On the benefits and boundaries of trust and trustworthiness

Schutter, M.

Chapter 1

Introduction

Trust has been described as the glue or social lubricant that binds people together (Arrow, 1974). This description highlights the function of trust in social relations. The 'glue or social lubricant' metaphor suggests a major role in keeping people together, but also in letting things run more smoothly. Trust has been demonstrated to have major benefits on both accounts and in a variety of social relationships, such as initiating and maintaining mutual cooperative relationships (Deutsch, 1958; McKnight et al., 1998; Yamagishi, 2011), stimulating a nation's economic growth (Knack & Keefer, 1997; Labonne & Chase, 2010; Zak & Knack, 2001), and strengthening people's norms in favor of cooperation and inclusion (Balliet & Van Lange, 2013; Putnam, 1993). It is therefore not surprising that the topic of trust has been widely studied across disciplines, such as economics and psychology.

So, what are the key ingredients of the glue or lubricant? In other words, what are the key characteristics of trust that bind us together and let things run more smoothly? Although everyone will intuitively have a sense of what trust is, it is less clear what the prerequisites of trust are. A widely accepted definition describes trust as "a psychological state comprising the intention to accept vulnerability based upon the positive expectations of the intentions or behavior of another" (Rousseau et al., 1998, p. 395). This definition captures the interpersonal aspect of trust but also captures that trust is based on positive expectations regarding others' behavior. It suggests that in order to trust others, people need to have the positive expectation that others will be trustworthy.

The question then is, what may drive such positive expectations? It is here, that theorizing on trust connects to the concept of reciprocity. Basically, the concept of reciprocity presumes that people respond positively to others' positive behaviors and negatively to others' negative behaviors (Falk & Fishbacher, 2006; Gouldner, 1960). A positive act like giving trust may be rooted in reciprocity expectations. This is also exemplified by Krueger (2021, p. 1), who described trust as "a trustor's investment of resources (e.g., money, time, energy) into another party (i.e., a trustee) that encompasses uncertainty regarding the benefits of reciprocation in the future". Scientific research also indicates that the behavior of the trustee is connected to the initial (trust) behavior of the trustor and therefore can be seen as a reciprocal act (e.g., Cox, 2004).

There are different scientific methods to assess trust in individuals. One way is to administer surveys and assess people's self-reported intentions to trust others or to be trustworthy. People in general, however, are not very accurate in predicting their own future behavior through self-reported intentions (Epley & Dunning, 2000, 2006). An important development in the field of trust research was the introduction of an experimental paradigm that included a behavioral measure of trust. This experimental paradigm was initially referred to as the 'investment game' but is now more often called the 'trust game' (Berg et al., 1995; for reviews, see e.g., Johnson & Mislin, 2011; Van den Akker et al., 2020). The trust game is mostly conducted in controlled laboratory settings. While the paradigm is used in a variety of forms, it typically entails an anonymous interaction between two persons who are simply called Person A and

Person B. Person A has some money that s/he can allocate between him/herself and Person B. Person A also has another option, and that is to let B allocate a higher amount of money. The decision to let B allocate the higher amount of money is taken as a measure of trust. How B then allocates this higher amount is taken as a measure of trustworthiness.

To illustrate, consider the following example of a trust game. Person A is provided with €20 and has two options: Person A can either allocate the €20 between him/herself and Person B (Option 1). Or Person A can choose to let Person B distribute the money between the two of them, with the beneficial consequence that the money will then be tripled to €60 (Option 2). If A chooses Option 1, this is considered a decision of 'no trust'. A's choice for Option 2, is seen as a decision of 'trust'. In the latter case, the amount of money (out of the €60) that B allocates to A is taken as a measure of 'trustworthiness'.

The trust game captures the main aspects of trust that fit the definitions of trust described above (Krueger, 2021; Rousseau et al., 1998): vulnerability and reciprocity. First, Person A (also referred to as the trustor) must accept vulnerability to possible exploitation by Person B (the trustee), who may or may not reciprocate the positive decision to trust. The clear structure also has the advantage of being easily comparable to other studies using the paradigm. Meta-analyses show a consistent pattern in the results (Johnson & Mislin, 2011; Van den Akker et al., 2020). For example, it has repeatedly been found that the vast majority of trustors is willing to trust and that most trustees act trustworthy in return.

Despite the obvious advantages of this paradigm, it should also be acknowledged that the trust game in its most basic form also imposes some restrictions. First, the trust game is highly suited to investigate how people react when they are trusted. But what if people are not trusted? The standard trust game does not allow us to investigate this question – simply because the game ends when Person A decides not to trust B. Second, the standard trust game presents participants with a situation that is very conducive to evoke reciprocal behavior. Participants usually know with certainty what the revenues of trust are (in the example above, the revenues are €60 upon the decision of trust), and they know that Person B will respond immediately after Person A. But what if people are uncertain what the revenue trust will be, and what if B cannot respond immediately? Under such conditions, will people still be that willing to trust? And will these trustees be just as trustworthy? In the current dissertation I aimed to address these unanswered questions by adapting the standard trust game paradigm in a series of experimental studies.

Overview of this dissertation

This dissertation contains three empirical chapters (Chapters 2–4) in which I modified the traditional set-up of the trust game to learn more about the importance of trust, as well as the boundary conditions of trust and trustworthiness.

Chapter 2 – Part I. Chapter 2, which focuses on (reactions to) decisions of no trust, consists of two parts. In Part I of Chapter 2 (published as Schutter et al., 2021), I present two experimental studies that modified the trust game to examine the consequences of no trust. More specifically, I focused on the effects of *not* being trusted. All the research using standard trust games cannot inform us on how people react to not being trusted. As such, we know a lot about the behavior of those who are trusted, but not about those who are not trusted, simply because if a Person A decides not to trust B, the game stops. This knowledge gap is unfortunate. After all, although research using trust games has revealed that the majority of participants decides to trust, it also shows that a non-negligible part of them chooses not to trust (Fetchenhauer & Dunning, 2009; Johnson & Mislin, 2011).

Studying how people react when they have not been trusted is important to paint a more complete picture of trust – it is informative to see what happens if the glue is not there. In trust games the game may end after no trust, but that does not mean that there are no consequences for those who have not been trusted. In Experiments 2.1 and 2.2, I studied the emotions and affective responses of participants (as Persons B) who learned that they had not been trusted by Person A. In Experiment 2.2, I added a behavioral option for the participants who were not trusted by letting them decide on an allocation in a subsequent dictator game. This allowed me to investigate a form of negative reciprocity (Falk & Fischbacher, 2006; Fehr & Gächter, 2000; Gouldner, 1960), a type of reciprocity that cannot be studied in the standard trust game. By letting participants play a dictator game with the Person A who had not trusted them, Persons B could retribute by allocating A low outcomes.

I also adapted the paradigm to allow for another form of reciprocity, generally referred to as 'paying it forward', in which negative reciprocal acts may also be directed at uninvolved others (Cardella et al., 2019; Gray et al., 2014). For this purpose, participants who learned that A had not trusted them subsequently played a dictator game with a person who had not been involved in the initial no-trust interaction.

The modifications described above enabled me to study the aftermath of no trust. In addition, I also modified the game in such a way that I could distinguish an active decision of no trust from an inactive decision of no trust. It can be argued that in the typical trust game, a decision not to trust is inactive, as A does not actively take away B's possibility to prove him/herself trustworthy. By the decision of Person A to distribute the outcomes him/herself, A does not provide B with the opportunity to divide a trust revenue. It is possible, however, to envisage a more active decision of no trust, e.g., when A *actively* decides to take away B's opportunity to act trustworthy. This could occur, for example, if B initially would be the one to allocate the money, but A decides to take away this possibility. The distinction is important, especially when it comes to how people react to no trust. The difference between active and inactive distrust is closely linked to the broader body of research on omissions and commissions (e.g., Kahneman & Miller, 1986; Kahneman & Tversky, 1982; Ritov & Baron, 1990, 1992; Spranca et al., 1991). Omissions describe decisions of inaction (e.g., not helping someone who is hurt),

whereas commissions usually involve decisions of action (e.g., hurting someone). Active decisions have in general been found to be more impactful than inactive decisions. I therefore examined whether this also applied to reactions to actively versus inactively not being trusted.

Chapter 2 – Part II. In Part I of Chapter 2, I focused on how people (i.e., Persons B) react when they learn that they were – actively or inactively – not trusted. In Part II, by contrast, I focus on Person A, and examine whether their willingness to trust is affected by whether the no trust decision is active or inactive. If people anticipate that active distrust is more impactful than inactive distrust, they might be more reluctant to actively distrust Person B. To study this, I used the set-up of Experiment 2.1, but now all participants were assigned to the role of Person A.

Chapter 3 – Part I. This chapter starts with the observation that the high levels of trust and trustworthiness in trust games are generally obtained in a setting that is highly conducive to reciprocal behavior. In particular, I note that in the traditional trust game the revenues of giving trust are exactly known to both the trustor and the trustee. For example, Person A knows that when s/he chooses to trust, B will divide €60. This also means that in the end, A will know exactly how much B kept and how much B was willing to allocate to A. The available evidence suggests that under these circumstances, A is often willing to accept the vulnerability to trust B, and that B often reciprocates trust with trustworthy behavior (i.e., high allocations to the trustee; see the meta-analyses of Johnson & Mislin, 2011, and Van den Akker et al., 2020). But what would happen if A could not assess whether or to what extent B was showing reciprocal behavior? Would trustees still be as trustworthy?

To study this, I modified the trust game by including unexpected trust revenues. Focusing on trustworthiness, all participants were assigned to the role of Person B, and all learned that Person A had trusted them to allocate a multiplied amount of money. But then participants also learned that the initial amount was not tripled (as initially stated), but unexpectedly doubled (i.e., a lower-than-expected trust revenue) or quadrupled (i.e., a higher-than-expected trust revenue). Moreover, this information was either revealed to both Person A and Person B, or concealed from A (the trustor) and thus only provided to the participants in the role of Person B (the trustee). In this latter case, participants thus had an information advantage. My interest was in how this would affect their trustworthiness, as measured by their allocations to the trustor.

I reasoned that it might affect their allocations because having an information advantage may offer people some "moral wiggle room" which allows them to feel and/or appear fair to others while actually furthering their own outcomes (Dana et al., 2006). In Experiment 3.1, I therefore investigated whether trustees would be tempted to allocate trustors a lower amount of money when they believed that the trustor was not informed about an unexpected increase in revenue. In Experiment 3.2, trustees again faced an unexpected lower or higher trust revenue, but they now had the choice to inform the trustor about the unexpected change in revenues.

This set-up provided trustees with the option to be transparent about the information they had, but also to leave the trustor in the dark about the change in revenues. This allowed me to test whether trustees would especially be tempted to withhold such information when trust revenues would unexpectedly be higher. The self-created moral wiggle room would then allow them to make lower allocations.

In Experiments 3.1 and 3.2, participants knew that the change of trust revenue was not anticipated by A. The change was introduced and implemented after the trustor had decided to trust. As such, Person A had no reason to believe that the revenue might change after deciding to trust Person B. Would reactions be different when participants thought that the trustor knew upfront that it was not certain that the money would be tripled? In Experiment 3.3, I examined whether participants – again in the role of Person B – thought that the trustor had been aware that the trust revenue was uncertain. For this purpose, participants learned that when A made the decision to trust, A was informed that the trust revenue was yet uncertain and could in the end be doubled, tripled, or quadrupled.

Chapter 3 – Part II. In Part I of Chapter 3, I focused on how trustees (i.e., Persons B) dealt with unexpected trust revenues and whether they used the moral wiggle room that an information advantage creates. While the results indeed supported this idea, Part I did leave two questions unanswered. The first is whether the findings are indeed best explained by a process of egocentric self-justification in which trustees use their information advantage to justify their self-interested behavior. The second is whether people (i.e., Persons A) would be willing to trust others in settings where the trust revenue would be uncertain.

In Experiment 3.4, I addressed the first question by examining how the observed behavior (i.e., allocations) of trustees would be evaluated by uninvolved others. For this purpose, I provided participants with a description of the set-up of Experiment 3.1. Participants evaluated different allocations as made by trustees (varying from very unequal to very equal) as well as evaluated the trustees who made the allocations. In Experiment 3.5, I studied the willingness to trust under uncertainty, as operationalized in Experiment 3.3. All participants were assigned to the role of Person A and had to decide whether they would trust Person B, while being uncertain about the trust revenue (i.e., upon the decision to trust the amount of money could be doubled, tripled, or quadrupled). The exact multiplier would eventually be revealed to Person A (but only if and after A had chosen to trust), or concealed from Person A. In the latter case, A would only learn how much Person B allocated to him/her, but not out of which total amount.

Chapter 4. In this chapter, I again start with the observation that the high levels of trust and trustworthiness in trust games are generally obtained in a setting that may be highly conducive to reciprocal behavior. This time, I focus on the fact that in the typical trust game, the decisions of trustors and trustees are closely connected in time; once the trustor has made her decision, the trustee responds immediately. Such a close temporal connection

may foster reciprocal behavior (e.g., Frederick et al., 2002), and thereby facilitate trust and trustworthiness. But what if it takes some time for the trustee to respond? Are trustors then still as willing to trust and are trustees then still so trustworthy? In an incentivized experiment, we examined both trustors and trustees who were either immediately coupled in time (i.e., Person B decided right after Person A) or faced a one-month interval between their decisions.

Chapter 5. This final chapter provides a review of the empirical work as reported in Chapters 2–4. It provides a conclusion and discussion of the separate chapters, but also provides an integration of the findings of this dissertation. Building on these insights, this chapter also identifies possible avenues for future research on trust and trustworthiness, and their relationships with reciprocity.