Reciprocity and Psychological Games

Pierpaolo Battigalli (Bocconi University, Milan)

Verbania, conference on "Reciprocity: Theory and Facts"
February 23, 2007

Abstract

I discuss how "intention-based" theories of intrinsic reciprocity can be expressed and analyzed using dynamic games with belief-dependent utilities. First, I propose relatively simple formulas to capture reciprocity motivations. Second, I argue that modeling plans of actions as beliefs about one's own contingent choices is reasonable, but if coupled with trembling-hand equilibrium ideas prevents to capture sequential reciprocity. Third, I consider different ways to weigh how kind a co-player would be at each node.

'Social preferences' help explain observed (in the lab.) behavior in Dictator, Ultimatum, Trust, Gift Exchange, Public Good and similar games. Two (non mutually exclusive) types of motivations inducing social preferences:

1. material payoffs of others matter $\longrightarrow$ distribution-dependent preferences (inequity aversion, status),

2. intentions matter $\longrightarrow$ belief-dependent motivations (some forms of reciprocity and guilt, social norms(?)).

Traditional game theory (GT) can address 1 (distribution-dependent preferences).

An extension of traditional GT, called "psychological GT", has been used to address 2 (e.g. Rabin [5], Dufwenberg&Kierchsteiger [2], Falk&Fischbacher [3], see also related paper by Segal&Sobel [6]).

I will focus on 2, in particular on reciprocity as a belief-dependent motivation.

Loosely speaking: in a *psychological game* utility functions depend on (actions and) *beliefs*, including beliefs about the beliefs of others. New framework put forward by Geanakoplos,Pearce&Stacchetti [4], then substantially expanded and refined by Battigalli&Dufwenberg [1] (BD) to deal with sequential games.

BD is needed because many (perhaps most) interesting examples where reciprocity can be plausibly assumed to play a role have a *sequential* structure, e.g. Ultimatum, Trust, and Gift Exchange games. Thus *revised beliefs about the beliefs of others* can play a role.

Here I want to discuss how such extension of GT can be used to model (intrinsic) reciprocity concerns relying on the idea that "intentions matter".

I try to capture some of the intuitions appeared in the literature (see papers mentioned above), but I differ on the details, and I discuss some issues I find problematic.

Following BD, I will also argue that it is plausible and analytically convenient to allow a player's utility function to depend on the *beliefs of others*.

Example 1: Distribution game

$$
\begin{tikzcd}
(\text{Ann}) \arrow[r, "Trust"] \arrow[d, "Don't"'] & (\text{Bob}) \arrow[r, "Give"] \arrow[d, "Take"'] & \begin{pmatrix} 2 \\ 0 \end{pmatrix} \\
\begin{pmatrix} 1 \\ 1 \end{pmatrix} & \begin{pmatrix} 0 \\ 2 \end{pmatrix} &
\end{tikzcd}
$$

Figure 1**.** Distribution Game with material payoffs.

Is *Trust* a "kind" action? It depends on intentions (Rabin, [5]):

Does Bob think that Ann intended to get $2, leaving $0 to him? Intention of Ann depends on her belief. Perception of intention by Bob depends on Bob's belief about Ann's belief.

$\alpha = \text{Pr}_{Ann}[Give \text{ if } Trust]$ initial 1st-order belief of Ann

$\beta = E_{Bob}[\alpha | Trust]$ conditional 2nd-order belief of Bob

Trust may be deemed "kind" if α is low.

Trust is perceived as kind by Bob if β is low.

If β is low and Bob is highly motivated by reciprocity considerations, he reciprocates and *Give*s.

Beliefs (about beliefs) affect preferences over consequences (material payoff distributions).

METHODOLOGY:

1) Start with a "material-payoff game" (or "game form"): specification of rules of interaction and of how distributions of monetary payoffs depend on (sequences of) actions.

2) Define the "kindness" of player i as a function of his "plan of action" and beliefs: i is kind (unkind) to j if he intends to make j get more (less) money than a context-dependent "equitable payoff" $\pi_j^{e_i}$.

3) Define "psychological utility functions" capturing the assumption that i is willing to sacrifice some of his monetary payoff to reciprocate the (perceived) (un-)kindness of j toward him.

4) Apply a solution concept to obtain behavioral implications [more generally, derive implications about behavior from assumptions about rationality and interactive beliefs]

The **"traditional" formulas** (cf. Rabin [5], Dufwenberg&Kirchsteiger [2])**:**

π_i=*material* (monetary) *payoff* of i, depends on actions

u_i=*psychological utility* of i

$\pi_j^{e_i}(\text{belief}_i)$="*equitable payoff*" ascribed by i to j, given i's belief; for example:

$$\pi_j^{e_i}(\text{belief}_i) = \frac{1}{2}\left(\max_{s_i} \mathsf{E}[\pi_j; \text{belief}_i, s_i] + \min_{s_i} \mathsf{E}[\pi_j; \text{belief}_i, s_i]\right)$$

$K_i(\text{belief}_i, \text{plan}_i)$=*kindness* of i toward j, it depends on i's intentions:

$$K_i(\text{belief}_i, \text{plan}_i) = \mathsf{E}[\pi_j; \text{belief}_i, \text{plan}_i] - \pi_j^{e_i}(\text{belief}_i)$$

θ_i=i's *sensitivity* to reciprocity, a *parameter*

psychological utility of player i:

$$\begin{aligned} u_i &= \pi_i + \theta_i \times \mathsf{E}_i[K_j] \times K_i = \pi_i + \theta_i \times \mathsf{E}_i[K_j] \times \left(\mathsf{E}_i[\pi_j] - \pi_j^e\right) \\ &= \pi_i + \theta_i \times \mathsf{E}[K_j; \text{belief}_i] \times \mathsf{E}[\pi_j; \text{belief}_i, \text{plan}_i] + f(\text{belief}_i) \end{aligned}$$

If $\mathsf{E}_i[K_j] > 0$ ($\mathsf{E}_i[K_j] < 0$), then i is willing to sacrifice some π_i to increase $\mathsf{E}_i[\pi_j]$ above the endogenous threshold π_j^e (decrease $\mathsf{E}_i[\pi_j]$ below π_j^e).

NOTE: Since K_j depends on the 1st-order belief of j (his belief about s_i), then $\mathsf{E}_i[K_j]$ depends on the 2nd-order belief of i (belief of i about belief of j)

I am intentionally ambiguous about exact functional forms. Many choices about details are arbitrary. Other things equal, I favour tractability and simplicity.

I propose to work with variations of the simpler formula

$$u_i = \pi_i + \theta_i \times \mathsf{E}_i[K_j] \times \pi_j$$

Back to Example 1

$$
\begin{tikzcd}
(\text{Ann}) \arrow[r, "Trust"] \arrow[d, "Don't"'] & (\text{Bob}) \arrow[r, "Give"] \arrow[d, "Take"'] & \begin{pmatrix} 2 \\ 0 \end{pmatrix} \\
\begin{pmatrix} 1 \\ 1 \end{pmatrix} & \begin{pmatrix} 0 \\ 2 \end{pmatrix} &
\end{tikzcd}
$$

Figure 1. Distribution Game with material payoffs.

Let $\theta_A = 0$, that is $u_A = \pi_A$ (Ann does not care for reciprocity).

Recall: $\alpha = \mathsf{Pr}_A[Give \text{ if } Trust]$ and $\beta = \mathsf{E}_B[\alpha | Trust]$

After some algebra:

$$K_A(Trust,\text{belief}_A) = 2(1-\alpha) - \left(\frac{3}{2} - \alpha\right) = \frac{1}{2} - \alpha$$

$$\mathsf{E}_B[K_A | Trust;\text{belief}_B] = \frac{1}{2} - \beta,$$

$$u_B(Trust, Give;\text{belief}_B) =$$

$$= \pi_B(Trust, Give) + \theta_B \times \left(\frac{1}{2} - \beta\right) \times \pi_A(Trust, Give) = 0 + 2\theta_B \left(\frac{1}{2} - \beta\right)$$

We obtain

$$\begin{array}{ccccc} & & Trust & & & Give & \\ & (\text{Ann}) & \longrightarrow & & (\text{Bob}) & \longrightarrow & \begin{pmatrix} 2 \\ 2\theta_B\left(\frac{1}{2}-\beta\right) \end{pmatrix} \\ Don't & \Big\downarrow & & Take & \Big\downarrow & & \\ & \begin{pmatrix} 1 \\ 1+* \end{pmatrix} & & & \begin{pmatrix} 0 \\ 2 \end{pmatrix} & & \end{array}$$

Figure 1p**.** Distribution Game with psychological utility.

(if $\theta_B < 2$, "standard" equilibrium $(Don't, Take)$, if $\theta_B > 2$ no pure equil.)

ISSUES:

- In games, the utility of player i depends on many unknowns. We might as well make psy-utility directly depend on the unknown kindness $\Rightarrow$ Beliefs$_j$ in u_i?

- In sequential games i only controls his action at current node, his *plan* comprises his current action and what he *believes* he would do later. Should we model *plans as beliefs about one's own strategy*? What does this imply for a sensible notion of sequential equilibrium?

- In sequential games K_i depends on node/subgame h *via* updating of beliefs$_i$ and options still open at h; how should we take this into account? How do players react to their perception of the co-player kindness at different nodes? Should we consider the perception of some "global/average kindness", or perception of "ex ante/initial kindness", or maybe "kindness on the path"?

BELIEFS OF OTHERS IN THE UTILITY FUNCTION

Psychological payoff function:

$$u_i = \pi_i + \theta_i \times K_j(\text{belief}_j,\text{plan}_j) \times \pi_j$$

Simpler functional form (first-order beliefs only) that yields the same best response correspondence as the previous one (cf BD [1]). Example:

$$\begin{tikzcd}[column sep=huge, row sep=large]
\text{(Ann)} \arrow[r, "Trust"] \arrow[d, "Don't"'] & \text{(Bob)} \arrow[r, "Give"] \arrow[d, "Take"'] & \begin{pmatrix} 2 \\ 2\theta_B\left(\frac{1}{2}-\alpha\right) \end{pmatrix} \\
\begin{pmatrix} 1 \\ * \end{pmatrix} & \begin{pmatrix} 0 \\ 2 \end{pmatrix} &
\end{tikzcd}$$

Figure 1p'. Distribution Game where utility depends on co-player belief.

PLANS OF ACTION AS BELIEFS ABOUT ONESELF?

Let plan$_i$=*belief* of i about how he would choose at each node

Sequential Equilibrium: for each pl. i and node/history h, max$\mathsf{E}_i[u_i|h]$+beliefs of all order are correct+deviations interpreted as "mistakes" (trembling hand)

Consequence: the sequentiality of reciprocity is lost, if i initially believes that j is unkind (or neutral), he would never change his mind, as deviations would be regarded as unintended mistakes.

Example 2: Trust Game

$$\begin{tikzcd} \text{Ann} \arrow[r, "Trust"] \arrow[d, "Don't"'] & \text{Bob} \arrow[r, "Share"] \arrow[d, "Grab"] & \begin{pmatrix}2\\2\end{pmatrix} \\ \begin{pmatrix}1\\1\end{pmatrix} & \begin{pmatrix}0\\4\end{pmatrix} & \end{tikzcd}$$

Figure 2. Trust Game with material payoffs.

$K_A(Don't, \alpha) = -\left(\frac{3}{2} - \alpha\right) < 0$, $K_A(Trust, \alpha) = \frac{3}{2} - \alpha$

Candidate equil.: $(Don't, Grab, \alpha = 0, \beta = 0)$.
Bob's perception of Ann's kindness is $K_A(Don't, \alpha) < 0$ *whatever Ann does*
$\Rightarrow$Bob $Grab$s.
This is an equilibrium for any reciprocity sensitivity θ_B.

Suppose instead: plan$_j$ =actual strategy$_j$
$\Rightarrow$even if i never changes his beliefs about *belief* $_i$ (sequential eq. assumption), yet i may be forced to change his beliefs about *intention*$_i$=(belief$_i$+actual strategy$_i$). In Trust Game:

$$
\begin{array}{ccccc}
 & Trust & & Share & \\
\text{Ann} & \longrightarrow & \text{Bob} & \longrightarrow & \begin{pmatrix} 2 \\ 2+2\theta_B(\frac{3}{2}-\alpha) \end{pmatrix} \\
| & & | & & \\
Don't \quad \downarrow & & \downarrow \quad Grab & & \\
\begin{pmatrix} 1 \\ * \end{pmatrix} & & \begin{pmatrix} 0 \\ 4 \end{pmatrix} & &
\end{array}
$$

Figure 2p. Trust Game with "meaningful" psy-utility.

For $\theta_B > \frac{2}{3}$, $(Don't, Grab, \alpha = \beta = 0)$ cannot be an equilibrium.
$\Rightarrow$ either let plan=actual strategy, or *modify equil. concept* (e.g. forward ind.).

EX ANTE *vs* GLOBAL *vs* PATH-KINDNESS

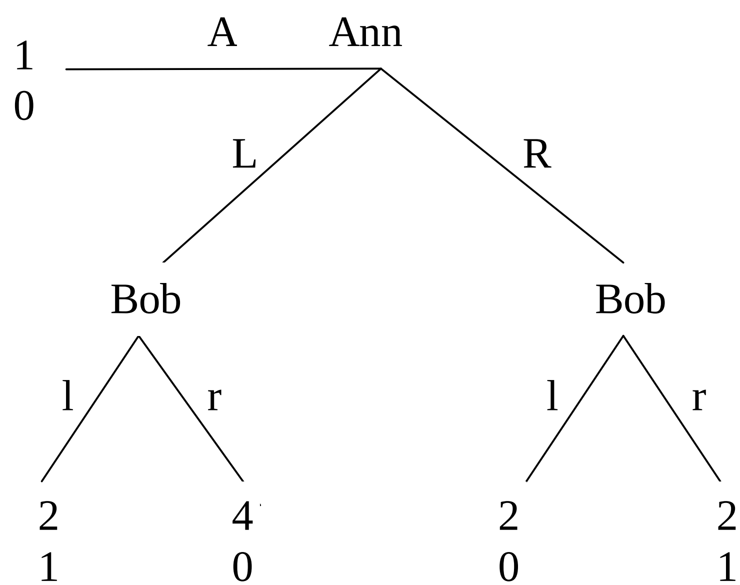

Figure 3. A material payoff game

Suppose Bob is known to be non-reciprocal ($\theta_B = 0$), whereas $\theta_A > 0$. Can *A*cross be a reciprocity equilibrium outcome?

NOTE: $\theta_B = 0 \Rightarrow$in eq., Bob plays *lr*$=$(*l* if L,*r* if *R*) and this is common belief

Kindness depends on node/history:
If Bob is initially certain of *A*cross $K_{B,root} = 0 \Rightarrow$If only the *"ex ante, or initial kindness"* $K_{B,root}$ matters for Ann, *A*cross cannot be chosen in equilibrium. Is this assumption reasonable? Should Ann regard Bob as neutral only because she does not allow him to play, even if she is certain that Bob would be "mean" if given the opportunity to play?

Kindness when Bob has to play (note: conditional beliefs of Bob about Ann's strategy are pinned down by the observed choice of Ann):

$K_{B,L}(lr) = 2 - (2+4)/2 = -1$, $K_{B,R}(lr) = 2 - (2+2)/2 = 0$

Define "*global, or average kindness*" of Bob: $\mathbf{K}_B(lr) = \frac{1}{2}K_{B,L}(lr) + \frac{1}{2}K_{B,R}(lr) = -\frac{1}{2}$. Assume

$$u_A = \pi_A + \theta_A \times \mathbf{K}_B \times \pi_B$$

We obtain: ...

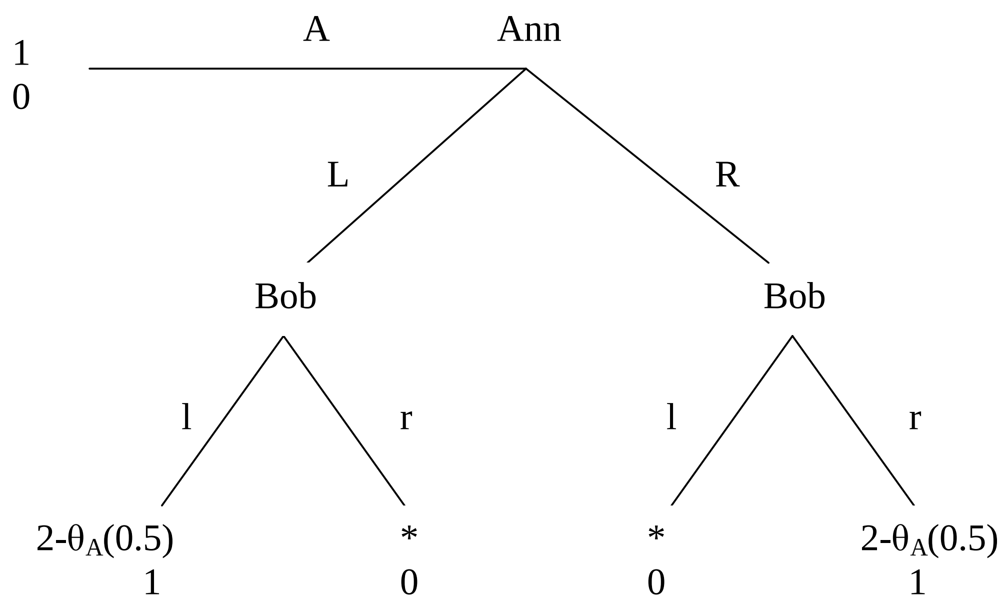

Figure 3p. Psychological game with reciprocity and "global kindness".

If $\theta_A > 2$ then *A*cross is a reciprocity equil. outcome: Ann gives up 1$ to preemptively punish Bob.

According to "*path-kindness*", Ann anticipates her final positive/negative feelings due to Bob's kind/unkind behavior for each of her possible moves:

$$u_A(a_A, a_B) = \pi_A(a_A, a_B) + \theta_A K_{a_A}(a_B).$$

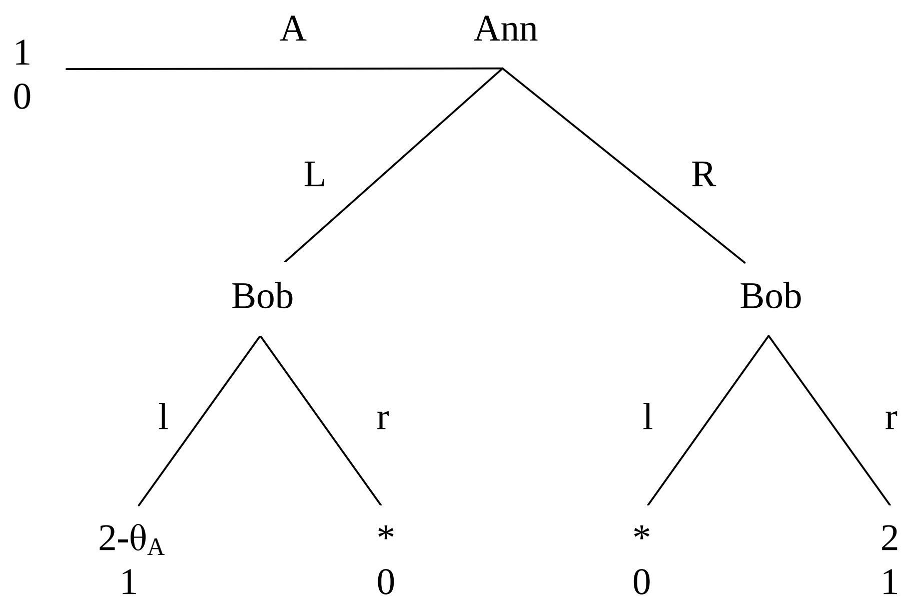

Figure 3p'. Psychological game with reciprocity and "path-kindness"

Clearly *A*cross cannot be a reciprocity equilibrium outcome with path kindness. Ann anticipates that she would feel hurt by Bob's unkindness if she chooses *L*eft, and she would not feel hurt if she chooses *R*ight. The unique equilibrium is (R, lr).

CONCLUSIONS

The idea that reciprocity is related to (perceived) intentions, and hence should be modelled *via* belief-dependent motivations is intuitively compelling, but hard to formalize: the devil is in the details.

All the existing models based on psychological game theory are very complex, and rely on some arbitrary modelling choices. I discussed three themes:

(1) "Technical tricks" (e.g. beliefs of others in the utility function) allow somewhat simpler reciprocity formulas.

(2) Intention$\Leftrightarrow$plan$+$belief. How should we model "plans"? How should we model updating of beliefs about intentions? Plan$=$"own-strategy-belief"$+$trembling-hand ideas on updating yield a completely uninteresting notion of sequential reciprocity. $\Rightarrow$Either let plan$=$actual strategy, or (better) replace trembling-hand ideas with solution concepts that allow for interesting updating about intentions of others.

(3) How should we weigh the co-player kindnesses at different histories? "Ex ante/initial kindness", "global/average kindness", or "path kindness"?

Hopefully, I provided a framework that allows to meaningfully discuss and elucidate different notions of intention-based reciprocity.

References

[1] BATTIGALLI, P. and M. DUFWENBERG (2005): "Dynamic Psychological Games", IGIER w.p. 287.

[2] DUFWENBERG, M. and G. KIRCHSTEIGER (2004): “A Theory of Sequential Reciprocity,” Games and Economic Behavior, **47**, 268-298.

[3] FALK, A. and FISCHBACHER (2006): "A Theory if Reciprocity", Games and Economic Behavior, **54**, 293-315.

[4] GEANAKOPLOS, J., D. PEARCE and E. STACCHETTI (1989): “Psychological Games and Sequential Rationality,” Games and Economic Behavior, **1**, 60-79.

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

[6] SEGAL, U. and J. SOBEL (2004,2006): "Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings", mimeo.