,a)=x_{i\in N}A_i(r,a)$.

Using the assignment of payoffs to endnodes, we can derive a payoff function for each player which depends on what profile in A is played. Let π_i: $A\rightarrow\mathbb{R}$ denote this function. We shall refer to π_i as player *i*'s *material payoff function*. We interpret material payoffs as representing money, or some other objectively measurable quantity. However, the material payoff is not the only payoff which we shall assume motivates *i* in his decision making. To get *i*'s *utility*, which is the function that *i* wants to maximize, we shall add a *reciprocity payoff* to *i*'s material payoff. We do this by rebuilding Γ as a psychological game in which the players have explicitly specified beliefs about one another's actions, and about one anothers beliefs about one another's actions.

We represent beliefs as behavior strategies. However, in order to avoid confusion, we introduce separate notation for beliefs. Let $B_{ij}=A_j$ be the set of possible beliefs of player *i* about the strategy of player *j*. Let $C_{ijk}=B_{jk}=A_k$ be the set of possible beliefs of player *i* about a belief of player *j* about the strategy of player *k*. For any $b_{ij}\in B_{ij}$, $c_{ijk}\in C_{ijk}$, and $r\in R$, define $b_{ij}(r)$ and $c_{ijk}(r)$ in a fashion completely analogous to that pertaining to strategies.

We wish to capture that each player *j* to some extent wants to be kind in return to any player *i* who is kind to *j*. What does it mean for *i* to be kind to *j*? Suppose that *i* chooses $a_i \in A_i$ and that he believes that all other players make choices according to the profile $(b_{ij})_{j\neq i}\in\times_{j\neq i}B_{ij}$. Following Rabin (1993), we note that player *i* then believes that he chooses in such a way that *j*'s material payoff will be $\pi_j(a_i,(b_{ij})_{j\neq i})$. He also believes that the feasible set of material payoffs for *j* is $\{\pi_j(a_i',(b_{ij})_{j\neq i})|a_i'\in A_i\}$. How kind *i* is to *j* can now be measured in terms of the relative size of $\pi_j(a_i,(b_{ij})_{j\neq i})$ within this set.

While this measurement may be done in several ways, there is one particular aspect that must handled carefully. Consider the following game Γ_3 which is related to Γ_2 in Figure 2:

(Insert Figure 3)

Suppose 1 plays the strategy *D*, and suppose he believes with probability one that player 2 is playing the strategy *cd* (any other belief will in fact also do to make our point). One sees that 1 believes he chooses the material payoff $\pi_j(D,cd)=0$ for player 2, from the feasible set of material payoffs for *j* which is [-1000,2]. Within this set, 0 is a rather large number. Should one therefore conclude that player 1 is rather kind by choosing *D*? We would find this unreasonable. The fact that *W* is a possible choice for 1 seems to be irrelevant for drawing conclusions regarding the kindness of the choices *C* and *D*. The choice of *W* guarantees an inefficient outcome which hurts both players. By contrast, each of the actions C and *D* may lead to outcomes that are efficient in terms of material payoff allocations.

We propose that 1's kindness if he chooses *D* in Γ_3 should be the same as if 1 chooses *D* in Γ_2, if 1 has the same beliefs in the two cases. That is, 1's kindness should be assessed with reference to the relative position of $\pi_2(D,cd)=0$ for player 2 in the set $[\pi_2(D,cd), \pi_2(C,cd)]=[0,2]$. Since 0 is the lowest number in this set, player 1 should be considered unkind if he chooses *D*.

In general, we proceed as follows. Define player *i*'s *efficient strategies* by

$E_i = \{a_i \in A_i |$ there exists no $a_i' \in A_i$ such that for all $r \in R$, $(a_j)_{j\neq i} \times_{j\neq i} A_j$, $k \in N$ it holds that
$\pi_k(a_i'(r),(a_j(r))_{j\neq i}) \geq \pi_k(a_i(r),(a_j(r))_{j\neq i})$, with strict inequality for some $(r,(a_j)_{j\neq i},k)\}$

Intuitively, a strategy is inefficient if there exists another strategy which, conditional on any history of play and subsequent choices by the others, provides no lower material payoff for any player, and a higher material payoff for some player for some history of play and subsequent choices by the others. For example, in Γ_1 and Γ_2 all strategies are efficient for both players. In Γ_3 all strategies are efficient, except those strategies of player 1 that assign positive probability to the choice *W*.

In order to define how kind *i* is to *j* we expand on an idea of Rabin's (1993): *i*'s kindness is zero if he believes that *j*'s material payoff will be the average between respectively the lowest and highest material payoff of *j's* that is compatible with *i* choosing an efficient strategy.[16] It

[16] In principle also other weighted averages could be used without any basic change of subsequent results.

is convenient to have a special notation which describes this number as a function of *i*'s beliefs about the profile being played. We call this function $\pi_j^{e_i}$, defined by

$$\pi_j^{e_i}((b_{ij})_{j\neq i}) = \tfrac{1}{2} \cdot [max\{\pi_j(a_i,(b_{ij})_{j\neq i})|\ a_i \in E_i\} + min\{\pi_j(a_i,(b_{ij})_{j\neq i})|\ a_i \in E_i\}]$$

Think of $\pi_j^{e_i}((b_{ij})_{j\neq i})$ as a norm for *i* describing the "equitable" payoffs for player *j* when *i*'s beliefs about other players' behavior are summarized by $(b_{ij})_{j\neq i}$. We use $\pi_j^{e_i}((b_{ij})_{j\neq i})$ as a reference point for measuring how kind *i* is to *j*. If *i* chooses a strategy a_i such that $\pi_j(a_i,(b_{ij})_{j\neq i}) = \pi_j^{e_i}((b_{ij})_{j\neq i})$, then his kindness is zero. Otherwise *i*'s kindness to *j* is proportional to how much more or less material payoff than $\pi_j^{e_i}((b_{ij})_{j\neq i})$ that *i* thinks will be the consequence for *j*. More specifically:

Definition 1. The kindness of player *i* to another player $j \neq i$ is given by the function

$\kappa_{ij}: A_i \times_{j\neq i} B_{ij} \rightarrow \mathbb{R}$ defined by

$$\kappa_{ij}(a_i,(b_{ij})_{j\neq i}) = \pi_j(a_i,(b_{ij})_{j\neq i}) - \pi_j^{e_i}((b_{ij})_{j\neq i})$$

Intuitively, Definition 1 reflects the idea i's kindness to *j* is proportional to "the size of his gift". There are many conceivable functional forms for κ_{ij} that could capture this idea. All examples in this paper, as well as analogues of the existence theorem in Section 5, work in much the same way with many such functions. We discuss this further in the Remark after the Theorem in Section 4, and also in Section 5 when we compare our model to that of Rabin (1993). Definition 1 is easiest to apply though, so we now proceed using this formulation.

Having defined kindness, we now turn to reciprocity—the idea that if *j* is kind (unkind) to *i*, then *i* wants to be kind in return (take revenge). Since *j*'s kindness depends on *j*'s beliefs, *i* cannot observe *j*'s kindness directly. However, *i* can consult his beliefs about *j*'s actions and beliefs and draw inferences concerning *j*'s kindness. We introduce a function λ_{iji} to keep track of how kind *i* believes that *j* is to *i*:

Definition 2. Player *i*'s beliefs about how kind player $j \neq i$ is to *i* is given by the function

$\lambda_{iji}: B_{ij} \times_{k\neq j} C_{ijk} \rightarrow \mathbb{R}$ defined by

$$\lambda_{iji}(b_{ij},(c_{ijk})_{k\neq j}) = \pi_i(b_{ij},(c_{ijk})_{k\neq j}) - \pi_i^{e_j}((c_{ijk})_{k\neq j})$$

Note that since $B_{ij}=A_j$ and $C_{ijk}=B_{jk}$, the function λ_{iji} is formally equivalent to the function κ_{ji} although it captures a psychological component that pertains to player *i*, not player *j*.

It is now time to specify the utilities which the players are assumed to maximize:

Definition 3. Player *i*'s *utility* is a function $U_i: A_i \times_{j\neq i}(B_{ij}\times_{k\neq j} C_{ijk}) \rightarrow \mathbb{R}$ defined by

$$U_i(a_i,(b_{ij},(c_{ijk})_{k\neq j})_{j\neq i}) = \pi_i(a_i,(b_{ij})_{j\neq i}) + Y_i \cdot \Sigma_{j\in N\setminus\{i\}}\, \kappa_{ij}(a_i,(b_{ij})_{j\neq i}) \cdot \lambda_{iji}\,(b_{ij},(c_{ijk})_{k\neq j}),$$

where Y_i is an exogenously given non-negative number.

Player *i*'s utility is the sum of *n* terms. The first term is his material payoff, the remaining terms express his *reciprocity payoff with respect to each player j≠i*. The constant Y_i measures how sensitive *i* is to reciprocity concerns. If $Y_i>0$ the following is true: If *i* believes that *j* is kind to him (i.e., $\lambda_{iji}(\cdot)>0$), then *i*'s reciprocity payoff with respect to *j* is increasing in *i*'s kindness to *j*. Furthermore, the higher is $\lambda_{iji}(\cdot)$, the more material payoff *i* is willing to give up in order to do *j* a favour. If *i* believes that *j* is unkind to him (i.e., $\lambda_{iji}(\cdot)<0$), then *i*'s reciprocity payoff with respect to *j* is decresing in *i*'s kindness to *j*. This is the way in which U_i reflects the idea that if *i* thinks that *j* is kind (unkind) to him, then *i* wants to be kind in return (take revenge). Of course, when *i* optimizes he may have to make tradeoffs between various reciprocity payoffs with respect to different players as well as his material payoff.

The specification in Definition 3 has a particular feature to which we shall return later. To facilitate reference, we now discuss this in a remark:

Remark. We choose to work with utilities as given by Definition 3 because this is the simplest formulation we can think of that invokes a concern for reciprocity. However, it is important to note that Definition 3 has the specific drawback that U_i will not represent *i*'s preferences in a way which is invariant with respect to the choice of monetary units. To see this, note that if *i*'s monetary payoff is measured in dollars, then the reciprocity payoff will have the dimension of dollars squared! In principle this problem could be solved by defining player *i*'s reciprocity payoff with respect to each player *j* as *Yi* times the square root of the absolute value of $\kappa_{ij}(\cdot)\cdot\lambda_{iji}(\cdot)$, adjusted so as to maintain the right sign. If one wanted to estimate a parameter like Y_i based on experimental data for different games, it would

probably be sensible to adopt such an approach. However, doing so greatly complicates the calculation of examples, so in what follows we work with Definition 3, and imagine that Y_i is selected to adequately reflect *i*'s sensitivity to reciprocity concerns *after* the choice of monetary unit has been made.[17]

We can now append any extensive game Γ with a vector of utilities $(U_i)_{i\in N}$ defined as above and get the tuple $\Gamma^*=(\Gamma, (U_i)_{i\in N})$. We refer to any Γ^* constructed in this fashion as a *psychological game with reciprocal incentives*. Note that such a Γ^* is not a "game" in the traditional sense, since the utility functions U_i are defined on domains that include subjective beliefs, and not only strategic choices.

We propose a solution concept which is applicable to any psychological game with reciprocal incentives. In the spirit of Rabin (1993), we look for equilibria in which each player chooses an optimal strategy given his beliefs which furthermore are required to be correct. We also wish to impose a requirement of sequential rationality in every subgame and make the following assumptions: Suppose player *i* plays the strategy $a_i\in A_i$, initially believing the others to play $(b_{ij})_{j\neq i}$. In a subgame Γ^r, we model player *i* as playing strategy $a_i(r)\in A_i$, believing the others to play $(b_{ij}(r))_{j\neq i}$. Of course, if *r* is the root of Γ, then $a_i(r)=a_i$ and $(b_{ij}(r))_{j\neq i}=(b_{ij})_{j\neq i}$. In any Γ^r which is a proper subgame of Γ, players are assumed to be motivated according to the beliefs they hold at *r* at all information sets that appear in the *r*-part of Γ^r.

Definition 4. The profile $a^*=(a_i^*)_{i\in N}$ is a *sequential reciprocity equilibrium* (*SRE*) if for all $i\in N$ and for all $r\in R$ it holds that

(1) $a_i^*(r) \in \text{argmax}_{a_i\in A_i(r,a^*)} U_i(a_i,(b_{ij}(r),(c_{ijk}(r))_{k\neq j})_{j\neq i}))$

(2) $b_{ij} = a_j^*$ for all $j\neq i$

(3) $c_{ijk} = a_k^*$ for all $j\neq i$, $k\neq j$

[17] There are in fact many alternative ways to specify the utility function of Definition 3 so as to capture the basic reciprocity properties we are after. All have their pros and cons. We discuss this further in the Remark following the Theorem in Section 4, and in Section 5 when we compare our model to that of Rabin (1993).

By condition (1) of Definition 4, a SRE is a strategy profile such that in the *r*-part of each subgame Γ^r each player makes choices which maximizes his utility given his beliefs and given that he follows his equilibrium strategy outside the *r*-part of Γ^r. If $\Gamma^r=\Gamma$, conditions (2) and (3) guarantee that his initial beliefs are correct. At any subsequent subgame Γ^s condition (1) requires that beliefs assign probability one to the sequence of choices that allow $s\in R$ to be reached, but are otherwise as the initial beliefs.

3. APPLICATIONS

In this section we apply our model to two well known and experimentally tested games. This serves the purpuse of showing how reciprocity motives shape the analysis. Furthermore, the applications exemplify how to calculate sequential reciprocity equilibria.

a) The Sequential Prisoners' Dilemma

The first game we analyze is the Sequential Prisoners' Dilemma Γ_2 of Figure 2,[18] which is a very simplified version of the Fair Wage-Effort model.

We first analyze player 2's behavior, which is summarized by two observations:

Observation 1: If player 1 defects (chooses *D*), player 2 also defects in every SRE.

To see this, note that only the reciprocity payoff can conceivably make 2 choose *c*, as the material payoff *per se* dictates a choice of *d* for 2. However, for any possible strategy of 2, player 2 gets less when 1 chooses *D* than when he chooses *C*. Whatever 1 believes about 2's strategy, 1's choice of *D* is unkind, and hence 2 must believe that 1 is unkind. Hence, the reciprocity payoff as well as the material payoff makes player 2 choose *d*.

[18]We restrict our attention to equilibria for reciprocity parameters Y_1 and Y_2 that are generic such that the conditions on those paramenters used for the characterisation of equilibria (see below) are never fulfilled with equality.

Observation 2: If player 1 cooperates, the following holds in all SRE:

a) If $Y_2 > 1$, player 2 cooperates.

b) If $Y_2 < 0.5$, player 2 defects.

c) If $0.5 < Y_2 < 1$, player 2 cooperates with a probability of $p = \frac{2 \cdot Y_2 - 1}{Y_2}$.

To see this, notice that if 1 cooperates, 2 can give 1 a material payoff of at least -1 and at most 1, so the "equitable" payoff of 1 is zero. If 2 chooses cooperation, 1 receives 1. Therefore, 2's kindness of cooperation is 1. Similarly, 2's kindness of defection is -1. In order to calculate how kind 2 believes 1 is after choosing *C* we have to specify 2's belief of 1's belief about 2's choice after *C*.[19] Denote this by *p*". Then 2's belief about how much payoff 1 intends to give to 2 by choosing *C* is $p'' \cdot 1 + (1 - p'') \cdot 2$, and since 2's payoff resulting from 1's choice of *D* would be zero,[20] 2's belief about 1's kindness from choosing *C* is $p'' \cdot 1 + (1 - p'') \cdot 2 - 0.5(p'' \cdot 1 + (1 - p'') \cdot 2 + 0) = 1 - 0.5 \cdot p''$. This implies that when 1 cooperates and the second order belief is *p*", 2's utility of cooperation is given by $1 + Y_2 \cdot 1 \cdot (1 - 0.5 \cdot p'')$, whereas 2's utility of defection is $2 + Y_2 \cdot (-1) \cdot (1 - 0.5 \cdot p'')$. The former is larger than the latter if $Y_2(2 - p'') > 1$. In equilibrium, the second order belief must be correct. Hence, if in equilibrium 2 cooperates, the condition must hold for $p'' = 1$. This is the case if $Y_2 > 1$. On the other hand, if in equilibrium 2 defects, the condition must not hold for $p'' = 0$. This implies that $Y_2 < 0.5$. For intermediate values of Y_2 $(0.5 < Y_2 < 1)$ neither cooperation nor defection can be part of an equilibrium. In order to have a mixed equilibrium, the utility of cooperation must be equal to the utility of defection. This is the case when $Y_2(2 - p'') = 1$. Since in equilibrium the second order belief must be correct, the actual probability of cooperation, *p*, must be such that the condition is fulfilled. This implies that $p = \frac{2Y_2 - 1}{Y_2}$.

---

[19] In principle we also need 2's belief about 1's behavior. However, we must only care about beliefs that are in accordance with reaching the node under consideration. After 1 has already choosen C, there is only one such belief, namely 1 choosing C. To put it differently: 2 already knows what 1 has done, and 2's belief has to be in accordance with her knowledge.

[20] Recall that in any SRE player 2 defects after a defection of 1 (see Observation 1).

Notice that the probability p=0 for $Y_2 = 0.5$, and p=1 for $Y_2 = 1$. Hence, Observation 1 and 2 together imply that for a given parameter Y_2 2's equilibrium behavior is unique. This is, however, in general not true for 1's behavior that can be characterized by three observations:

Observation 3: If $Y_2 < 0.5$, defection is 1's unique equilibrium behavior.

To see this, notice that for $Y_2 < 0.5$ player 2 always defects (see Observation 1 and 2). Hence, only the reciprocity part of the utility function can make 1 choose *C* (the material payoff alone would dictate for 1 to choose *D*). However, for any second order belief about 1's behavior 2's strategy of always defecting is unkind. Hence, the reciprocity payoff as well as the material payoff makes player 1 choose *D*.

Observation 4: If $Y_2 > 1$, 1's equilibrium behavior is characterized by one of the three following possibilities:

a) Player 1 cooperates (regardless of Y_1).

b) $Y_1 > 1$ and player 1 defects.

c) $Y_1 > 1$ and player 1 cooperates with probability $q = \frac{Y_1 - 1}{2Y_1}$.

To see this, note that $Y_2 > 1$ implies that 2 cooperates when 1 cooperates and defects when 1 defects (see Observation 1 and 2). Hence, 1 can give 2 a material payoff of at least 0 and at most 1. Hence, the "equitable" payoff of 1 is 0.5. If 1 chooses cooperation, 2 receives 1. Therefore, 1's kindness of cooperation is 0.5. Similarly, 1's kindness of defection is -0.5. In order to calculate how kind 1 believes that 2 is we have to specify 1's belief about what 2 believes that 1 will do. Denote by *q"* this second order belief of 1 choosing *C*. Then 1 believes that 2 believes that she gives player 1 a material payoff of $q'' \cdot 1 + (1 - q'') \cdot 0$ by choosing her equilibrium strategy. If 2 always cooperates, 1's payoff is $q'' \cdot 1 + (1 - q'') \cdot 2$, whereas if 2 always defects, 1's payoff is $q'' \cdot (-1) + (1 - q'') \cdot 0$. Hence, 1's belief about 2's kindness from choosing *c* after *C* and *d* after *D* is given by: $q'' \cdot 1 + (1 - q'') \cdot 0 - 0.5 \cdot (q'' \cdot 1 + (1 - q'') \cdot 2 + q'' \cdot (-1) + (1 - q'') \cdot 0) = 2 \cdot q'' - 1$. This implies that when 2 plays the equilibrium strategy and the second order belief is *q"*, 1's utility of cooperation is given by $1 + Y_1 \cdot 0.5 \cdot (2 \cdot q'' - 1)$, whereas 1's utility of defection is

$0 + Y_1 \cdot (-0.5) \cdot (2 \cdot q'' - 1)$. The former is larger than the latter if $1 + Y_1 \cdot (2 \cdot q'' - 1) > 0$. In equilibrium, the second order belief must be correct. Hence, if in equilibrium 1 cooperates, the condition must hold for $q'' = 1$, which is always the case.

On the other hand, if in equilibrium 1 defects, the condition must not hold for $q'' = 0$. This implies that $Y_1 > 1$.

In order to have a mixed equilibrium, the utility of cooperation must be equal to the utility of defection. This is the case when $1 + Y_1 \cdot (2 \cdot q'' - 1) = 0$. Since in equilibrium the second order belief must be correct, the actual probability of cooperation, *q*, must be such that the condition is fulfilled. This implies that $q = \frac{Y_1 - 1}{2Y_1}$.

Observation 4a corresponds to the intuitively most plausible equilibrium — since 2 is using strategy *cd*, 1's material payoff as well as his reciprocity payoff leads him to cooperate. If, however, reciprocity is important enough, there also exists other equilibria that are characterized by "self-fulfilling prophecies". If 1 believes that 2 initially believes that 1 chooses *D*, and that 2 defects in that case, then 1 believes that 2 is unkind. This in turn leads 1 to be unkind, i.e. to play *D* (or to mix). Of course, such a mechanism only works when 1 is motivated enough by reciprocity - if this is not the case, 1's material payoff together with the reciprocal behavior of 2 make him cooperate.

Next we turn to the equilibrium behavior when 2 is moderately motivated by reciprocity and hence answers a cooperative choice of 1 with mixing.

Observation 5: If $0.5 < Y_2 < 1$, 1's equilibrium behavior is characterized by one of the three following possibilities:

a) $Y_2 > \frac{2}{3}$ and player 1 cooperates.

b) $Y_1 > 3Y_2 - 2$ and player 1 defects.

c) $\frac{(3Y_2 - 2)Y_2}{5Y_2 - 2} < Y_1 < 3Y_2 - 2$ and player 1 cooperates with probability $q = \frac{Y_2(3Y_2 - Y_1 - 2)}{2Y_1(2Y_2 - 1)}$.

To see this, notice that $0.5 < Y_2 < 1$ implies that 2 cooperates with probability $p = \frac{2}{Y_2}$ when 1 cooperates, and 2 defects when 1 defects (see Observations 1 and 2). Hence, 1 can

give 2 a material payoff of at least 0 and at most $p \cdot 1 + 2 \cdot (1-p)$. Hence, the "equitable" payoff of 1 is $0.5(p + 2 \cdot (1-p)) = \frac{2-p}{2}$. If 1 chooses cooperation, 2 receives $p \cdot 1 + 2 \cdot (1-p)$. Therefore, 1's kindness of cooperation is $\frac{2-p}{2}$. Similarly, 1's kindness of defection is $-\frac{2-p}{2}$. In order to calculate how kind 1 believes 2 is we have to specify 1's belief about what 2 believes that 1 will do. Denote by *q"* this second order belief of 1 choosing C. Then 1 believes that 2 believes that she gives player 1 a material payoff of $q''(p \cdot 1 + (1-p) \cdot (-1)) + (1-q'') \cdot 0$ by her equilibrium strategy. If 2 always cooperates, 1's payoff is $q'' \cdot 1 + (1-q'') \cdot 2$, whereas if 2 always defects, 1's payoff is $q'' \cdot (-1) + (1-q'') \cdot 0$. Hence, 1's belief about 2's kindness of her equilibrium strategy is $q''(p \cdot 1 + (1-p) \cdot (-1)) + (1-q'') \cdot 0 - 0.5(q'' \cdot 1 + (1-q'') \cdot 2 + q'' \cdot (-1) + (1-q'') \cdot 0) = 2q''p - 1$. This implies that when 2 plays the equilibrium strategy and the second order belief is *q"*, 1's utility of cooperation is given by $p + (1-p) \cdot (-1) + Y_1 \frac{2-p}{2}(2q''p - 1)$, whereas 1's utility of defection is $0 + Y_1(-\frac{2-p}{2})(2q''p - 1)$. The former is larger than the latter if $2p - 1 + Y_1(2-p)(2q''p - 1) > 0$. In equilibrium, the second order belief must be correct. Hence, if in equilibrium 1 cooperates, the condition must hold for $q'' = 1$, which happens if $p > 0.5$. This in turn implies that $Y_2 > \frac{2}{3}$ (see the calculation of *p* in Observation 2 c).

On the other hand, if in a SRE 1 defects, the condition must not hold for $q'' = 0$. Inserting for p and rearranging terms this leads to $Y_1 > 3Y_2 - 2$.

In order to have a mixed equilibrium, utility of cooperation must be equal to the utility of defection. This is the case when $2p - 1 + Y_1(2-p)(2q''p - 1) = 0$. Since in equilibrium the second order belief must be correct, the actual probability of cooperation, *q*, must be such that the condition is fulfilled. Substituting for *p* this implies that $q = \frac{2 - 2 + 1 - }{2Y_1(2Y_2 - 1)}$. The other conditions of Observation 5c are necessary to guarantee that *q* is larger than zero and smaller than 1.

Like in Observation 4, the first of these cases is the intuitively plausible one — if 2 reciprocates with a high enough probability, 1 cooperates because of his material payoff as well as because of reciprocity reasons. If, however, reciprocity is important enough, there also exists other equilibria that are characterized by self-fulfilling prophecies: If 1 believes that 2 initially believes that 1 chooses *D*, and that 2 defects in that case, 1 expects an unkind action of 2. This in turn leads 1 to be unkind, i.e. to play indeed *D* (or to mix). Of course, this

mechanism only works when 1 is enough motivated by reciprocity — if this is not the case, 1's material interest together with the reciprocal behavior of 2 make him cooperate.

To summarize, SREs in the Sequential Prisoners Dilemma are characterized by the fact that - depending on 2's reciprocity inclination - the equilibrium behavior of player 2 is unique. If confronted with defection 2 also defects. If 1 cooperates, 2 defects when the psychological payoff plays only a little role for him, she cooperates when she is very motivated by reciprocity considerations, and she plays a mixed strategy for intermediate levels of her reciprocity parameter. Player 1's equilibrium behavior is only unique when 2 always defects. In this case, he also defects irrespectively of his own type. If, however, 2 reciprocates with a high enough probability then there exists a SRE where hhe cooperates irrespectively of his own type. But besides of that there also exists two self-fulfilling prophecy equilibria when the impact of reciprocity considerations on 1 is large enough. These self-fulfilling prophecy equilibria are characterized by a mutual distrust and hence noncoopertion.

b) The Ultimatum Game

The next game we analyse is the Ultimatum Game. In this game a monetary “pie” of c units of money has to be divided between two persons. Player 1 offers player 2 an integer amount x between zero and c. Hence, player 1's strategy set is given by $\{0,1,...,c-1,c\}$. Player 2 may either accept *(A)* or reject *(R)* the offer. Hence, a pure strategy of 2 is a function from the set of possible offers into $\{A,R\}$. If the responder accepts the offer, she gets the offered amount and the proposer the rest. In case of rejection, both get nothing. Hence, as long as player 2 is only motivated by her material payoff, i.e. by the money she gets, she should accept every positive offer. Knowing that, a purely "materialistic" proposer should offer zero (or one monetary unit) in any subgame perfect equilibrium.

In order to characterize the sequential reciprocity equilibria,[21] notice first that in every SRE player 2 will accept the highest possible offer of $x=c$. In case of such an offer, player 1 gets

[21] In this application we restrict our attention to pure strategy equilibria. Furthermore, like in the previous application we characterize the equilibria only for reciprocity parameters Y_1 and Y_2 that are generic such that

zero irrespectively of 2's decision. Hence, 2 can neither be kind nor unkind to 1, and hence only the material payoff matters in that case. Therefore, 2 accepts an offer of c. On the other hand, 1 can always guarantee that 2 earns nothing (by offering zero). Hence, in any equilibrium 2's equitable payoff $\pi_2^{e_1}$ used for the calculation of 1's kindness $\kappa_{12}(\cdot)$ or 2's belief of 1's kindness $\lambda_{212}(\cdot)$ (see Definitions 1 and 2) is *c/2*. Furthermore, notice that each strategy of 2 which involves rejecting any offer is not efficient. Hence, 2 has only one efficient strategy, namely accepting every possible offer, and 1's payoff from 2's efficient strategy is *(c-x)*. Therefore, 1's equitable payoff for any possible offer *x* is *(c-x)*. Hence, in case 1 makes an offer that is rejected by 2, 2's kindness equals *0-(c-x)=-(c-x)*, whereas in case of acceptance 2's kindness is zero.

We now turn to the behavior of player 2. It can be characterized by three observations.

Observation 1: In every SRE, 2 accepts all offers larger than $\frac{Y_2 c^2}{2+cY_2}$.

To see this, assume to the contrary that there exists a SRE where such a high offer would be rejected. Since beliefs are correct in equilibrium, this requires that 2 believes that 1 believes that his offer will be rejected. Hence, for all second order beliefs that are consistent with rejecting the offer, the 2's belief in 1's kindness, $\lambda_{212}(\cdot)$, equals *-c/2*. Recall that 2's kindness from rejection is *-(c-x)*, and from acceptance is zero. Therefore, 2's utility from a deviation to acceptance is given by $x + Y_2 \cdot 0 \cdot (-c/2) = x$, whereas rejection gives 2 a utility of $0 + Y_2(-c+x)(-c/2)$. For *x* larger than $\frac{c}{2+cY_2}$, the former is higher than the latter. Therefore, to maximize her utility 2 should deviate to acceptance - a contradiction.

Observation 2: If $Y_2 > 0$, 2 rejects all offers smaller than $\frac{2+3Y_2c-\sqrt{4+12Y_2c+Y_2^2c^2}}{4Y_2}$.

To see this, assume to the contrary that there exists a SRE such that such a small offer would be accepted. In equilibrium beliefs are correct, which requires that 2 believes that 1 believes that his offer will be accepted. Hence, for all second order beliefs that are consistent with

---

the conditions on those paramenters used for the characterisation of equilibria (see below) are never fulfilled with equality.

accepting the offer in equilibrium, 2's belief in 1's kindness, $\lambda_{212}(\cdot)$, equals *x-c/2*. Recall that 2's kindness from rejection is $-(c-x)$, and from acceptance is zero. Therefore, 2's utility from accepting is given by $x + Y_2 \cdot 0 \cdot (x - c/2) = x$, whereas deviating to rejection gives 2 a utility of $0 + Y_2(-c + x)(x - c/2)$. For *x* smaller than $\frac{2 + 3Y_2 c - \sqrt{4 + 12Y_2 c + Y_2^2 c^2}}{4Y_2}$, the latter utility is higher than the former. Therefore, the 2 should deviate to rejection - a contradiction.

One can show that the boundary value for rejection in Observation 1 is larger than the boundary value for acceptance in Observation 2[22]. Hence, there exists an intermediate range of offers where the 2's reaction is not yet characterized. We do this in

Observation 3: Any possible reaction of 2 following an offer of *x* with $\frac{2 + 3Y_2 c - \sqrt{4 + 12Y_2 c + Y_2^2 c^2}}{4Y_2} < x < \frac{Y_2 c^2}{2 + cY_2}$ is part of an SRE.

To see this, assume that 2 accepts in equilibrium. Correctness of beliefs requires that 2 believes that 1 believes that his offer will be accepted. Hence, for all second order beliefs that are consistent with accepting the offer, the 2's belief in 1's kindness 2, $\lambda_{212}(\cdot)$, equals *x-c/2*. Recall that 2's kindness from rejection is *-(c-x)*, and from acceptance is zero. Therefore, 2's utility from accepting is given by $x + Y_2 \cdot 0 \cdot (x - c/2) = x$, whereas deviating to rejection would give 2 a utility of $0 + Y_2(-c + x)(x - c/2)$. For *x* larger than $\frac{2 + 3Y_2 c - \sqrt{4 + 12Y_2 c + Y_2^2 c^2}}{4Y_2}$, the former utility is higher than the latter. Therefore, acceptance is indeed a best reponse.

[22] This can be most easily seen by transforming the conditions in Observations 1 and 2 to be conditions on Y_2.

Now assume that in equilibrium 2 rejects the same offer *x*. Correctness of beliefs requires that 2 believes that 1 believes that his offer will be rejected. Hence, for all second order beliefs that are consistent with rejecting the offer, the 2's belief in 1's kindness, $\lambda_{212}(\cdot)$, equals *-c/2*. Recall that 2's kindness from rejection is *-(c-x)*, and from acceptance is zero. Therefore, 2's utility from a deviation to acceptance is given by $x + Y_2 \cdot 0 \cdot (-c/2)$, whereas rejection gives 2 a utility of $0 + Y_2(-c + x)(-c/2)$. For *x* smaller than $\frac{2}{2 + cY_2}$, the latter is larger than the former. Therefore, rejection is also a best reponse.

This intermediate range of offers is characterized by self-fulfilling prophecies. Suppose the beliefs are such that the offer is rejected, which makes the offer a very unkind action of 1 (2's expected material payoff is zero). This in turn leads 2 to be unkind to 1, and the only way she can do so is by rejecting the offer. On the other hand, if the beliefs imply acceptance of the offer, the offer is not so unkind (since in case of acceptance 2's material payoff is *x*), and hence 2 does not opt for the unkind action - 2 accepts. Because of these selfullfilling prophecies the SRE of this game is not unique even for generic material payoffs.

Notice further that if Y_2 converges to zero, the boundary value for 2 to accept in Observation 1 converges to zero - 2's SRE-behavior converges to the responder's subgame perfect equilibrium behavior in the game without reciprocity considerations.

We now turn to the behavior of the proposer which is characterized by the two observations:

Observation 4: For all Y_1, Y_2 there exists an SRE such that 1 makes the lowest offer that is acceptable for 2 (according to 2's equilibrium strategy).

To see this, assume that 1 makes the lowest acceptable offer and that this is part of an SRE. Correctness of beliefs requires that 1 believes that 2 will accept it. Hence 1 believes that 2's kindness is zero (recall that 2 can never be kind, only unkind or "neutral" to 1). Therefore, 1's reciprocity payoff is zero irrespectively of what he does, and hence only the material payoff shapes 1's decision. 1's material payoff is of course largest if 1 makes the lowest acceptable offer. Hence, such an offer maximizes 1's utility.

Observation 5: If the reciprocity parameters of both players are large enough, then there exists a SRE where 1 makes an offer that is not accepted by 2.

Notice first that if Y_2 is large enough, it follows from Observation 2 there exists offers that will be rejected for sure. Now assume that in equilibrium 1 makes such an offer, denoted by *x*. Since in equilibrium beliefs are correct, 1's belief in 2's kindness, $\lambda_{121}(\cdot)$, is $-c+x$. Furthermore, 1's kindness from an unacceptable offer is given by *-c/2* (since 2's material payoff is zero and the equitable payoff, as already noted, is *c/2*). Hence, 1's utility from making the unacceptable offer is $0+Y_2(c-x)c/2$. Deviating to another unacceptable offer does not change 1's utility: each player's material payoff remains the same, hence 1's kindness does not change, and a deviation never changes the kindness expected from the other player. On the other hand, a deviation to an acceptable offer changes 1's utility. Since 2's kindness is never positive, 1 will never make an acceptable offer higher than the lowest acceptable offer that we denote by $\underline{x}$. If 1 offers $\underline{x}$, his kindness changes to $\underline{x}-c/2$. Hence, if 1 deviates from *x* to $\underline{x}$, his utility changes to $(c-\underline{x})+Y_1(-c+x)(\underline{x}-c/2)$. If Y_1 is large enough, i.e. if $Y_1 > \frac{c-\underline{x}}{(c-x)\underline{x}}$, this utility from deviating is lower than the utility of the equilibrium offer *x*.

To summarize, in each SRE the responder accepts high offers and rejects low offers. Due to self-fulfilling prophecies, acceptance as well as rejection is part of SREs for intermediate offers. For all possible parameter values there exists SREs such that the proposer makes the lowest offer acceptable for the responder. Furthermore, if reciprocity is important enough there exists equilibria that are characterized by rejected offers. In these equilibria players hold beliefs that make them view one another as unkind, which in turn leads the players to be unkind in return.

4. EXISTENCE

THEOREM. *There exists a SRE in every psychological game with reciprocal incentives.*

Proof. The idea of the proof is to construct a particular profile $\alpha \in A$ which turns out to be a SRE. Consider a subgame Γ^r that has zero depth. For each $i \in N$ define a correspondence β_i^r: $A \to A_i$ by $\beta_i^r(a)$=argmax $_{a_i' \in A_i} U_i(a_i'(r),(a_j(r),(a_k(r))_{k\neq j})_{j\neq i}))$. Note two things (i) and (ii): (i) By standard arguments one sees that the combined correspondence β^r: $A \to A$ defined by

$\beta^r(a)=(\beta_i^r(a))_{i\in N}$ admits a fixpoint.[23] (ii) By inspection of β_i^r one sees that if $a_i' \in \beta_i^r(a)$ and if $a_i'' \in A_i$ prescribes the same choices as a_i' at all information sets in Γ^r, then it must hold that $a_i'' \in \beta_i^r(a)$. Combining (i) and (ii) one infers that there exist some profile a such that for *any* Γ^r with $r \in R$ that has zero depth it holds that $a \in \beta^r(a)$. Pick any such profile and call it $\alpha^0=(\alpha_i^0)_{i\in N}$.

If Γ itself has zero depth, set $\alpha=\alpha^0$. Else, let $x>0$ be the lowest number such that there exists some subgame with depth x. Consider a subgame Γ^s that has depth x. For each $i \in N$ define the correspondence β_i^s: $A(s,\alpha^0) \rightarrow A_i(s,\alpha^0)$ by $\beta_i^s(a)=\text{argmax}_{a_i' \in A_i(s,\alpha^0)} U_i(a_i',(a_{ij}(s), (a_{ijk}(s))_{k\neq j})_{j\neq i}))$. Note two things (iii) and (iv): (iii) By standard arguments one sees that the combined correspondence β^s: $A(s,\alpha^0) \rightarrow A(s,\alpha^0)$ defined by $\beta^s(a)=(\beta_i^s(a))_{i\in N}$ admits a fixpoint. (iv) By inspection of β_i^s one sees that if $a_i' \in \beta_i^s(a)$ and if $a_i'' \in A_i$ prescribes the same choices as a_i' at all information sets inside Γ^s, then it must hold that $a_i'' \in \beta_i^s(a)$. Combining (iii) and (iv) one infers that there exist some profile a such that for *any* subgame Γ^s that has a depth of x it holds that $a \in \beta^s(a)$. Pick any such profile and call it $\alpha^x=(\alpha_i^x)_{i\in N}$

If K itself has depth x, set $\alpha=\alpha^x$. Else, proceed backwards through the tree in an analogous fashion (let y be the lowest number such that there exists some subgame that has depth $y>x$; consider a subgame that has depth y; et cetera) until Γ itself is considered and it is inferred that there exist some profile $a \in A$ such that $a \in \beta_i^t(a)$, where t is the root of Γ. Pick any such profile and call it α.

It is easy to see that α so defined must be a SRE. This is because in the above construction of α, the choices assigned to information sets in the r-part of any subgame Γ^r were selected so as to maximize the respective players' utilities given the choices assigned to information sets succeeding the r-part of Γ^r, and what choices are assigned to information sets preceeding the r-part of Γ^r is irrelevant for the determination of optimal choices in the r-part of Γ^r. ■

[23] Since A_i is non-empty (recall that if i owns no information set then A_i is taken to be singleton) and compact and U_i is continuous (as π_i, κ_{ij}, and λ_{iji} are all continuous) β_i^r is non-empty, closed-valued, and upper-hemi continuous by Berge's maximum principle. Since A_i is convex and U_i is quasi-concave (in fact linear) in i's own strategy, β_i^r must furthermore be convex-valued. Hence Kakutani's fixed point theorem can be applied to β^r.

Remark. Analogous theorems go through also with alternative definitions of the utilities, as long as these are quasiconcave in each player's own strategy. For example, this is the case with the formulation mentioned in Remark of Section 2 where each player *i*'s reciprocity payoff with respect to each player *j* is Y_i times the square root of the absolute value of $\kappa_{ij}(\cdot)\cdot\lambda_{iji}(\cdot)$, adjusted so as to maintain the right sign. Then each players reciprocity payoff is concave in the his strategy, and so is the utility. Another possibility is to leave the appearance of Definition 3 intact, but to modify Definitions 1 and 2 and redefine κ_{ij} and λ_{iji} as follows:

$$\kappa_{ij}(a_i,(b_{ij})_{j\neq i}) = [\pi_j(a_i,(b_{ij})_{j\neq i}) - \pi_j^{e_i}((b_{ij})_{j\neq i})] / [Q + max\{\pi_j(a_i,(b_{ij})_{j\neq i})|\, a_i\in A_i\} - min\{\pi_j(a_i,(b_{ij})_{j\neq i})|\, a_i\in A_i\}]$$

$$\lambda_{iji}(b_{ij},(c_{ijk})_{k\neq j}) = [\pi_i(b_{ij},(c_{ijk})_{k\neq j}) - \pi_i^{e_j}((c_{ijk})_{k\neq j})] / [Q+max\{\pi_i(b_{ij},(c_{ijk})_{k\neq j})|b_{ij}\in B_{ij}\} - min\{\pi_i(b_{ij},(c_{ijk})_{k\neq j})|b_{ij}\in B_{ij}\}]$$

where *Q* is a positive constant. This specification is similar to Rabin's (1993) formulation. We discuss this further in the next section.

5. COMPARISON WITH RABIN (1993)

Rabin (1993) develops a theory of reciprocity for normal form games with two players. In most two-player extensive games that lack proper subgames our model generates predictions that are similar to those that would obtain if Rabin's (1993) model was applied to the normal form of any such extensive game. Hence, if we rewrite the normal form games analyzed in Rabin's paper as simultaneous move extensive games we get qualitatively similar conclusions as he does in almost all cases. This indicates that the main difference between our model and that of Rabin (1993) is the requirement of sequential reciprocity we impose in games with an interesting dynamic structure. Yet the two models are different also in some

other ways. In this section we review these differences and attempt to justify our modeling choices.

If we were to make the following three changes to our model, then if we applied it to any two-player extensive game that has no proper subgames we would get *exactly* the same solutions as does Rabin in the normal form of that game:

Change (i). Substitute $(1+\kappa_{ij}(a_i,(b_{ij})_{j\neq i}))$ for $\kappa_{ij}(a_i,(b_{ij})_{j\neq i})$ in Definition 3.

Change (ii). Use the κ_{ij} and λ_{iji} functions discussed in the Remark of Section 4, except that $Q=0$ and that these function take a zero value if the respective denominators are zero.

Change (iii). Redefine the notion of an efficient strategy such that $a_i \in A_i$ is an *efficient strategy given beliefs* $(b_{ij})_{j\neq i}$ if there exists no $a_i' \in A_i$ such that for all $r \in R$, and $k \in N$ it holds that $\pi_k(a_i'(r),(b_{ij}(r))_{j\neq i}) \geq \pi_k(a_i(r),(b_{ij}(r))_{j\neq i})$, with strict inequality for some (r, k).

Change (i) incorporates an additional motivational element which Rabin (1993, p. 1287) argues is realistic. However, for the sake of simplicity we avoid it. In principle Change (i) can be applied to our model without adverse consequences, and we will not discuss this any further here. Change (ii) represents a kind of normalization of the players kindness such that the reciprocity payoff will have zero dimension. By contrast we measure kindness in the same unit as the material payoffs (for example dollars). Change (iii) makes the definition of an efficient strategy dependent on what is a players belief, whereas according to our definition efficiency is a belief-independent property.

Changes (ii) and (iii) are problematic in the context of general extensive games. To argue this point, consider the following example Γ_4:

(Insert Figure 4)

Assume first that only Change (ii) is made in our theory. Suppose that in equilibrium it holds that $a_2=b_{12}=p\cdot f+(1-p)\cdot d$ with $p<1$. A direct calculation involving the relevant function

outlined in Change (ii) shows that 1's kindness is $(2\cdot(1-p)-½\cdot(2\cdot(1-p)+0))/(2\cdot(1-p)-0)=½$. If Y_2 is high enough, 2 must choose *f*, which is a contradiction. Suppose instead that $a_2=b_{12}=f$. Player 1's kindness is now zero, so 2 must choose *d*. Again this is a contradiction. This proves that invoking (ii) in our theory would would preclude an existence theorem like that in Section 4.

Note that the culprit here is the discontinuity exhibited by player 1's kindness function as $p\rightarrow 1$. In fact, for all values of $p<1$, given change (ii), 1's kindness is *constant* (=½). We find this feature unreasonable, since the higher is *p* the more likely 1 figures it to be that 2 chooses *f* (since in equilibrium $b_{12}=p\cdot f+(1-p)\cdot d$), and the less material payoff 1 then believes that he gives to 2. We find it natural that 1's kindness in equilibrium is *decreasing* in *p*, as is the case in our theory (with $b_{12}=p\cdot f+(1-p)\cdot d$, 1's kindness if he chooses *F* is $(1-p)$).

Also Change (iii) would lead to existence problems. To see this, consider again Γ_4, and assume that only Change (iii) is made in our theory. Suppose that in equilibrium it holds that $a_2=b_{12}=pf+(1-p)d$ with $p\geq ½$. Given Change (iii), *only F* is an efficient strategy for 1.[24] Hence 1 is not kind when choosing *F*. But then 2 chooses *d*, which is a contradiction. Suppose instead that $a_2=b_{12}=pf+(1-p)d$ with $p<½$. Then *all* 1's strategies are efficient. Player 1 is now kind choosing *F*, and since $p<½$ his kindness is bounded away from zero. If Y_2 is high enough 2 must choose *f*, which again is a contradiction. This proves that invoking (iii) in our theory leads to non-existence of equilibria in some games. Note that in our theory this problem does not arise because efficiency of a strategy is a belief-independent property. According to our definition, in Γ_4 there are no inefficient strategies regardless of b_{12}.

A final difference between our model and Rabin's is that our theory applies also to games with more than two players. We close this section by analyzing Γ_5, a modified version of a three-player game which Nalebuff & Stiglitz (1988) use to discuss certain aspects of vengefulness.

---

[24] Rabin (1993) does not give an explicit definition of an efficient strategy. Our argument here presumes a definition corresponding to Change (iii), so that a strategy is efficient if no other strategy is at least as no worse for any player, and better for some player. Alternatively a strategy may be defined as efficient if no other strategy is strictly better for all players. It is easy to verify that also then can Γ_4 be used to illustrate non-existence.

(Insert Figure 5 here)

With $\varepsilon=0$, Γ_5 may be thought of as a model of a strategic situation in which a \$4-pie is to be divided between three players. First player 1 has to choose which one of the other two players must get a zero payoff. Then the player who was "unfavorably" treated by 1 is called upon to decide which one of the other two will get which of the two positive monetary payoffs. Intuition may suggest that player 1 is a priori worst off of the three. Whoever he treats unfavourably will feel badly treated, and hence take revenge on 1 by awarding him the lowest possible monetary payoff. Effectively, player 1 will get a payoff of one, while players 2 and 3 look at expected payoffs of 1.5.

If each player was motivated solely by his or her own monetary income this outcome would not be guaranteed (in subgame perfect equilibrium), as players 2 and 3 would be indifferent between all their choices. In order to accommodate revenge, Nalebuff & Stiglitz append the usual selfishness assumption, and assume that the players have lexicographically ordered objectives. Each player primarily maximizes his monetary rewards, but in case many choices yield *exactly the same* monetary payoff ties are broken so as to allow a player to take revenge. In Γ_5, this works to 1's disadvantage.

Our model of sequential reciprocity allows a similar conclusion, evoking also certain emotions on behalf of player 1. This is true also when 2 and 3 incur some monetary cost $\varepsilon>0$ if they "punish" player 1. For *any* $\varepsilon\geq 0$, at 2's decision node 2 believes that 1 is unkind to 2 ($\lambda_{121}(\cdot)<0$), and that 3 is neither kind nor unkind to 2 ($\lambda_{323}(\cdot)=0$). Player 2 can get a positive reciprocity payoff only by choosing r_2, since $\kappa_{21}(r_2,\cdot)<0<\kappa_{21}(l_2,\cdot)$. For large enough Y_2 player 2 will choose r_2 as her material cost is swamped by the sweetness of revenge.

Analogous remarks apply at player 3's node, so in any sequential reciprocity it is true that if Y_2, Y_2 are high enough, then $a_2=r_2$ and $a_3=r_3$. Yet, there are multiple equilibria which are characterized by "self-fulfilling prophecies" much like in the examples of Section 3. Both the pure strategy profiles (L,r_2,r_3) and (R,r_2,r_3) are equilibria (and with $Y_2>0$ the only remaining

equilibrium is ($\frac{1}{2}\cdot L+\frac{1}{2}\cdot R, r_2, r_3$)). The following calculations for player 1 confirm this for (L, r_2, r_3):

$\kappa_{12}(L, (r_2, r_3)) = \kappa_{13}(R, (r_2, r_3)) = -1.5$;

$\kappa_{13}(L, (r_2, r_3)) = \kappa_{12}(R, (r_2, r_3)) = 1.5$;

$\lambda_{121}(r_2, (L, r_3)) = -1$; $\lambda_{131}(r_3, (L, r_2)) = 0$

Hence, it holds that

$u_1(L, (r_2, r_3)) = 1+Y1\cdot[(-1.5)\cdot(-1)+ (1.5)\cdot 0)] < 1+Y1\cdot[(1.5)\cdot(-1)+ (-1.5)\cdot 0)] = u_1(R, (r_2, r_3))$,

which shows that (L, r_2, r_3) is indeed a SRE.

6. CONCLUDING REMARKS

In this paper we focus on modeling a concern for reciprocity, and disregard distributional concerns like altruism, equity, and envy. As noted in the Introduction, it is clear that this omission is not innocuous. For example, in experimental Dictator games individuals often give away lots of money (see Davis & Holt 1993, pp 263-269 for a discussion), something which cannot be explained by the model we propose in this paper. In reality people seem to be motivated in many different ways, and perhaps this all depends not only on the strategic nature of a situation but also on other aspects of the context where the situation occurs. For example, in the case of Dictator games the evidence reported by Hoffman, McCabe & Smith (1996) suggests that "social distance" is important in that context. We leave for future research the delicate tasks of integrating multiple concerns in one unified model, and of determining percisely in what context one or another motivational concern is of particular importance. However, it seems clear that when these issues are tackled, experimental and theoretical work should go hand in hand.

REFERENCES

Akerlof, George A. (1982), "Labour Contracts as a Partial Gift Exchange", *Quarterly Journal of Economics* 97, 543-569.

Akerlof, George A. and Yellen, Janet L. (1988), "Fairness and Unemployment", *American Economic Review* 78, 44-49.

Akerlof, George A. and Yellen, Janet L. (1990), "The Fair-Wage Effort Hypothesis and Unemployment", *Quarterly Journal of Economics* 195, 255-284.

Berg, Joyce, Dickhaut, John and McCabe, Kevin (1995), "Trust, Reciprocity and Social History", *Games and Economic Behavior* 10, 122-142.

Bewley, Truman (1995), "A Depressed Labor Market as Explained by Participants", *American Economic Review* 85, 250-259.

Bolton, Gary E., Brandts, Jordi, and Katok, Elena (1996), "A Simple Test of Explanations for Contributions in Dilemma Games", mimeo.

Bolton, Gary E, and Ockenfels, Axel (1997) "A Theory of Equity, Reciprocity and Competition", mimeo.

Charness, Gary (1996), "Attribution and Reciprocity in a Simulated Labor Market: An Experimental Investigation", mimeo.

van Damme, E. (1991), *Stability and Perfection of Nash Equilibria*, Springer-Verlag.

Davis, Douglas D. and Holt, Charles A. (1993), *Experimental Economics*, Princeton University Press.

Falk, Armin, Gächter, Simon, and Kovacs, Judith (1997), "Reputation and Reciprocity", mimeo.

Fehr, Ernst, Gächter, Simon and Kirchsteiger, Georg (1996), "Reciprocal Fairness and Noncompensating Wage Differentials", *Journal of Institutional and Theoretical Economics* 152(4), 608-640.

Fehr, Ernst, Gächter, Simon, and Kirchsteiger, Georg (1997), "Reciprocity as a Contract Enforcement Device: Experimental Evidence", *Econometrica* 65(4), 833-860.

Fehr, Ernst and Kirchsteiger, Georg (1994), "Insider Power, Wage Discrimination and Fairness", *Economic Journal* 104, 571-583.

Fehr, Ernst, Kirchsteiger, Georg, and Riedl, Arno (1993), "Does Fairness Prevent Market Clearing? An Experimental Investigation", *Quarterly Journal of Economics* 108(2), 437-460.

Fehr, Ernst, Kirchsteiger, Georg and Riedl, Arno (1998), "Gift Exchange and Reciprocity in Competitive Experimental Markets", *European Economic Review* 42(1), 1-34.

Fehr, Ernst, and Schmidt, Klaus (1997), "A Theory of Fairness, Competition, and Cooperation", mimeo.

Geanakoplos, John, Pearce, David and Stacchetti, Ennio (1989), "Psychological Games and Sequential Rationality", *Games and Economic Behavior* 1, 60-79.

Goranson, Richard E. and Berkowitz, Leonard (1966), "Reciprocity and Responsibility Reactions to Prior Help", *Journal of Personality and Social Psychology* 3(2), 227-232.

Greenberg, Martin S. and Frisch, David M. (1972), "Effect of Intentionality on Willingness to Reciprocate a Favor", *Journal of Experimental Social Psychology* 8, 99-111.

Hoffman, Elizabeth, McCabe, Kevin and Smith, Vernon L. (1996), “Social Distance and Other-Regarding Behavior in Dictator Games”, *American Economic Review* 86(3), 653-660.

Kahneman, Daniel, Knetsch, Jack L. and Thaler, Richard H. (1986), "Fairness as a Constraint on Profit Seeking: Entitlements in the Market", *American Economic Review* 76 (4), 728-741.

Kirchsteiger, Georg (1994), "The Role of Envy in Ultimatum Games", *Journal of Economic Behavior and Organisation* 25(3), 373-390.

Komter, Aafke E. (eds) (1996), *The Gift: An Interdiciplinary Approach*, Amsterdam University Press, Amsterdam.

Levine, David K. (1997), "Modeling Altruism and Spitefulness in Experiments", mimeo.

Nalebuff, B. and Shubik, M. (1988), "Revenge and Rational Play", Woodrow Wilson School, Discussion paper #138.

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

Roth, Alvin E. (1995), "Bargaining Experiments", in: John Kagel and Alvin E. Roth (eds), *Handbook of Experimental Economics*, Princeton University Press.

Sen, Amartya K. (1979), "Utilitarianism and Welfarism", *Journal of Philosophy* 76, 463-489.

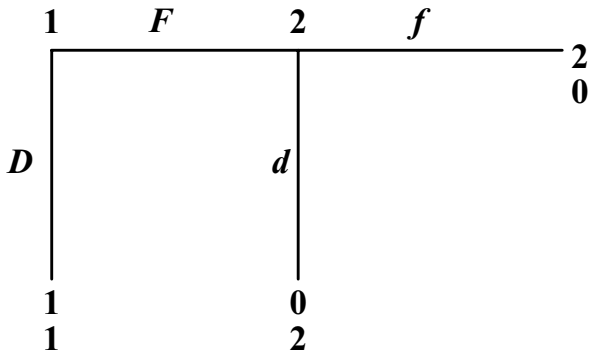

Figure 1: Game Γ_1

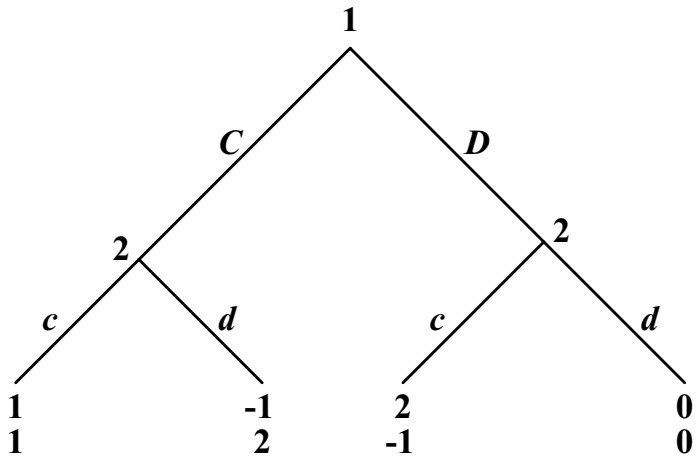

Figure 2: Game Γ_2—The Sequential Prisoners' Dilemma

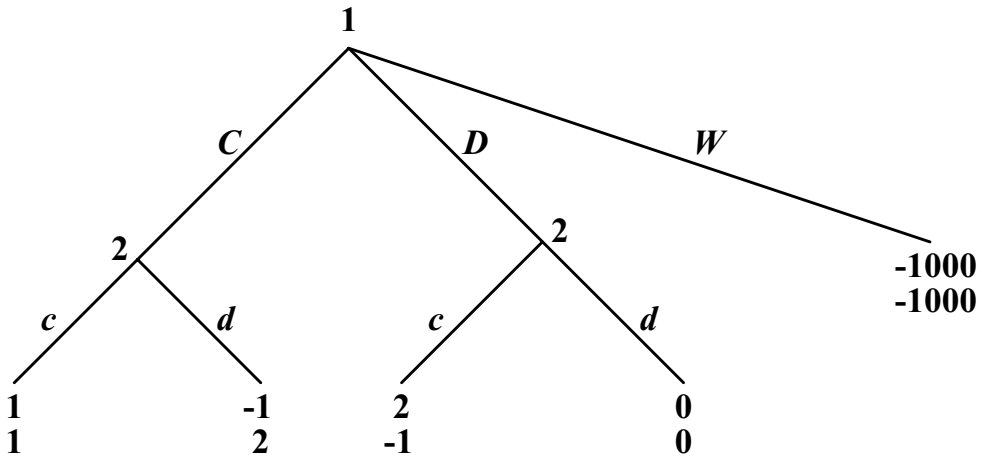

Figure 3: Game Γ_3

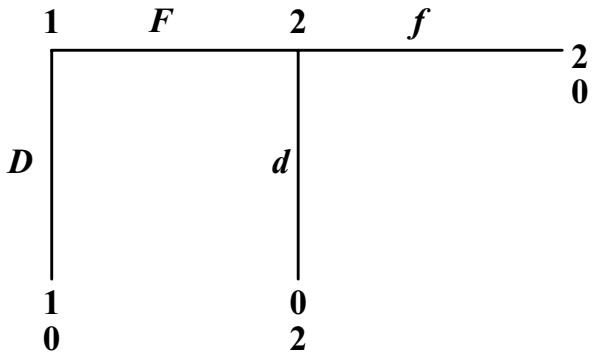

Figure 4: Game Γ_4

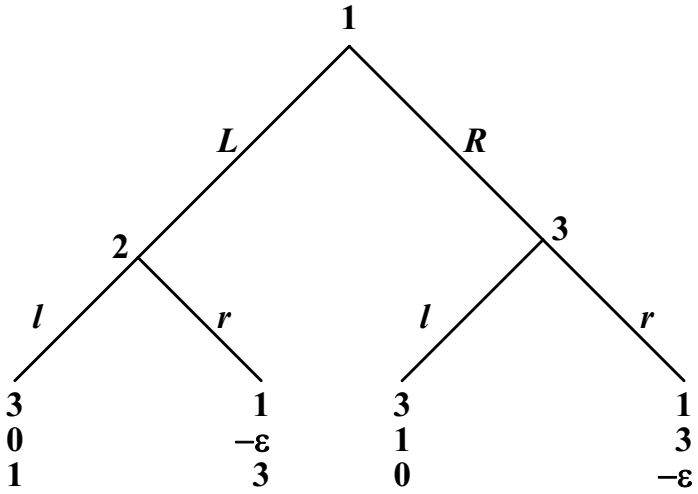

Figure 5: Game Γ_5