Contribution to *Abstracta* symposium on *Knowledge by Trust*, by Paul Faulkner

Katherine Hawley, April 2012

Paul Faulkner has given us a very rich and complex book, one which will repay careful study both now and in years to come, for anyone interested in trust, testimony or epistemology more widely. Rather than attempting to summarise his arguments, or to grapple with them as a whole, I will focus here on Faulkner's development of his keystone 'problem of cooperation', which he generates through discussion of the 'Testimony Game', and its parallels with the 'Trust Game'.

1. The Trust Game

The Trust Game is an experimental set-up which is used by economists and social psychologists to investigate our behaviour and choices under certain constrained circumstances. In the standard Trust Game, a first player is given £10, and can choose either to keep all of this, or else to transfer some or all of the £10 to a second player. Whatever the first player decides to transfer is quadrupled by intervention from the experimenter. For example, if the first player decides to keep £6 the second player receives £16. The second player can then opt either to keep all she receives, or else to send some or all of it to the first player.

This is a peculiarly artificial situation, where the stakes are low (no-one risks their own personal funds), and players are somewhat distanced from ordinary social norms (it's just a game, after all). But social scientists like it because it allows them to put numbers on people's behaviour, by recording how much is transferred in each direction, and then to see how these numbers vary as the experimental set-up is varied. For example, the Trust Game can be played as a one-off, or repeatedly between the same two players. The game can be used with players of different ages, or different nationalities, or different genders, either in matching pairs or across these categories. Players may be allowed to interact face-to-face, to see photographs of one another, to speak by phone, or to have no direct contact whatsoever. Players may be told they are members of the same fraternity or club. Players may be given a whiff of oxytocin, the hormone sometimes known as the 'cuddle chemical'. Players' behaviour in the game may be compared with their responses to poll questions about trustworthiness in society, and so on.

The first puzzle about the Trust Game concerns the behaviour of the second player. Even when the game is played as a one-off, a proportion of second players choose to transfer some money to the first player. And this behaviour is hard to understand in terms of self-interest. Suppose you have £16. You can either take it home, or else give some of it to a stranger with whom you will never interact again. Why do anything other than keep the whole £16? What is motivating those second players who decide to transfer some of their cash?

The second puzzle concerns the behaviour of the first player. Even when the game is played as one-off, many first players choose to transfer some money to the second player. This suggests that either the first players are predicting that the second players will not be motivated by pure self-interest, or else that the first players themselves are not motivated by pure self-interest. You have £10, which you can either keep, or else share with a stranger who will have no obvious motivation to send you any money in return. Why do anything other than keep the whole £10?

Collectively, these puzzles raise both a normative and a non-normative issue. The non-normative issue concerns the empirical mismatch between the assumption that people are motivated by narrow self-interest, and expect others to be likewise motivated, and the actual behaviour of people who play these games. Some players do not behave as these assumptions would predict: what, then is their additional or alternative motivation? (We might address this question by exploring other set-ups, such as the dictator game, in which the first player simply chooses whether or not to transfer some money to the second player. Either way, that's the end of the game. A substantial minority of first players transfer money even in the dictator game.)

The normative issue arises from the fact that the first player ends up better off (takes home more than £10) if she takes a risk on the second player, and the risk pays off. And of course the second player benefits from any degree of transfer from the first player. If the first player takes a risk, then the experimenter injects more cash into the game, which can potentially benefit both players. So it seems intuitively that there's good reason to take this initial risk, which conflicts with the idea that there is no good reason to do this. What is the rational choice for the first player?

The difference between the non-normative and the normative puzzle can be brought out by considering other experimental situations in which people commonly make mistakes in reasoning or judgement. (One example: when a character, Linda, is described as active in radical student politics, people often judge that in later life it's more likely that she becomes a feminist bank teller than that she becomes a bank teller. Another example: confronted with two-sided cards, people are bad at judging which cards they need to turn over in order to test the rule that if a card has a vowel on one side, it has an even number on the other.) In such situations, there is a non-normative puzzle: why do people commonly make this mistake? But there is no corresponding normative puzzle, no temptation to think that in some sense there is a genuinely good reason to make the judgement in question.

For Faulkner, the Trust Game illustrates the 'problem of cooperation'. He says 'For the investor [i.e. the first player] to be acting reasonably in making a transfer, he needs to think, for whatever reason, that the trustee [i.e. the second player] will make a back-transfer. This game then illustrates how cooperation can be problematic because it is arguable that we often lack grounds for thinking this, but make transfers nevertheless. We trust and yet appear to be unreasonable in doing so.' (p.4)

Is this the normative or the non-normative issue? I think it's a bit of both. There's the empirical fact that many first players make a transfer (i.e. 'trust' the second player), which they would not do if they were motivated purely by narrow self-interest and assumed that the second player was likewise motivated (and they were able to reason through the consequences of this assumption). So there's the non-normative question of what motivates such first players. But then there's also the normative question of whether what motivates such first players does after all make their choice a reasonable one. Faulkner suggests that we cannot see such choices as reasonable unless the first player is motivated by the belief that the second player will cooperate (or at least by a belief that this is fairly likely), and that the first player has grounds for this belief.

2. The Testimony Game

Faulkner then argues that the Testimony Game is importantly analogous to the Trust Game. Here is a description of the Testimony Game which attempts to make that

analogy as closely as possible. The first player is the audience, and the second player is the speaker, who asserts that *p*, let's say. The first player begins, not with £10, but with no opinion as to whether *p*. This neutrality is worth something epistemically, in that it is better than a false belief as to whether *p*. The first player then has the option of keeping this neutrality (i.e. ignoring what the speaker says), and ending the game, or else giving some credence to what the speaker says. If he gives some credence to what the speaker says, the speaker (the second player) then has two options. The first is to take the credence and walk away – that is, to take the benefits of being believed, but not to return the favour by being trustworthy. The second option is to take the credence, and return something of benefit to the first player, the audience, by being trustworthy.

In the Trust Game, the experimenter injects extra cash iff the first player decides to make a transfer, and this extra cash can benefit both parties iff the second player also makes a transfer. In the Testimony Game, the analogous idea must be that the benefit to the second player (the speaker) of being given credence is greater than the potential loss carried by the first player (the audience) in moving from neutrality to giving credence; if the second player (the speaker) responds by being trustworthy, then both parties end up better off than they would have been had the audience ignored the speaker. Generalising, we are all better off if as a rule audiences give credence and speakers are trustworthy.

Recall that there are two initial puzzles about the Trust Game. Why does the second player transfer cash, as opposed to taking the whole lot home, as narrow self-interest seems to dictate? And why does the first player transfer cash, given the assumption that the first player is motivated by narrow self-interest, and assumes that the second player is likewise motivated? These then generate a non-normative question: why do people behave in this way? And they generate a normative question: narrow self-interest seems to make the first player's transfer unreasonable, but the overall benefits of cooperation seem to make the first player's transfer reasonable. So which is it?

In the Testimony Game we may likewise ask why the speaker (the second player) decides to be trustworthy, as opposed to saying whatever suits her immediate self-interest best. And we may ask why the audience (the first player) offers credence, on

the assumption that she is motivated by narrow self-interest, and assumes the speaker is likewise motivated. Then there is the non-normative puzzle: what motivates audiences to offer credence, and speakers to decide to be trustworthy? And the normative puzzle: narrow self-interest seems to make the audience's decision unreasonable, but the overall benefits of cooperation seem to make it reasonable. So is it reasonable or unreasonable?

Faulkner articulates the problem of cooperation in the context of the Testimony game as follows: "the acceptance of testimony must be backed by reasons if it is to be reasonable. The problem of cooperation is then the problem of giving an account of the satisfaction of this condition. It is the problem of explaining the rationality of testimonial cooperation. This is then problematic to the extent that this condition cannot be satisfied; that is, to the extent that we lack reasons – or have a psychological tendency to trust that outstrips our possession of reasons. For the moment, I will leave it open whether, if at all, testimonial cooperation is problematic." (pp. 6-7)

3. Differences between the Trust Game and the Testimony Game.

I think that the differences between the Trust Game and the Testimony Game are too great for Faulkner to be able to draw on the analogy between the two, as he wishes to do. Moreover, the differences between the Trust Game and real-life cases of testimonial exchange, even between strangers, are even greater. I will outline these differences before going on to explore their significance for Faulkner's arguments.

The first crucial difference is that the first player (the investor) makes the first move in the Trust Game, before any other interaction between the two players, and the second player then reacts to the first player's decision. But in the Testimony Game, matters begin with the 'second' player (the speaker) making an assertion; the 'first' player (the audience) then decides whether or not to abandon neutrality of belief, and give some credence to the speaker.

This temporal issue is significant in two ways. The 'second' player decides whether to be trustworthy before knowing what the audience will decide to do. This may affect the way people think about what to do: in the lab-based Trust Game, second

players are more likely to return cash if the first player has made a relatively generous initial transfer. (This tendency is also illustrated in the Ultimatum Game, in which first players can choose either to keep their stake, or else to transfer some of it to a second player. The second player can either accept what's transferred, or else reject it, in which case both first and second player lose everything. Second players show a marked tendency to 'punish' small initial offers, even at financial cost to themselves.) In testimonial situations, speakers often have to decide whether to be trustworthy even before they know whether their audience will trust them, and cannot react to the the audience's decision.

Moreover the temporal issue means that the audience decides what to do in light of the fact that the speaker has volunteered an assertion. This fact itself is a significant piece of evidence about the speaker, of a kind which is unavailable to first players in the Trust Game.

Ironically, the speaker's volunteering an assertion makes it more accurate to talk of trust in connection with the Testimony Game than in connection with the so-called Trust Game. Second players in the Trust Game who decide to keep what they've been given may perhaps be described as mean, or selfish, or spoilsports, but they are not *untrustworthy*. After all, they make no prior agreement to return any part of the cash, and in general there is no obligation to give money to (non-destitute) people who would like you to, even if they have previously given you money. First players in the Trust Game are not genuinely trusting, they are opting to take a risk: handing over money without being asked, then complaining if it is not returned with interest, is otherwise known as loan-sharking (or sub-prime mortgage mis-selling). By contrast, in the Testimony Game the speaker effectively asks for credence by choosing to make an assertion, and it makes sense to think of the audience trusting or distrusting in response.

The second crucial difference is that, in the Trust Game, 'never take a risk' is a reasonable strategy for someone acting as first player with a sequence of different partners. This may not be the income-maximising strategy, but on every occasion the first player will get to keep the £10, which is not too bad. In the Testimony Game, it is much less clear that 'never give credence' is a reasonable long-term strategy for

audiences: maintaining neutrality about every proposition you cannot check for yourself will leave you with very few beliefs indeed. This fact may alter the 'pay-off matrix' for the Testimony Game, if the relative value of neutrality over false belief shrinks as we increase the number of cases. (One might try to model this in the Trust Game by telling the first player that the stake will be reduced by £1 on each iteration.)

The third crucial difference is that in the Trust Game, the currency (cash) is fully meaningful even in the one-shot version, where players interact just once. In the Testimony Game, as Faulkner makes clear, the 'currency' offered by the speaker is not truth but trustworthiness. This is because the currency offered needs to be something which the speaker could benefit from retaining: in a given instance, it might suit the speaker to offer the truth (in boasting about an achievement, for example), whereas trustworthiness involves commitment to speak the truth both when this is convenient and when it is inconvenient. This currency of trustworthiness is thus only fully meaningful when we consider a sequence of interactions

The fourth crucial difference is that in the Trust Game, players do not switch roles. Even when the interaction is repeated a number of times, individual people stick with their roles as first player or as second player. In real-life testimonial situations, we often switch roles between speaker and audience, even within a given pair: this is otherwise known as a conversation.

4. Consequences of these differences

Whatever we make of these differences, there is no doubt that Faulkner has given us a fresh, fruitful way of thinking about the challenges we face in testimonial exchange. But what are the consequences of these disanalogies for Faulkner's broader arguments? So far as I can see, the analogy and attendant problem of cooperation are used in two main ways in *Knowledge on Trust*: to undermine the nonreductionist view that we have a default entitlement to accept testimony, and to make plausible Faulkner's views about social norms of trust and trustworthiness. I will briefly discuss these in turn.

For Faulkner, the situation of the first player in the Trust Game, and by extension the situation of the audience in the Testimony Game indicates that trust (transferring

money, giving credence) is unreasonable unless the first player has some positive reason to think that this will pay off (that the second player will back-transfer money, that the speaker will be trustworthy). But a number of the crucial differences shed some doubt on this. The fact that the 'second player' begins, by volunteering an assertion, in the Testimony Game, may mean that the very set-up makes it reasonable for the first player to trust. Admittedly, we might think of this as evidence available to the first speaker, which tells in favour of the reductionist view, but the nonreductionist might instead think of this as a feature of the situation which facilitates default entitlement. Moreover the fact that 'never give credence' (unlike 'never transfer cash') is not even a moderately-good long term strategy might also tell in favour of a default entitlement. And perhaps the nonreductionist might make something of the fact that we often switch, unpredictably, between the roles of speaker and audience, unlike players of the Trust Game.

Finally, though I find Faulkner's emphasis on our awareness of social norms of trust and trustworthiness very compelling, I suspect that he is drawing additional, unwarranted support for his view from the rather weak analogy between the two games. Faulkner's discussion of trust is subtle, and he draws out the importance of normative expectations, of the trustee's recognition that the truster makes herself dependent through her trust, and of the trustee's being motivated by concern for this dependency. None of this applies to the Trust Game, in which normative expectations are inappropriate, and the second player's actions are what count, not her motivations; as I argued above the Trust Game is not really about trust. (In Faulkner's terms, the Trust Game involves at most predictive trust, not affective trust.) It is in fact more plausible that the Testimony Game involves trust of the rich kind which is governed by social norms, and so it is unclear what, if anything, Faulkner has to gain by beginning with the Trust Game.