Journal of Economic Behavior and Organization 000 (2018) 1–9

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Reciprocity and the Paradox of Trust in psychological game theory☆

Andrea Isoni[a], Robert Sugden[b,*]

[a] Warwick Business School, Coventry CV4 7AL, UK; University of Cagliari, Italy
[b] School of Economics and Centre for Behavioural and Experimental Social Science, University of East Anglia, University Plain, Norwich NR4 7TJ, UK

ARTICLE INFO

Article history:
Received 15 January 2018
Accepted 19 April 2018
Available online xxx

JEL classifications:
C72 (noncooperative games)
D91 (role and effects of psychological, emotional, social, and cognitive factors on decision making)

Keywords:
Reciprocity
Paradox of Trust
Kindness
Cooperation
Psychological game theory
Mutual benefit

ABSTRACT

Rabin's psychological game-theoretic model of 'fairness' has been the starting point for a literature about preferences for reciprocity. In this literature, reciprocity is modelled by defining an individual's 'kindness' or 'unkindness' in terms of the consequences of his actions for others, and assuming a motivation to reward (punish) other people's kindness (unkindness). Contrary to intuition, this form of reciprocity cannot explain mutually beneficial trust and trustworthiness in a simple Trust Game. We formalise and offer a diagnosis of this 'Paradox of Trust'. We distinguish between two kinds of reciprocity. Rabin's concept of *reciprocal kindness* is a psychologically plausible motivation, and the paradox is an informative result about the implications of this motivation. However, trust is better understood in terms of *reciprocal cooperation* – the motivation to play one's part in mutually beneficial practices, conditional on others playing their parts. We show that a theory of reciprocal cooperation can avoid the paradox.

Matthew Rabin's model of preferences for 'fairness' (Rabin, 1993) is one of the earliest applications of psychological game theory. It has also been one of the most influential, providing the starting point for an important strand in the social preference literature. Rabin's model has two fundamental features which, in this literature, have come to be seen as characteristic of reciprocity. The first feature is a concept of *kindness*. In a two-player game, the degree to which one player i is kind or unkind to the other player j is assessed by taking i's beliefs about j's strategy as given and then considering the decision problem faced by i as if it were a non-strategic choice among the alternative distributions of material payoffs between the players that are feasible for i.[1] Since this is the kind of problem that is faced by the active player in a Dictator Game, we will call it a *dictator problem*.[2] Player i shows kindness (unkindness) towards j by choosing a distribution that is relatively

☆ An earlier version of this paper was presented at a workshop on behavioural game theory held at the University of East Anglia in 2017. We thank participants in this workshop, and particularly Martin Dufwenberg and Amrish Patel, for comments. Our work was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 670103.

* Corresponding author.
E-mail addresses: a.isoni@warwick.ac.uk (A. Isoni), r.sugden@uea.ac.uk (R. Sugden).

[1] Throughout the paper (except when discussing public goods in Section 4) we will consider only two-player games. In such games, whenever players are indexed by i and j, we assume $i \neq j$ and refer to i as 'he' and to j as 'she'.

[2] We apply the term 'Dictator Game' to any game in which one player chooses between alternative distributions of material payoffs between himself and a co-player. We do not require that the sum of the two payoffs is constant.

https://doi.org/10.1016/j.jebo.2018.04.015

favourable (unfavourable) towards her. The second feature is that each player has a preference for *rewarding* the other player for acting on kind intentions and *punishing* her for acting on unkind intentions. Because *i*'s kindness is defined in terms of his beliefs about *j*'s strategy, and because *j*'s preferences for rewarding or punishing *i*'s good or bad intentions are defined in terms of her beliefs about *i*'s kindness or unkindness, players' utilities depend on their first- and second-order beliefs: hence the need for psychological game theory. Models with these two properties will be called models of *reciprocal kindness*.

The present paper is concerned with a surprising property of Rabin's model, which we will call the *Paradox of Trust*. This is that, in a simple Trust Game, there cannot be an equilibrium in which trust by one player and trustworthiness by the other are mutually beneficial. This result is surprising because Rabin's model is generally understood as a model of reciprocity, and because the relationship between trustworthiness and trust is naturally described as one of reciprocity.[3] Rabin (1993: 1296–1297) hints at this sense of surprise when, discussing a variant of the paradox, he acknowledges that his model does not adequately represent the motivation to return trust. Our aim is to identify the source of this problem.

We must make clear that the problem is not that there are observable regularities in other-regarding behaviour that a model of reciprocity does not explain. Rabin (1993: 1296–1297) makes clear that his model is not intended to represent all the emotions that induce other-regarding behaviour: the intention is only to isolate one such emotion.[4] Nor is the problem that behavioural economics lacks any explanation of observations of mutually beneficial actions in Trust Games. For example, such actions are consistent with the hypothesis, proposed by the theories of *normative expectations* (Sugden, 1998), *trust responsiveness* (Pelligra, 2005; Bacharach et al., 2007) and *guilt aversion* (Battigalli and Dufwenberg, 2007), that individuals are motivated to behave in ways that confirm other people's expectations of benefit. But intending to confirm another person's expectations of your kindness to him is not the same thing as intending to reciprocate his kindness to you. The problem posed by the Paradox of Trust is to explain how mutually beneficial trust and trustworthiness can arise *from reciprocity*.

In Section 1, we characterise the Paradox of Trust in terms of a description of the behaviour and beliefs of two players in two related games. We argue that this description represents a psychologically credible scenario in which, in an intuitive sense, the players act on intentions for reciprocity. We show that this scenario is inconsistent with Rabin's model. In Section 2, we show that close analogues of this paradox also occur in the models of reciprocal kindness proposed by Charness and Rabin (2002) and Falk and Fischbacher (2006). We also consider the generalisation of Rabin's original model proposed by Dufwenberg and Kirchsteiger (2004). Dufwenberg and Kirchsteiger make an amendment to Rabin's model which eliminates the paradox. We argue that this amendment lacks a convincing psychological rationale and has implications that are not compatible with intuitive understandings of kindness and reciprocity. In Section 3, we present our diagnosis of the paradox. We argue that Rabin's conception of reciprocal kindness is a psychologically plausible motivation, and that the Paradox of Trust is an informative result about the collective consequences that follow when individuals are motivated in this way. However, we distinguish between this conception of reciprocity and *reciprocal cooperation* – playing one's part in mutually beneficial practices, conditional on others playing their parts. In Section 4, we give a sketch of one of the earliest behavioural models of reciprocal cooperation, that of Sugden (1984). The idea that individuals can sometimes be motivated by this form of reciprocity is psychologically plausible too, and can explain practices of mutually beneficial trust and trustworthiness. Section 5 concludes.

1. Rabin's model of reciprocity and the Paradox of Trust

As an experimental paradigm, the Trust Game originates in the work of Berg et al. (1995), but it has a much longer history as a theoretical model. Hobbes (1651/ 1962: 110–115) discusses an example in which a prisoner of war can be released in return for a promise to pay a ransom on return to his home country. The captor performs first, trusting the captive to perform second. Hobbes argues that, because of the value of reputation, it is rational for the captive to pay the ransom. A much more recent precursor of Berg et al.'s experiment is the core of Akerlof's (1982) model of labour contracts as 'partial gift exchange'. In this model, an employer pays a worker more than the worker's reservation wage; the worker responds by supplying costly effort when effort cannot be monitored.

Our discussion will focus on the specific Trust Game shown in Fig. 1 and denoted by G_1. This game involves two players, P1 (the first mover) and P2 (the second mover).

Payoffs are expressed in units of some material good that is valued by both players, measured relative to some reference point. For compactness, we will write the Trust Game as $G_1 = \{_1\ (0, 0), \{_2\ (-1, 3), (1, 1)\}\}$. The inner pair of brackets $\{_2\ \ldots\}$ denote the decision problem faced by P2 if her decision node is reached; her choice is between alternative distributions of material payoffs. The outer pair of brackets $\{_1\ \ldots\}$ denote the decision problem faced by P1 at his decision node: he can choose either to bring about the distribution (0, 0) or to allow P2 to choose from $\{_2\ \ldots\}$. Our analysis of this game can be

[3] Although 'kindness', 'unkindness', 'reward' and 'punishment' are terms used by Rabin himself, the word 'reciprocity' does not appear in Rabin's original paper. However, its use to describe motivations to reward kindness and punish unkindness is now standard (e.g. Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). When reporting the results of the first Trust Game experiment, Berg et al. (1995) describe their findings as evidence of 'reciprocity'.

[4] Charness and Rabin (2002: 851) and Dufwenberg and Kirchsteiger (2004: 272) make similar remarks about their models of reciprocity, which we will discuss later.

Please cite this article as: A. Isoni, R. Sugden, Reciprocity and the Paradox of Trust in psychological game theory, Journal of Economic Behavior and Organization (2018), https://doi.org/10.1016/j.jebo.2018.04.015

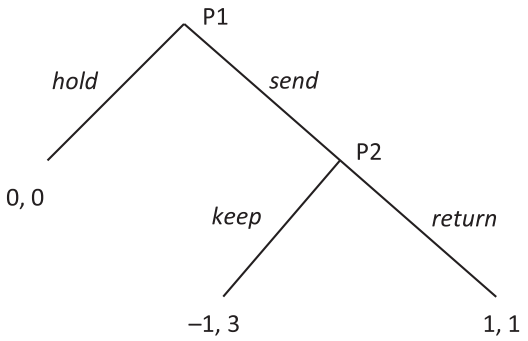

Fig. 1. The Trust Game (G_1).

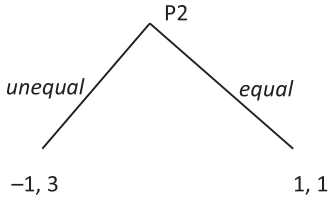

Fig. 2. P2's Dictator Game (G_1').

generalised to any game of the form $\{_1\ (0, 0), \{_2\ (x_1, x_2), (y, y)\}\}$ where $x_1 < 0 < y < x_2$. By setting the two players' payoffs equal to one another, both after *hold* and after (*send, return*), we screen out the effects of preferences for equality.[5]

We will also consider the game shown in Fig. 2 and denoted by $G_1' = \{_2\ (-1, 3), (1, 1)\}$, which we will call *P2's Dictator Game*. It is identical to the decision problem faced by P2 in the Trust Game, except that it is not preceded by any action by P1. The person who is P1 (respectively P2) in one game is also P1 (respectively P2) in the other, and so acts on the same preferences in both games. P2's actions in her Dictator Game will be called *unequal* (corresponding with *keep*) and *equal* (corresponding with *return*). Comparisons between these two games will allow us to analyse whether P2's behaviour in the Trust Game is contingent on P1's choice of *send*.

Behaviour strategies for P1and P2 in the Trust Game are fully described by the respective probabilities q_1 and q_2 with which P1 chooses *send* and (conditional on P1 having made this choice) P2 chooses *return*. To allow a psychological game-theoretic analysis, we also need to specify first- and second-order beliefs. We will say that there is *first-order consistency* of beliefs if P1 (respectively P2) believes that P2's (respectively P1's) behaviour strategy is q_2 (respectively q_1). In other words, players hold correct beliefs about each other's strategy choices. There is *second-order consistency* of beliefs if P1 (P2) believes that P2 (P1) believes that P1's (P2's) behaviour strategy is q_1 (q_2). In other words, players hold correct beliefs about each other's first-order beliefs. In P2's Dictator Game, the only behaviour strategy to consider is that of P2. The probability with which P2 chooses *equal* is denoted q_2'; first- and second-order consistency of beliefs are defined as before.

Now consider the following scenario, which we will call *Trust World*. In this scenario, $q_1 = q_2 = 1$ in the Trust Game and $q_2' < 1$ in P2's Dictator Game; in both games, there is first- and second-order consistency of beliefs. This scenario describes a world in which, if P1 and P2 play the Trust Game, P1 is fully trusting (he chooses *send* with probability 1) and P2 is fully trustworthy (she chooses *return* with probability 1). Each player believes that the other is fully trusting or trustworthy, and that the other believes that he or she is fully trustworthy or trusting. However, if P2's Dictator Game is played, P2 chooses *equal* with probability less than 1. Thus, P2's choice of *return* in the Trust Game cannot be explained simply by assuming that she has a history-independent preference for (1, 1) over (−1, 3), for example because of altruism or inequality aversion. The explanation of that choice must involve the idea that P2 chooses *return* in response to P1's choice of *send*.[6] Deliberately, our description of Trust World provides no explanation of the players' actions or beliefs. Our claim is that it provides an idealised representation of a psychologically intelligible interaction between a trusting first mover and a trustworthy second mover. Intuitively, Trust World seems to describe the workings of some kind of reciprocity. A model of reciprocal kindness falls prey to the Paradox of Trust if it cannot represent this description as an equilibrium.

We begin with Rabin's seminal model. Strictly interpreted, this model applies only to normal-form games, while the Trust Game is sequential. Extending Rabin's model to sequential games in general is not a trivial task, because of the need to

[5] This assumption is significant only for our discussion of the model proposed by Falk and Fischbacher (2006).

[6] There is experimental evidence that individuals are in fact more likely to choose *return* as second movers in Trust Games than to behave contrary to self-interest in corresponding Dictator Games (McCabe et al., 2003). For the purposes of our argument, however, we need to claim only that a person might plausibly have a motivation to reciprocate trust that is independent of altruism or inequality aversion.

Please cite this article as: A. Isoni, R. Sugden, Reciprocity and the Paradox of Trust in psychological game theory, Journal of Economic Behavior and Organization (2018), https://doi.org/10.1016/j.jebo.2018.04.015

track changes in first- and second-order beliefs over the course of a sequential game.[7] However, Trust World has properties that allow an uncontroversial translation of Rabin's assumptions to behaviour in that world. Because the Trust Game has only one decision node for each player, there is no need to consider how players' beliefs might change with changes in information. Because each of these nodes is reached with non-zero probability, we do not face the problem of having to justify assumptions about what a player would believe, were a zero-probability event to occur.[8] And because the probability of each action is either 0 or 1, it does not matter whether the kindness of a co-player's previous actions is assessed in terms of ex ante probabilities or ex post realisations.[9]

In Rabin's model, players get utility from their material payoffs, from being kind to co-players who are kind to them, and from being unkind to co-players who are unkind to them. They also get utility directly from the belief that a co-player is being kind to them, and get disutility directly from the belief that a co-player is being unkind to them. In a two-player game, the kindness of a player *i* is defined by reference to the dictator problem he faces, given his beliefs about the behaviour and beliefs of his co-player *j*. In the most general statement of the model, *i*'s kindness is defined relative to some unspecified normative rule for 'sharing along the Pareto frontier' of his dictator problem; that rule picks out an *equitable payoff* for *j* which lies strictly between the highest and the lowest payoffs for *j* on that frontier, unless there is only one such payoff on the frontier, in which case that payoff is the equitable one (pp. 1286, 1297). In the specific form of the model that Rabin favours, *j*'s equitable payoff is the average of the two extremes. If *i*'s chosen strategy implies an expected payoff for *j* that is higher (lower) than the equitable benchmark, *i* is kind (unkind) to *j*. The extent of *i*'s kindness or unkindness is measured entirely in terms of the effect of his decision on *j*'s payoff; no account is taken of how much *i* sacrifices in benefiting (or harming) *j*.[10]

In Trust World, P2 chooses *return*. By doing this, she chooses (1, 1) from the set {(–1, 3), (1, 1)}. Clearly, that choice is kind. Given his belief that P2 will choose *return* with probability 1, P1 chooses (1, 1) from the set {(0, 0), (1, 1)}. Since the Pareto frontier for this dictator problem is the singleton {(1, 1)}, P1's chosen action is neither kind nor unkind. Because of this, Rabin's model implies that P2 merely maximises her material payoff, and so chooses *keep* – contrary to the description of Trust World. The implication is that Rabin's model is not consistent with that scenario. In other words, the most famous social-preference model of reciprocity is unable to explain what, intuitively, seems to be a paradigm case of reciprocity. This illustrates the Paradox of Trust.

Rabin's paper includes a brief discussion of a normal-form version of the Trust Game ('Leaving a Partnership'). Rabin notes that the 'cooperative' strategy combination that corresponds with (*send, return*) is not an equilibrium in his model. He says that, contrary to this implication of his model, 'it seems plausible that cooperation would take place'. His response to this variant of the Paradox of Trust is to say that his model is not intended to represent all psychological factors that can affect behaviour in games; theorists may need to consider modelling 'additional emotions'. The additional emotion he has in mind is a desire 'to reward trust' (pp. 1296–1297). The idea seems to be that, in choosing *return*, P2 rewards P1 for the goodness of *send*, and that *send* is worthy of reward because it is an act of trust. Rabin does not pursue the issue of how trust should be defined, or why it might be thought to deserve to be rewarded. Nevertheless, Rabin's suggestion that trust and reciprocal kindness result from different psychological mechanisms is significant. If that suggestion is correct, the Paradox of Trust may not be a limitation of his model: it may be an informative result about the psychology of reciprocal kindness.

As an illustration of how the paradox might be informative, consider Akerlof's hypothesis about gift exchange in labour markets. Akerlof's model is effectively a Trust Game in which P1 is an employer and P2 is a worker; *send* is paying more than the minimum necessary wage and *return* is performing more than the minimum necessary effort. The gift-exchange equilibrium in this model is a version of Trust World. According to Akerlof, the relationship between employer and worker is one of reciprocal kindness or 'sentiment', analogous with the relationship between people who exchange gifts at Christmas (pp. 549–550). But imagine a worker who, after the employer has paid her more than the minimum wage, says: 'My employer has predicted that this wage will cause me to feel sentiment for the firm and so put in more effort. This isn't kindness; it is a sophisticated strategy for increasing profit. But now he is expecting me to incur a sacrifice to benefit him. He has acted on self-interest, so why shouldn't I do the same?' This response encapsulates the logic of Rabin's conception of reciprocal kindness. If workers predictably reasoned in this way, an opportunity for mutual benefit would be lost. That might be unfortunate, but the attitude we have attributed to the worker does not seem psychologically implausible. The implication, we suggest, is that the reciprocity expressed in practices of mutually beneficial trust and trustworthiness must be something other than reciprocal kindness.

[7] Dufwenberg and Kirchsteiger (2004) offer such a generalisation, but their model also includes a significant amendment to Rabin's assumptions. We will discuss this model in Section 2.3.

[8] This is a deep problem in epistemic game theory: see, for example, Binmore (1987), Pettit and Sugden (1989) and Reny (1992).

[9] In Rabin's model of the normal form of the Trust Game, P2's strategy choice depends on the perceived kindness or unkindness of P1's chosen strategy, which in general is a probability mix of *hold* and *send*. In models of sequential reciprocity, P2's choice at her decision node depends on the perceived kindness or unkindness of P1's *actual* choice (either *hold* with probability 1 or *send* with probability 1).

[10] This feature might be seen as a limitation of Rabin's original model, but it is orthogonal to our distinction between kindness and cooperation. It can be avoided by defining *i*'s kindness and unkindness in terms of implicit weights in *i*'s utility function, as in Charness and Rabin's (2002) model.

Please cite this article as: A. Isoni, R. Sugden, Reciprocity and the Paradox of Trust in psychological game theory, Journal of Economic Behavior and Organization (2018), https://doi.org/10.1016/j.jebo.2018.04.015

2. Other models of reciprocal kindness

We have argued that the Paradox of Trust is inherent to the psychology of reciprocal kindness. So far, however, our argument has been presented in relation to Rabin's model of reciprocal kindness. In this section, we examine other psychological game-theoretic models of reciprocal kindness to see whether they offer ways of avoiding the paradox.[11]

2.1. Charness and Rabin's model of reciprocity

Charness and Rabin (2002: 851–858) propose a psychological game-theoretic model of social preferences that applies to sequential games with material payoffs. In a two-player game, the utility of each player i is a weighted average of i's material payoff and a measure of 'social welfare'. In the baseline definition of this concept, social welfare is a weighted average of the minimum and the sum of the two players' payoffs. Negative reciprocity is modelled by defining, for each player i, an endogenous variable $d_i \leq 0$ which represents i's *demerit*. This is interpreted as a measure of 'how much [that player] deserves'; the more negative the value of d_i, the less i deserves (p. 853). If i has zero demerit, j's utility function uses the baseline measure of social welfare. The greater the absolute value of i's demerit, the less weight j's utility function gives to i's payoffs. Like unkindness in Rabin's original model, the demerit of player i is assessed by considering the dictator problem faced by i, given i's beliefs about j's behaviour. Player i has demerit to the extent that his decision implies that the weight he is giving to social welfare is less than some socially given 'selflessness standard'. Thus, if i's action reveals insufficient selflessness, it will induce a negative response from j. There is no positive reciprocity in the model.

This model is not compatible with Trust World. If P1 has zero demerit in the Trust Game, the decision problems faced by P2 in the Trust Game and in her Dictator Game are equivalent to one another, and so the model cannot explain why P2's behaviour is different in the two games. But if P1's demerit in the Trust Game is non-zero,[12] P2 will give less weight to P1's payoffs in that game than in her Dictator Game, and so P1 will be *less* likely to choose *return* in the former than to choose *equal* in the latter. Again, this is inconsistent with the description of Trust World and leads to the Paradox of Trust.

2.2. Falk and Fishbacher's model of reciprocity

Falk and Fischbacher (2006) propose a 'theory of reciprocity' which uses the framework of psychological game theory and applies to sequential games with material payoffs. This model combines elements of inequality aversion, trust responsiveness and reciprocal kindness. In a two-player game, the kindness of a player i at a given decision node is assessed in terms of (j's beliefs about) the payoff distribution implied by i's choice in the dictator problem he faces at that node, given his beliefs about j's behaviour. He is 'intentionally kind' (intentionally unkind) if he chooses a distribution in which j has advantageous (disadvantageous) inequality, and if he had an alternative option which would have given j a lower (higher) payoff.[13] The assumed form of reciprocity can be roughly expressed as: 'If your co-player is intentionally kind (unkind), try to reward (punish) her by ensuring that she gets a better (worse) outcome than she expects'. This model is not compatible with Trust World. In Trust World, P1's choice is between the payoff distributions (0, 0) and (1, 1). In choosing (1, 1), P1 gives P2 neither advantageous nor disadvantageous inequality, and so is neither kind nor unkind. Since Falk and Fischbacher's concept of reciprocity does not come into play, P2 acts on self-interest. But that implies that she chooses *keep*, contrary to the description of Trust World.

2.3. Dufwenberg and Kirchsteiger's model of reciprocity

Each of the models we have considered so far leads to some form of the Paradox of Trust. The model proposed by Dufwenberg and Kirchsteiger (hereafter DK; 2004) is an exception. DK retain the main features of Rabin's original model and extend it to sequential games. Because of the special features of Trust World described in Section 1, we do not need to discuss all the subtleties involved in moving from normal-form to sequential games. However, DK make a number of amendments to Rabin's model. We will focus on one of these amendments, which DK present as having the merit of eliminating the Paradox of Trust (pp. 289–290).[14]

[11] We omit a detailed analysis of Levine's (1998) attempt to represent reciprocal kindness without using psychological game theory. In Levine's model, players have 'types', distinguished by their degrees of altruism or 'spite', and get utility from being kind (unkind) to co-players who are altruistic (spiteful). If players are initially uncertain about each other's types, a kind action can signal altruism, and hence induce reciprocal kindness. In a Trust Game in which (as in Trust World) P2 believes with probability 1 that P1 will choose *send, send* has no information content, and so P2's decision problem is equivalent to that of her Dictator Game. This is a variant of the Paradox of Trust.

[12] Intuitively, it seems unreasonable to treat P1's choice of *send* as an indication of demerit. Formally, however, that choice can be rationalised by any utility function in which P1's payoff and social welfare both have positive weight – including utility functions that fail to meet the selflessness standard.

[13] Falk and Fischbacher also allow the possibility that actions that lead to unequal outcomes are perceived as 'kind' or 'unkind' even if they were unintended. For our purposes, it is not necessary to consider this possibility.

[14] DK argue that this amendment has a further advantage. Because of the way Rabin defines the concept of 'equitable payoff', a straightforward extension of his model to sequential games would imply the non-existence of equilibrium in some games. By making each player's equitable payoff independent of players' beliefs, DK's amendment avoids this problem (Dufwenberg and Kirchsteiger, 2004, p. 289).

Please cite this article as: A. Isoni, R. Sugden, Reciprocity and the Paradox of Trust in psychological game theory, Journal of Economic Behavior and Organization (2018), https://doi.org/10.1016/j.jebo.2018.04.015

Consider a sequential game for two players. At any given decision node for a player *i*, *i*'s kindness is defined in relation to the set of payoff distributions that are feasible for *i*, given his updated beliefs about *j*'s behaviour. So far, this is a straightforward generalisation of the dictator problem that underlies the concept of kindness in Rabin's model. Recall that in Rabin's model, *i*'s kindness to *j* is measured relative to an 'equitable payoff' for *j* that is defined as the average of the highest and lowest payoffs for *j* on the Pareto frontier of this problem. DK use a different definition of 'equitable payoff'.

This definition rests on the more basic concept of an *efficient* strategy. Whether any specific behaviour strategy is efficient is a property of that strategy in relation to the game as a whole; it is independent of players' beliefs and is invariant with respect to the history of play. In general, a behaviour strategy for a given player 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' (p. 276). Roughly speaking, a strategy for a given player is inefficient if some other strategy Pareto-dominates it in material payoffs for *all possible* decisions by other players. At any given decision node for player *i*, *j*'s equitable payoff is 'the average between the lowest and the highest material payoff of *j* that is compatible with *i* choosing an efficient strategy' (pp. 276–277). As in Rabin's model, *i* is kind (unkind) to *j* if, given his beliefs about *j*'s strategy, his action induces a payoff for *j* that is greater than (less than) *j*'s equitable payoff.

Notice a fundamental difference between Rabin's and DK's measurements of kindness. In both models, the dictator problem that is used to assess *i*'s kindness is defined in relation to *i*'s beliefs about other players' strategies. In Rabin's model, *i*'s kindness to *j* is measured relative to an equitable payoff for *j* that is defined solely in terms of the properties of *i*'s dictator problem, and thereby in relation to *i*'s beliefs. In DK's model, in contrast, the benchmark from which *i*'s kindness is measured is independent of *i*'s beliefs. DK do not give any psychological intuition for this construction except for an implicit appeal to the psychological plausibility of (*send, return*) as an equilibrium in the Trust Game. But for this to be satisfactory, the amendment must have similarly plausible implications for other games.

Before exploring this issue, we show that DK's model is compatible with the properties of Trust World. To do this, we assume that actions and beliefs are as specified by the description of Trust World, and check that there is no contradiction with the assumptions of DK's model. Notice that all strategies in the Trust Game are efficient. In particular, *hold* is efficient because its outcome, (0, 0), is not Pareto-dominated by every possible outcome of *send*. If P1 were to choose *send*, P2 would have the option of choosing *keep*, leading to (−1, 3). In P1's belief, the probability of *keep*, conditional on *send*, is zero; but DK's concept of efficiency is belief-independent. Given his actual beliefs, P1 faces the dictator problem {(0, 0), (1, 1)}. Since both P1's strategies are efficient, the equitable payoff for P2 is 0.5, and so *send* is kind; it also maximises P1's material payoff. Thus, if (in P1's belief) *return* is not unkind, P1's choice of *send* is consistent with the model. Given that P1 has chosen *send*, P2's dictator problem is {(−1, 3), (1, 1)}, and the equitable payoff for P1 is 0. If P2 believes that *send* is kind, and if she has a sufficiently strong preference to reciprocate kindness, her choice of *return* is kind and utility-maximising. Thus, the combination of *send* and *return* is consistent with the model. In P2's Dictator Game, in contrast, reciprocity does not come into play, and so DK's model implies that P2 will choose *unequal*, consistently with the description of Trust World.

Given the underlying conceptual framework of social preference theory, DK's classification of *send* as kind might seem psychologically plausible. It seems clear that P1's choice of *send* has *some* property that might (but not necessarily will) induce P2 to choose *return*, forgoing material payoff in a way that benefits P1. *If* one thinks in terms of reward and punishment, and *if* one thinks of kindness as the characteristic feature of actions that deserve reward, it is tempting to conclude that *send* must be kind. For the moment, we set aside our reservations about these 'if …' clauses, and ask whether DK's kindness classifications are psychologically plausible in other cases.

Consider game G_2, defined by $G_2 = \{_1\ (0, 0), \{_2\ (0.5, -0.5), (1, 1)\}\}$. For convenience, we label actions as in the Trust Game. That is, the first option in $\{_1\ \ldots\}$ is *hold* and the second is *send*; the probability of *send* is q_1. Similarly, the first option in $\{_2\ \ldots\}$ is *keep* and the second is *return*; the probability of *return* is q_2. Consider the scenario in which $q_1 = q_2 = 1$, and in which first- and second-order beliefs are consistent in the sense defined in Section 1. This scenario differs from the Trust World scenario for G_1 only with respect to the consequences of the zero-probability strategy combination (*send, keep*). In G_2, it is in P2's interest to choose *return*, and so P1's expectation of this action does not require a belief that P2 acts on an other-regarding motivation. And whatever expectations P1 had held about P2's action, it would still have been in his interest to choose *send*. In this case, it seems natural to say that *send* is neither kind nor unkind, but merely prudent or rational. Rabin's model delivers this classification. Just as in the Trust World scenario for G_1, P1's dictator problem is {(0, 0), (1, 1)}; the Pareto frontier for this problem is the singleton {(1, 1)}. According to Rabin's definitions, *send* is therefore neither kind nor unkind. Notice, however, that neither of P1's strategies is inefficient. Thus, according to DK's definitions, the equitable payoff for P2 is 0.5, just as in G_1, and *send* is kind.

Now consider game G_3, defined by $G_3 = \{_1\ (0, 0), \{_2\ (0.5, 0.5), (1, 1)\}\}$. Again, actions are labelled as in the Trust Game. Consider the scenario in which $q_1 = q_2 = 1$, and in which first- and second-order beliefs are consistent. The only difference between the G_2 and G_3 scenarios is that G_3 gives P2 a higher payoff in the zero-probability event that *send* is followed by *keep*. But now (0, 0) is not only below the Pareto frontier for P1's dictator problem; it is also inefficient in DK's sense. Thus, Rabin and DK agree in classifying *send* as neither kind nor unkind. Considered in isolation, this classification makes intuitive sense. But the comparative statics of DK's model are puzzling. Why is *send* kind in G_2 but not in G_3? The only difference between the scenarios concerns an opportunity that P1 gives to P2 by choosing *send*. In each case, this is an opportunity that P2 has no reason to choose, and in fact does not choose. This opportunity has a higher material payoff for P2, and the

Please cite this article as: A. Isoni, R. Sugden, Reciprocity and the Paradox of Trust in psychological game theory, Journal of Economic Behavior and Organization (2018), https://doi.org/10.1016/j.jebo.2018.04.015

same material payoff for P1, in the second case than in the first. How can this difference make P1's action *less* deserving of reward in the second case?

Thus, although DK's amendment avoids the Paradox of Trust, it does not seem to offer a psychologically plausible account of how reciprocal kindness can explain trust and trustworthiness.

3. Reciprocal kindness *versus* reciprocal cooperation

In models of reciprocal kindness, kindness (or its opposite, unkindness) is the most basic other-oriented motivation. Recall that, in a two-player game, the kindness of a player *i* is understood as if he were responding to a dictator problem in which he could choose from a fixed set of alternative payoff distributions for himself and his co-player *j*. The composition of this set depends on *i*'s beliefs about *j*'s behaviour and beliefs, but *i*'s beliefs are taken as given. It follows that *i*'s kindness to *j* cannot affect *i*'s beliefs about *j*'s behaviour: *j*'s role in the problem is entirely passive. Thus, although *i*'s kindness expresses an attitude towards *outcomes for j*, it cannot be intended as the first move in a cooperative interaction with *j*: it must be *gratuitous*, a free gift. Reciprocity enters these models only as a second-order preference for rewarding a co-player's kindness or for punishing her unkindness. Since kindness is gratuitous, rewarding it cannot be intended as the second move in a cooperative interaction. In this sense, reward is gratuitous too. The contrast between gratuity and cooperation is the source of the Paradox of Trust.

It seems inescapable that in any credible model of reciprocal kindness in which there is consistency of first- and second-order beliefs, P2's choice of *return* in the Trust Game must be classified as kind and her choice of *keep* as unkind. If such a model is to be compatible with Trust World, P2's choice of *return* must be a response to the kindness of *send*. The difficulty is to explain how *send* can express kindness, given that P1 believes that P2 will choose *return* with probability 1. Since P1's dictator problem is {(0, 0), (1, 1)}, it seems that the only inference that can be drawn from his choice of *send* is that he is not so gratuitously malevolent as to want to impose equal losses on himself and P2.

Against this claim, it might be objected that the mere possibility that P2 might choose *keep* matters for an assessment of the kindness of *send*, because P1 is exposing himself to the possibility of loss. This would be contrary to the basic principles of both psychological and conventional game theory, in which what matter are players' beliefs, because in Trust World, P1 assigns zero probability to that event. However, it might be said that, in assessing the kindness of P1's action, we should ignore any benefits that he receives (however predictably) as a reward for its kindness. That thought fits with the familiar idea that a genuinely kind person does not consider how his kindness might be rewarded: virtue should be its own reward, even if in fact it brings material rewards too.[15] This argument would perhaps have some force in Trust World if P2 believed that P1 would have chosen *send* even if he had expected P2 to choose *keep*. That belief is not wholly implausible, since (–1, 3) might be judged to generate more social welfare than (0, 0), and P1 might have a high degree of 'selflessness'. If that were the case for P1, his hypothetically selfless decision might be deemed to be gratuitously kind. By the same token, however, it would not be an act of trust in the normal sense of the word, since P1 would not be relying on P2 to perform any particular action. It would certainly be true to say that P2's being free to choose *keep* is an essential part of what makes *send* an act of trust, and not just a response to a dictator problem. But if trust and trustworthiness are to be understood as reciprocal, P1's choice of *send* must be in response to his belief that (with sufficiently high probability) P2 will in fact choose *return*.

As this example illustrates, trust is *not* a form of kindness. Kindness is gratuitous, but trust is construed by the first mover as the beginning of an interaction with a trustee. The intuition that the Trust Game really is a model of trust and trustworthiness seems to depend on the thought that, if P1 chooses *send*, he thinks of himself as playing his part in a *joint* action, the other part of which is P2's choice of *return*. Similarly, if P2 chooses *return* in response to P1's choice of *send*, she does not think of herself as gratuitously rewarding P1 for his gratuitous kindness to her: she is playing her part in the joint action that P1 has initiated. The relationship between P1's trust and P2's trustworthiness is a kind of reciprocity, but it is not reciprocal kindness. It is *reciprocal cooperation*, the reciprocity of playing one's part in a cooperative practice when one believes that the other party to that joint action will play (or has played) hers.

The Paradox of Trust is that expecting to benefit from one's own act of kindness can undermine the kindness of the act. Reciprocal cooperation does not run into a similar paradox: expecting to benefit from an act of cooperation does not undermine the cooperativeness of the act. To the contrary: the whole point of cooperation is that both parties benefit. Interestingly, when Berg et al. (1995, p. 124) describe their Trust Game experiment as a study of 'reciprocity', they seem to be thinking of reciprocal cooperation rather than reciprocal kindness: 'If the [second mover] interprets the [first mover's] decision to send money as an attempt to improve the outcome for both parties, then the [second mover] is more likely to reciprocate'.

Think of the worker in Akerlof's model. She knows that the employer is not being gratuitously kind: he expects to benefit by paying a relatively high wage. Nor does the employer perceive himself to be incurring any risk in doing so: he believes the probability of benefiting to be 1. His intention is to initiate a relationship with the worker that he expects to be mutually

[15] A related idea is developed by Roel (2017) in a recent paper about trust and reciprocity. Roel proposes an alternative to DK's definition of an 'efficient' strategy. The essential idea is that a strategy for player *i* is *trust-efficient* if it is Pareto-efficient, given the strategy that (in *i*'s belief) *j* would have played, had she not been 'generous'. In the Trust Game, a non-generous P2 would choose *keep*, and so *hold* is trust-efficient and *send* is kind.

Please cite this article as: A. Isoni, R. Sugden, Reciprocity and the Paradox of Trust in psychological game theory, Journal of Economic Behavior and Organization (2018), https://doi.org/10.1016/j.jebo.2018.04.015

beneficial. If the worker supplies the effort that the employer expects, she is not *rewarding* him for that intention. It would be truer to say that she is *joining in* with or *completing* that intention.

4. A model of reciprocal cooperation

In previous sections, we have argued that the behaviour exhibited by the Paradox of Trust should be understood as reciprocal cooperation rather than as reciprocal kindness. We now offer a sketch of how a theory of reciprocal cooperation might explain that behaviour. To be more precise, we show how such a theory might explain a Trust Game scenario in which $q_1 = q_2 = 1$ without having implications about behaviour in any Dictator Game. Our approach is to adapt one of the first behavioural models of reciprocal cooperation, that of Sugden (1984).[16]

Sugden's model is concerned with voluntary contributions to public goods. In this model, there is a set N of players. Each player i simultaneously chooses a non-negative *contribution* towards the provision of a public good from which all players benefit, not necessarily to the same extent. Consider any given profile of contributions by players other than i, and any set of players $S \subseteq N$ that contains i. Player i's *Kantian obligation to S* is the solution to the following problem: Given the actual contribution of each player (if any) who is not a member of S, and subject to the hypothetical constraint that each member of S must make the same contribution as every other member, what level of contribution would maximise i's utility? If S contains at least one player in addition to i, i's *reciprocal obligation* to S is whichever is the smaller of (1) his Kantian obligation to S and (2) the smallest contribution made by any member of S other than i himself. If $S = \{i\}$, i's reciprocal obligation to S is defined to be the same as his Kantian obligation to that set or, equivalently, i's utility-maximising response to the actual behaviour of his co-players. An equilibrium of the model is a profile of contributions such that each player's contribution takes the smallest value that is consistent with his reciprocal obligation to every set S of which he is a member. The intuitive idea is that each individual maximises utility subject to a self-imposed moral constraint that requires him to match other people's contributions to mutually beneficial arrangements to supply the public good.

Now consider how this model might be adapted to apply to a simple two-person sequential Public Good game. Suppose that P1 and P2 are the potential beneficiaries of a public good. P1 moves first, choosing whether to contribute two units of material payoff to a public account (*contribute*) or not to contribute (*decline*). If P1 chooses *decline*, the game ends. If he chooses *contribute*, P2 faces the same choice as P1 did. Each player earns 0.75 units for every unit contributed to the public account by either player. In our notation, this corresponds to the Trust Game $G_4 = \{_1\ (0, 0), \{_2\ (-0.5, 1.5), (1, 1)\}\}$.

In terms of Sugden's model, each player's Kantian obligation to {P1, P2} is the action *contribute*. If P1 chooses *contribute*, he has met this obligation, and so P2 has a reciprocal obligation to choose *contribute* too. If P1 expects that his choice of *contribute* will be reciprocated by P2, that choice maximises his utility and ensures that he meets his reciprocal obligation to {P1, P2}. Thus, (*contribute, contribute*) is an equilibrium in the sense of the model. The model is therefore compatible with the Trust World scenario of the Trust Game. Notice that in this equilibrium, neither player is gratuitously kind, and neither acts with the intention of rewarding the other for his or her good behaviour. P2's action is contrary to her self-interest, given the action that P1 has already taken, but she performs it as her part of a combination of actions that benefits them both.

5. Conclusion

The Paradox of Trust has sometimes been viewed as revealing a limitation of Rabin's model of reciprocity that can be repaired by some minor amendment, such as Dufwenberg and Kirchsteiger's change to Rabin's definition of 'equitable payoff'. We have argued against this interpretation. Rabin's model is based on a conceptually and psychologically coherent notion of reciprocal kindness. If this model is interpreted as a formalisation of this motivation, considered in isolation, the Paradox of Trust is not a deficiency. To the contrary, it can be seen as one of the significant results of psychological game theory. It shows that if individuals' other-regarding motivations are solely those of reciprocal kindness, there can be situations in which mutually beneficial trust and trustworthiness are not sustainable.

One way of reading this result is as a potential explanation of situations in which opportunities for mutual benefit are *not* realised. For example, as we suggested in Section 1, it identifies a psychological mechanism that might frustrate an employer's attempt to induce unobservable effort by paying a worker more than her reservation wage. But it can also be read as a demonstration that reciprocal kindness is not the only form that reciprocity can take.

As we noted in Section 1, the latter interpretation of the Paradox of Trust can be found in Rabin's original paper. Rabin left open the question of how cooperation should be explained, but suggested that trustworthiness might be modelled as a desire to reward the intentions that lie behind acts of trust – an approach that would be in the spirit of psychological game theory. We have argued that the motivation to be trustworthy should not be understood in terms of rewarding another person's meritorious action. Instead, acts of trust and trustworthiness should be seen as complementary components of cooperative practices in which each participant is motivated to play his or her part, conditional on other parties playing theirs. If behavioural economics is to explain mutually beneficial trust and trustworthiness, it needs a theory of reciprocal

[16] As the focus of our paper is on psychological game theory, we do not try to review the various ways in which reciprocal cooperation and related concepts have been modelled. These include the models of 'team reasoning' proposed by Sugden (1993, 2015), Bacharach (1999, 2006) and Karpus and Radzvillas (2017), and the model of 'virtual bargaining' proposed by Misyak and Chater (2014).

Please cite this article as: A. Isoni, R. Sugden, Reciprocity and the Paradox of Trust in psychological game theory, Journal of Economic Behavior and Organization (2018), https://doi.org/10.1016/j.jebo.2018.04.015

cooperation that is not ultimately grounded in concepts of gratuitous kindness and unkindness and of reward and punishment.

References

Akerlof, G., 1982. Labor contracts as partial gift exchange. Q. J. Econ. 97, 543–569.
Bacharach, M., 1999. Interactive team reasoning: a contribution to the theory of cooperation. Res. Econ. 53, 117–147.
Bacharach, M., 2006. Beyond Individual Choice. Princeton University Press, Princeton.
Bacharach, M., Guerra, G., Zizzo, D., 2007. The self-fulfilling property of trust: an experimental study. Theory Dec. 63, 349–388.
Battigalli, P., Dufwenberg, M., 2007. Guilt in games. Am. Econ. Rev. 97, 171–176.
Binmore, K., 1987. Modeling rational players: Part 1. Econ. Philosoph. 3, 179–214.
Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games Econ. Behav. 10, 122–142.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Q. J. Econ. 117, 817–869.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47, 268–298.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games Econ. Behav. 54, 293–315.
Hobbes, T., 1651/ 1962. Leviathan. Macmillan, London.
Karpus, J., Radzvillas, M., 2017. Team reasoning and a measure of mutual advantage in games. Forthcoming. Econ. Philosoph. https://doi.org/10.1017/S0266267117000153.
Levine, D., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dyn. 1, 593–622.
McCabe, K., Rigdon, M., Smith, V., 2003. Positive reciprocity and intentions in trust games. J. Econ. Behav. Org. 52, 267–275.
Misyak, J., Chater, N., 2014. Virtual bargaining: a theory of social decision-making. Philosoph. Trans. R. Soc. B 369, 20130487.
Pelligra, V., 2005. Under trusting eyes: the responsive nature of trust. In: Gui, B., Sugden, R. (Eds.), Economics and Social Interaction. Cambridge University Press, Cambridge 195--124.
Pettit, P., Sugden, R., 1989. The backward induction paradox. J. Philosoph. 86, 169–182.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Am. Econ. Rev. 83, 1281–1302.
Reny, P., 1992. Backward induction, normal form perfection and explicable equilibria. Econometrica 60, 627–649.
Roel, M., 2017. A theory of reciprocity with trust. http://personal.lse.ac.uk/roel/roel_jmp.pdf. Accessed 23 November 2017.
Sugden, R., 1984. Reciprocity: the supply of public goods through voluntary contributions. Econ. J. 94, 772–787.
Sugden, R., 1993. Thinking as a team: toward an explanation of nonselfish behavior. Social Philosoph. Policy 10, 69–89.
Sugden, R., 1998. Normative expectations: the simultaneous evolution of institutions and norms. In: Ben-Ner, A., Putterman, L. (Eds.), Economics, Values, and Organization. Cambridge University Press, Cambridge, pp. 73–100.
Sugden, R., 2015. Team reasoning and intentional cooperation for mutual benefit. J. Social Ontol. 1, 143–166.

Please cite this article as: A. Isoni, R. Sugden, Reciprocity and the Paradox of Trust in psychological game theory, Journal of Economic Behavior and Organization (2018), https://doi.org/10.1016/j.jebo.2018.04.015