Why do people sometimes not keep their promises?[*]

Giovanni Di Bartolomeo,[†] Stefano Papa[‡] and Francesco Passarelli[§]

April, 2014

Abstract

This paper investigates two theories that account for promise keeping. The first theory is motivated by guilt aversion when people dislike to let others' expectations down; the second theory argues that promises generate a sense of moral obligation, which is independent of others' expectations. A crucial aspect for testing these theories is understanding how communication yields to a promise and how the latter shapes expectations. We show that flawed promise definitions may lead to biased interpretation of the experimental data. Moreover, we analyze in details the belief formation process and individuate the effects of cognitive biases.

By defining a promise in a way that is consistent to data, we find that making a promise is not sufficient to make people feel a moral obligation. Sometimes people can go back on their promises. It is necessary that the one who promises also believes that he raised the partner's expectations. Only in this case, she feels guilty and honors her obligation.

Keywords: Cheap talk, promises, moral obligations, cognitive biases, guilt aversion, beliefs, psychological games.

[*] The authors are grateful to XXXX for useful comments on earlier drafts. Giovanni Di Bartolomeo acknowledges financial support by Sapienza University of Rome. The authors also thank G. Cresci, V. Di Simone, G. Orati for assistance.

[†] Department of Economics and Law, Sapienza University of Rome, via del Castro Laurenziano, 9 – 00161, Rome, Italy, email: *giovanni.dibartolomeo@uniroma1.it.*

[‡] Department of Communication, University of Teramo, email: *spapa@unite.it.*

[§] Department of Communication, University of Teramo, email: *fpassarelli@unite.it* and Bocconi University, Milan.

1 Introduction

Why do people keep promises? And why do people sometimes "forget" to keep them? There are two theories which emphasize the role of communication among individuals. The first one states that those who make promises dislike letting others down. This psychological pain is caused by guilt, when the promise raises others' expectations. The second theory looks at promises as a moral obligation. Promises in the Kantian philosophical tradition are not interpersonal, they define moral obligations that one owe to everybody under all circumstances and mainly towards herself. Thus people keep promises in order to bear the psychological cost of a reduced self-esteem. Both theories argue that promises are relevant, but only the second one assumes that promises are important *per se*, independently of their impact on others' expectations. When trying to connect communication with behavior, a crucial aspect is determining if communication gives rise to promise or not. One may expect that only when communication ends up to an engagement to follow a possible agreement up, then future behavior is affected. In this paper we test the two theories by looking at how communication shapes promises, and then future behavior. This has already been done, among others, by Charness and Dufwenberg (2006) and Vanberg (2008), who look at one-way and two-ways communication, respectively.[1]

Charness and Dufwenberg find evidence that promises resulting from *unilateral* messages raise expectations, and thus claim that promise keeping by those who make promises is due to their guilt aversion. Vanberg observes that the correlation between promises and expectations is consistent with both theories. Using an alternative design he finds that only the second theory holds: promise are kept because of a moral obligation, rather than guilt aversion. However, despite two-ways communication, Vanberg assumes that a *promise* arises only when one party has made it, independently of what the other party said. Arguably, the relationship between promises and expectations in two-ways communication is more articulated. Through proposals and counter-proposals the parties tend to reach some form of agreement. Then a promise is possibly the result of an articulated, bidirectional communication process leading to an agreement, rather then of a simple unilateral proposal. In this case a promise is a non-binding commitment to keep the agreement. Vanberg does not consider this point, which however is an important one.

We claim that the impact of communication on expectations is different when only one party sends a promise rather than when an agreement

[1] Experimental evidence on the positive effects of promises in trust games is provided by, e.g., Ellingsen and Johannesson (2004), Charness and Dufwenberg (2006, 2010), Vanberg (2008), Ben-Ner and Putterman (2009), Ben-Ner *et al.* (2011).

has been reached. In fact we find evidence that only agreements raise expectations. The emotional feelings arising from communication arise only when a bilateral agreement is reached. Thus people's expectations arise only when an agreement occurs, and only in this case a possible feeling of guilt or moral obligations emerge. Broadly speaking, a unilateral proposal that is not accepted is simply a bubble, and does not give rise to any feeling. This interpretation is consistent with Ben-Ner and Putterman (2009).

Aiming at testing the two theories, Vanberg (2008) switches some subjects randomly after communication takes place and compares the behavior of an agent when she has to choose whether to keep her own promise or not, with her behavior when she has to choose if keeping a promise made by another agent. He finds that, all other things being equal, only in the former case promise keeping occurs. Thus, he claims that people keep promises only because they care about their own commitments, and not because they feel guilty.

A crucial aspect of his design and in general for testing these theories is understanding how communication yields to a promise and how the latter shapes expectations. In order to compare the effects of own promises and promises made by others, he correctly assumes that expectations of people who have received a promise are independent of being switched or not if they ignore it. However, he omits to verify this assumption. We do it and find significant differences for both his and our (using his definition) data. This proofs that the Vanberg's (2008) definition of a promise is flowed. By contrast, if one defines, as we do, a promise as a commitment to keep an agreement, this inconsistency does not emerge. By a definition of promises consistent with the data, we find a different result: people keep their own promises, but only because these promises lead to high expectations. Thus, promise keeping is due to guilt aversion.

Specifically, we find that people do not always keep their own promise. They do it only when the promises raise other's expectations. Instead people do not care of a promise made by another guy, even when that promise has lead to high expectations. Moreover, we find that people evaluate others' expectations systematically smaller when the promise has been made by another guy. We claim that this systematic difference is due the fact that "reading" own promises is a different task compared to interpret messages written by others. This occurs because in the latter case subjects did not actively participate to the communication process. Finally, follow an intuition of Vanberg (2008), we test if cognitive biases are at work when people interpret promises made by others. We consider two kinds of biases: false consensus effect and the confirmatory bias. The false consensus effect occurs when a person tends to overestimate the number of the people agree with her. The confirmatory bias is related to how subjects interpret new information: they tend to interpret ambiguous evidence as supporting their existing beliefs. We find evidence of a false consensus effect, but we cannot exclude that a confirmatory bias also matters.

The rest of the paper is organized as follows. In the next section we develop a simple model that describes the communication process, i.e. how communication yields to a promise and how the latter shapes expectations. Section 3 describes the experiment design and the procedure used. Section 4 shows our main results; some conclusions are entrusted to the last section.

2 Belief formation, signals and cognitive biases

Beliefs play a crucial role in our experimental design; therefore to formally discuss communication and belief formation, this section develops a simple, general model that describes the communication process as a signal extraction game. It describes our ideas about second-order beliefs formation and how some cognitive bias may affect them. We assume that reading own messages can be a different task compared to interpret messages written by others. This occurs because in the latter case agents did not directly participate to the communication process that has/or not has generated an agreement. Thus we consider that the task of switched agents can be different from that of no switched one.

As stressed by Rabin and Scharg (1999: p. 37), psychological research also indicates that people have a cognitive bias that leads them to misinterpret new information as supporting previously held hypotheses. Cognitive biases may have a relevant role for our experiment as also stressed by Vanberg (2008). In particular, we consider two biases: false consensus effect and the confirmatory bias. The former consists in the fact that people to overestimate the extent to which others agree with them.[2] The second type of bias is not tied to what agents do, but how they interpret new information in terms of what they did. We refer to this kind of cognitive bias as a confirmatory bias.[3] It implies that subjects tend to interpret ambiguous evidence as supporting their existing choices or beliefs. Since in our context, agents must form expectations about what others expect, both biases seem to be potentially relevant.

[2] It has long been recognized by psychologists since Ross *et al.* (1977).

[3] It is worth noting that in psychology and experimental economics, the term "confirmation bias" is used but with several slightly different meanings.

We assume two equally sized groups (A and B) and consider a simple two-stage game (Game Γ^1). The game can be described as follows. In the first stage, players of the two groups are randomly paired and can communicate. In the second stage, payoffs of paired subjects only depend on the behavior of agent A. We assume that *i)* second-order beliefs are affected by communication in the first stage; *ii)* in stage 2, player A will choose according to her second-order belief determining the game payoffs as her preferences depends on the her beliefs formed during the game—along the lines of psychological game theory.[4] These assumptions imply that communication is effective and guilt aversion explains the agents' behaviors.

The game can be summarized by two states of the world, $\theta \in \{\theta_C, \theta_D\}$, where θ_C and θ_D indicates high or low second-order beliefs after the communication ($\theta = \theta_C$ if subjects form an agreement that implies cooperation "C" during this stage; $\theta = \theta_D$ i.e. disagreement "D", otherwise). We assume that prior beliefs at the beginning of the game are $prob(\theta = \theta_C) = \sigma$ and $prob(\theta = \theta_D) = 1 - \sigma$. The second-order beliefs associated with the two states are C in θ_C and $D < C$ in θ_D.

The outcome of the game is trivial. If agent A forms an agreement during the communication, she knows that player B's first-order belief is C and $prob(\theta = \theta_C) = 1$; otherwise she knows that the B's first-order belief is D and $prob(\theta = \theta_C) = 0$. As agent A observe the outcome of the communication, priors are useless.

Now we consider a variant of the above game where players A are paired with new partners after the communication stage (Game Γ^2). Therefore, they should form second-order beliefs about partners who have not communicate with them. We assume that they can read the communication of the new partners. It is worth noticing that reading own messages can be a different task compared to interpret messages written by others; thus, by using Bayesian updating, we model the information about the new partner's as a signal which is correlated to the state of the world. Formally, agents A read the new partner's messages as an agreement or not $m \in \{m_C, m_D\}$ and know that: $prob(m = m_C|\theta = \theta_C) = p > 1/2$ and $prob(m = m_D|\theta = \theta_D) = q > 1/2$.

The agents then revise their prior about the states of the world in line with the received signal:

$$\begin{aligned} prob(\theta &= \theta_C|m = m_C) = \frac{p\sigma}{1 - p\sigma\,(2p-1)} && (1) \\ prob(\theta &= \theta_D|m = m_D) = \frac{q\sigma}{1 - q\sigma\,(2q-1)} && (2) \end{aligned}$$

[4] See Geanakoplos *et al.* (1989) and Battigalli and Dufwenberg (2007, 2009).

Observe that $prob(\theta = \theta_C|m = m_D) = 1 - prob(\theta = \theta_C|m = m_C)$ and $prob(\theta = \theta_D|m = m_C) = 1 - prob(\theta = \theta_D|m = m_D)$.

Defining EC (ED) as the second-order belief of an agent who received a m_C (m_D) signal, it follows that $EC < C$ and $ED > D$. Thus agents A involved in game Γ^1 and Γ^2 have different second-order beliefs and differences come from different information sets. Thus they are fully rational.

Proposition 1 *Second-order beliefs in* Γ^1 *and* Γ^2 *may be different: for* $p < 1$, *second-order beliefs of subjects observing cooperation in* Γ^1 *are larger than those in* Γ^2*; for* $q < 1$, *second-order beliefs of subjects observing no cooperation in* Γ^1 *are smaller than those in* Γ^2*.*

Let's us now revise Γ^2 assuming that agents may be subjected to two cognitive biases (game $\tilde{\Gamma}^2$): false consensus and confirmatory bias. The former means that agents consider their own choice more representative than those of randomly chosen others; their beliefs are thus biased towards their own behavior or opinions. It follows that an agent who forms (or does not form) an agreement during the communication stage assumes that other agents are more likely to behave as she did. A confirmatory bias, emphasized by Rabin and Scharg (1999), may instead emerge when a subject tends to misinterpret ambiguous evidence to confirm her own belief about the world, i.e. people tend to interpret ambiguous evidence as supporting their existing position. These bias imply that agents may have different beliefs even if they have the same information set.

The false consensus effect is formalized as a different prior assigned to $prob(\theta = \theta_C)$ between agents who formed and those who did not form an agreement during the communication stage. We assume that $prob^C(\theta = \theta_C) = \sigma + \delta_C > \sigma$ and $prob^D(\theta = \theta_C) = \sigma - \delta_D < \sigma$, where the subscripts indicate agents who formed (C) and those who did not form (D) an agreement. It follows

$$prob^i(\theta = \theta_C|m = m_C) = \begin{cases} \frac{p(\sigma+\delta_C)}{1-p(\sigma+\delta_C)(2p-1)} & i = C \\ \frac{p(\sigma-\delta_D)}{1-p(\sigma-\delta_D)(2p-1)} & i = D \end{cases} \tag{3}$$

A false consensus effect clearly implies that second order beliefs of agents who received the same message differs according to the outcome of the communication stage; i.e., $EC^C > EC^D$ as $prob^C(\theta = \theta_C|m = m_C) > prob^D(\theta = \theta_C|m = m_C)$. Similar results hold for $prob^i(\theta = \theta_C|m = m_D)$ with $i \in \{C, D\}$.

The confirmation bias is modeled by considering that after receiving the same signal, agents interpret it differently according to their previous choices.

We assume that $prob^C(m = m_C|\theta = \theta_C) = p_C \geqslant prob^D(m = m_C|\theta = \theta_C) = p_D > 1/2$, where again subscripts indicate agents who formed (C) and those who did not form (D) an agreement. It follows that

$$prob^i(\theta = \theta_C|m = m_C) = \begin{cases} \frac{p_C(\sigma+\delta_C)}{1-p_C(\sigma+\delta_C)(2p_C-1)} & i = C \\ \frac{p_D(\sigma-\delta_D)}{1-p_D(\sigma-\delta_D)(2p_D-1)} & i = D \end{cases} \tag{4}$$

i.e., $prob^C(\theta = \theta_C|m = m_C) \geqslant prob^D(\theta = \theta_C|m = m_C)$. This bias also implies that $EC^C > EC^D$ as again $prob^C(\theta = \theta_C|m = m_C) > prob^D(\theta = \theta_C|m = m_C)$.

The following proposition can be stated from confirmatory bias and false consensus effect.

Proposition 2 *Second-order beliefs in* $\tilde{\Gamma}^2$ *may differ between subjects who decided to form an agreement during the communication phase and those who do not because of either a false consensus effect or a confirmatory bias.*

It is worth noticing that although both biases have the same effects on second-order beliefs, but they affect the beliefs at different time. Specifically, false consensus bias operates before reading the new partner's communication, whereas the confirmatory bias operates after.

Summarizing, we have shown how rational agents may have different second-order beliefs (and thus follow different choices) assuming that they read they own communication or that be (Proposition 1). Moreover, by introducing some cognitive bias we show how the agents' second-order beliefs can be anchored to their initial choices even if those are irrelevant for their expectations (Proposition 2). The belief anchoring can occur after the agents choices (false consensus effect), after receiving new information (confirmatory bias) or both.

3 Experiment description

3.1 Design

We consider the mini dictator game with random dictatorship and pre communication proposed by Vanberg (2008). The game can be illustrated by considering two stages: *i) communication* and *ii) action*. During the communication stage, subjects can communicate by sending of two alternate messages in a chat. The action stage is a random dictator game. More in details, at the beginning of the communication stage, N subjects are matched in pairs. Each subject can communicate with the partner. At the end of this stage, for each pair, Nature decides if players change their partners (i.e., they are switched) or not. Then the Nature randomly determines the role of the players (i.e., dictator or recipient). Both subjects know that they can be switched with probability $\frac{1}{2}$, but only the dictator observes the Nature's choice. If switched, the dictator can read the chat of the new partner. In the action stage, the dictator must choose between two actions, "roll" or "don't roll." If the dictator chooses the latter, she receives 14 tokens and the recipient (old or new partner) receives nothing. If the dictator instead chooses to roll, she receives 10 tokens and the recipient receives 12 tokens with probability 5/6 and nothing with probability 1/6. The second stage of the game replicates the payoffs of Charness and Dufwenberg (2006).

The payoffs of the random-dictator game are summarized in the figure below (borrowed from Vanberg, 2008).

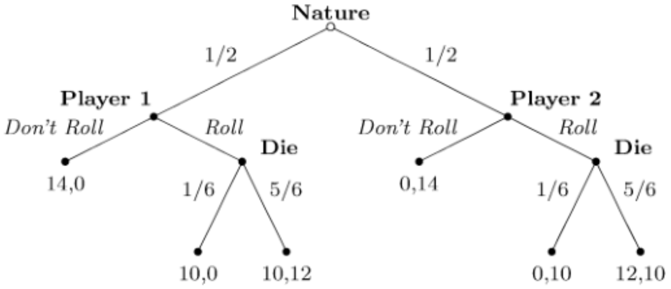

Figure 1 – Mini-dictator game with random dictatorship.

Once that messages are classified as a promise or not (see next subsection), our setup can be easily used to test the communication effects and test the two explanations for promise keeping (see subsection 3.3) eliciting first- and second-order beliefs (see subsection 3.4).

3.2 Message classification

Charness and Dufwenberg (2011) point out that free-form messages can be classified in a variety of ways. Here, for the sake of comparison, our setting is as close as possible to Vanberg's (2008). We consider a dichotomous classification (promise and no promise) and treat each pair of messages sent by a single subject as a unique message.

In our sample, the typical message (proposal) is: "Let's agree to roll if either of us is drawn as the dictator." The reply (counterproposal) to this kind of message may consist in an agreement or a dismissal. Vanberg (2008) evaluates the proposal as a promise independently of the counterproposal. In evaluating messages, we depart from him by considering both sides of communication in line with Ben-Ner and Putterman (2009). Specifically, we classified a message as a promise when, not only a proposal of cooperation was made, but also the latter was followed by an agreement. Moreover, message interpretation is clearly a subjective activity, we attempt to reduce this subjectivity by involving several evaluators.[5]

As we will show in more details, the core of our design is a comparison between the group of those who have to choose if keeping her own promises or not, and the group of those who have to choose if keeping a promise made by someone else. Then the definition of promise is crucial. Biases can derive from flawed promise definitions by the experimenter, which may lead to a wrong interpretation of the experiment results as we show in the example below (see Table 1).[6]

In particular, one can define a promise either as a unilateral proposal or as the result of proposal and counterproposal. Suppose that the first definition is correct (i.e., unilateral proposals lead to an increase in expectations) and the second one is wrong. However, classifying messages according to the second (wrong) definition would imply that the two groups' average beliefs are unbiased. The reason is that bilateral agreements are a subset of unilateral claims.[7] By contrast, suppose that the second definition is correct while the first one is wrong. In this case, using the first (wrong) definition to classify messages might lead to biases, since average beliefs now depend on draws. In fact, if in one of the two groups, by chance, there was a high number of misinterpreted promises (i.e., unilateral proposals), then average beliefs of that groups would be low. Thanks to this inconsistency, possible flaws in promise definition can be tested.

Let's provide an example. In Table 1 couples on the same row engage in a chat. Assume that a chat affects beliefs only when it leads to a promise. Suppose, as we do, that a promise when there is a proposal and a counterproposal to cooperate. Then the beliefs of subjects in each couple are those reported in the last column.

[5] Each evaluator assigned a value of one to a proposal (or counterproposal) for an agreement and zero otherwise. Thus, we consider a promise a "1+1" classification. According to Vanberg a promise is "1+1" or "1+0."

[6] Note that, differently from flawed definitions, evaluators' subjectivity equally affects the interpretation of messages of subjects of both groups, then it should not alter the comparison on average.

[7] Of course, in this case, a problem may arise if one compare beliefs of people who made a promise to those who did not.

Table 1 – Communication and beliefs

	Message		Message	Beliefs
Al	Let's agree	Dorothy	Let's agree	1
Bob	Let's agree	Eloise	No, thanks	0
Carl	No, thanks	Fanny	No, thanks	0

If the promise definition is flawed then first-order beliefs depend on draws. This may lead to a logical inconsistency.

Suppose that men are drawn as recipients in the non-switched case and women subjects are drawn as recipients in the switched one. If the promise definition is correct, no difference among beliefs of switched subjects and non-switched ones, conditional to receiving a promise, will occur on average. In this case, Al's belief (i.e. the only switched man who received a promise) is one, and Dorothy's belief (i.e. the only switched woman who received a promise) is one, as well. Thus if our promise definition is correct, data should be consistent with this prediction.

If instead the promise definition is flawed, then also Bob's message is classified as a promise. Al's belief (i.e. the only switched man who received a promise) is one, but Dorothy and Eloise's average belief (i.e. the two switched woman who received a wrongly classified promise) is one half. It is apparent that using a wrong promise definition, a difference between average beliefs of switched subjects and non-switched ones, conditional to receiving a promise, may occur as a result of random draws, which is clearly a logical inconsistency.

Finally, note that, assuming a flawed promise definition, a difference in the first-order between the two groups (e.g., average non-switched first-order beliefs larger than those of switched ones) determined by the draw should also implies an opposite difference in the second-order one (e.g., switched second-order beliefs larger than those of non-switched subjects). However, if reading messages from other is a different task, second-order beliefs can be the same. The high expectations of non-switched may be in fact underestimated by dictators who read promise made by other people.[8]

Given a promise definition and the corresponding message classification, we can test the above problem by comparing the (average) first-order beliefs between switch and no switch subjects who received a promise. Since all of them ignore if they were switched or not, averages should be equal. If this is not the case, promise definition is not consistent with the data, and thus, it is flawed. If data do not pass the test, then the way we defined a promise is flawed—the contrary is not necessarily true.

[8] As we will show, this seems to explain the Vanberg's (2008) outcomes.

3.3 Hypotheses

As discussed above, before investigating the two theories for promise keeping, we test the consistency of our definition (and thus message classification), according to which a promises is a proposal and a counterproposal to cooperate (Ben-Ner and Putterman, 2009). We also test the consistency with the data of a definition based on a unilateral claim to cooperate as Vanberg (2008). Formally, we verify this consistency by comparing the (average) first-order beliefs conditional to a promise between switch and no switch sub-samples. We expect to find equal averages.

Evidence for promise keeping is then tested by comparing the proportions of roll of dictators who have formed an agreement (promise) to the proportion of roll of the other dictators (no promise) in the sub-sample of non-switched pairs. If found, this evidence can be motivated by either an expectation- or a commitment-based theory (indeed, even by both). Of course, by the former we refer to an explanation based on guilt aversion and by the latter to one that refers to moral obligation. As shown by Vanberg (2008), motivations for promise keeping can be investigated by using the second-order beliefs and the switched dictator sub-sample. If the two groups of non-switched and switched dictators observing an agreement have the same averages in second-order beliefs, then a difference in the roll proportions between these two groups can be interpreted as evidence for commitment-based explanation for promise keeping.

Second-order beliefs the above two groups will be the same when subjects are rational (no cognitive biases) and there are no problem in message interpretation, as in the Game Γ^1 described in Section 2. But it is not true in general. Thus, investigating the motivations for promise keeping, it should be account for the second-order beliefs formation and the testable assumptions related to this process.

As shown, second-order beliefs between switched and non-switched dictators may be in fact different even agents are rational as in game Γ^2 (see Proposition 1). A difference in the average second-order beliefs of non-switched and switched dictators observing an agreement could provide evidence for subjects' difficulties in message interpretation. Different second-order beliefs may simply reflect the fact that different capabilities in reading signals associated to the messages between dictators who evaluate their own promise (no switched) and those who evaluate a promise written by others (switched). Everything equal, second-order beliefs of switched dictators who observe an agreement should be lower that those of non-switched dictators.

Furthermore, as second-order beliefs may also be affected by cognitive biases. In such a case, beliefs of switched dictators may also dependent on their own messages (see Proposition 2). Considering in the switched subsample, differences in the average second-order beliefs explained by *coeteris paribus* differences in the results of their previous communication will provide evidence for cognitive biases. It is worth noticing that switched dictators known that they have been switched and thus that their messages do not affect the new partner expectations as they have not read them. To investigate the eventual nature of a cognitive bias, we extend the approach of Vanberg (2008) by considering an additional treatment (T2).

The treatment T2 is the same of T1; i.e., a mini dictator game with random dictatorship and pre communication. It only differs in the dictators' second-order beliefs elicitation. In T2, in fact, second-order beliefs are collected after dictators are informed about their role and if they are switched of not, but before they can read the communication of the new partner. As a result, second-order beliefs of non-switched dictators and first-order beliefs are as in T1, but second-order beliefs of switched dictators are collected under a different information set. If in T1 we observe a cognitive bias in second-order beliefs of switched dictators, by using the data from T2 we can verify if this bias emerges before reading new partner's chat (false consensus effect) or after (confirmatory bias).

Formally, we can index second-beliefs as follows.

Definition 3 (Second-order beliefs) *We indicate average second-order beliefs of dictators by using three indexes: Z, i, j (i.e., Z_i^j). By $Z \in \{S, N\}$ we individuate dictators switched or not; by $i \in \{C, D\}$ we mean that dictators have formed an agreement during the communication (C) or not (D); similarly, by $j \in \{C, D\}$ we indicate if, in the action stage, the dictators have been matched with recipients who have formed an agreement during the communication.*

Then we can test the assumptions supporting of Game Γ^1, i.e.

$$\begin{cases} S_D^C = S_C^C = N_C^C \\ S_D^D = S_C^D = N_D^D \end{cases} \tag{5}$$

and game Γ^2

$$\begin{cases} S_D^C = S_C^C \\ S_D^D = S_C^D \end{cases} \tag{6}$$

In game $\tilde{\Gamma}^2$ neither (5) or (6) is instead expected.

3.4 Procedures

The experiment was conducted at Sapienza University of Rome. Undergraduate students were recruited by e-mail using lists compiled in advance by using University mailing lists and advertisements placed on the University notice boards and participants were randomly selected from the database. We verified that there was no repeat participation of students who can not use twice their matriculation number to enroll. At the beginning of each session, subjects were required to provide their registration cards. We ran 9 sessions, 6 for treatment (T1) and 3 for treatment (T2), with 32 subjects participating in each session, for a total amount of 288 participants. Subjects were randomly assigned to 32 visually isolated computer terminals of laboratory. Instructions were distributed by experimenter and two controllers who checked that the instructions were correctly followed by participants. Questions were answered individually at the subjects' seats. With the start of the experiment, subjects filled out a short questionnaire testing comprehension of the rules and of the game and they insert their age and gender information. Each session consisted of 8 rounds, with perfect stranger matching. After the experiment, one of the 8 rounds was chosen for payment of the decision. For elicitation of beliefs, we gave an extra incentive in tokens, as later explained. Beliefs were paid in all rounds except the one chosen for payment of the decision. All subjects received a fixed participation fee of 2.50 tokens.[9] The experiment was conducted with the software z-Tree (Fischbacher, 2007). The procedure is described in more detail below.

In each round 32 subjects were randomly matched to form 16 chatting pairs. Subjects were chosen randomly to start to send two alternate messages (a couple each one) to other of at most 90 characters. During the communication phase, no pair knew which of the two would be the dictator. After communication, one subject was randomly chosen as dictator. Then, half of all chatting pairs (i.e., 8) were randomly switched with other pairs, and the other half of pairs remain the same (no switch). No switched dictators remained with their chatting partners during the action phase. Switched dictators were matched with a new recipient from another pair. Prior to the action phase, subjects were informed to their role. Dictators were informed whether or not their partner had been switched and they could also observe the new partners' previous conversations. Recipients were informed only of their role, not whether their partners were switched or not.

Dictators were asked to roll or not, whereas recipients had to indicate their first-order beliefs (from the five point scale, see Table 2) about the choice of a feasible dictator to whom they were now matched. Each column

[9] The echange rate is 1 token equal to 1/2 euro.

is associated with payoffs that depend on the decision made by the partner. This procedure yields a five point scale for first-order beliefs.

Table 2 – Incentive for belief elicitation

The dictator will	Certainly	Probably	Unsure	Certainly	Probably
	choose roll			choose don't roll	
Earnings if the dictator					
chooses roll	0.65	0.60	0.50	0.35	0.15
chooses don't roll	0.15	0.35	0.50	0.60	0.65

Then, dictators had to indicated their second-order beliefs in order to guess the recipient's belief about their own behavior. If dictators guessed marking correctly the box in Table 2, they earned 0.50 tokens. This yields a five point scale for second-order beliefs. At the end of each round, payoffs were shown to both subjects that they would receive if the current round was chosen to payment. Recipients were not informed of the dictator's choice (and the number of dice rolled) or whether their partners had been switched. Feedback regarding the payoffs from guessing the other beliefs was given only at the end of the experiment.

4 Experiment results

We ran 6 experimental sessions to perform T1, each involving 32 subjects for 8 rounds. Each session represents one independent observation. We have a total of 768 dictator's decisions, equally splitted between "switch" and "no switch" condition. Three additional sessions have been run to perform T2. In T1, three research assistants (message evaluators) analyzed 1536 messages. They classified 994 messages either as a proposal or a counterproposal. The rest of the messages were bare messages. Among proposal or a counterproposal, they identified 394 messages as promises. According to our definition, promise occurs only when a proposal is made and counterproposal which is an agreement follows.[10]

We tested our promise definition. As pointed out earlier, since recipients do not know if they were switched or not, the first-order beliefs of those who received a promise must be equal on average, independently of being switched or not. If we found any differences in average first-order beliefs, then our definition would be flawed. But we did not. In our sample, average

[10] We randomly draw the evaluations made by one of the three assistants. We made the draw before of the experiment without informing the assistants. Correlations among classifications are however close to 0.99.

first-order beliefs are 0.65 and 0.62 in the case of no switch and switch, respectively. The averages first-order beliefs of recipients who did not receive a promise are 0.46 and 0.49 in the case of non-switch and switch, respectively. It is apparent that averages are not significantly different in both cases.

We used Vanberg's (2008) promise definition with our data and tested it. But we obtained statistically significant differences.[11] As said, this test shows that there are some flaws with Vanberg's promise definition. We found that the same problem arises in Vanberg (2008).[12]

The main outcomes of T1 are reported in Table 3 and 4. The first row of each table refers to dictators whose communication ended up with a promise; the second row refers to all other dictators. Columns indicate if dictators have been switched (columns 2 and 3) or not (1). When switched, columns indicate if the dictator's new partner ended up her communication with a promise (2) or not (3).

Table 3 shows average rolls. Under the no switch condition (column (1)), dictators who promised chose to roll in 59% of times. This fraction is significantly greater than the average 30% roll rate observed among all other dictators ($Z = 4.90$, $p = 0.00$). As expected, communication matters.

Proposition 4 *Dictators who agree to roll during the communication are more likely to roll than dictators who did not: People keep promises.*

Table 3 – Average rolls (T1)[13]

	No Switch	Switch observing	
Dictator's deal		promise	no promise
	(1)	(2)	(3)
promise	0.59	0.39	0.37
	(0.49)	(0.49)	(0.49)
no promise	0.30	0.35	0.30
	(0.46)	(0.48)	(0.46)

The above result is related to second-order beliefs. In fact, the first column of Table 4 shows that dictators who promised and were not switched also had significantly higher second-order beliefs, 0.77, than those who did

[11] We also test average first order beliefs of subjects who formed an agreement during the communication compared to those who decline an agreement proposal (i.e. interpreting a promise as the intention of a party only). We find a statistically significant difference.

[12] By using Vanberg's public available data in the Econometrica website, we checked that first-order beliefs of switched and no switched recipients are significantly different on average when recipients have played with dictators who has made them a promise. Averages are 0.63 in the switch case and 0.70 in the no switch case.

[13] In brackets we report the standard deviations.

not promise and were not switched, 0.49 ($Z = 6.72$, $p = 0.00$). This observed pattern mirrors the results of Charness and Dufwenberg (2006) and Vanberg (2008).

Table 4 – Average second-order beliefs (T1)[14]

	No Switch	Switch observing	
Dictator's deal		promise	no promise
	(1)	(2)	(3)
promise	0.77	0.68	0.54
	(0.29)	(0.33)	(0.37)
no promise	0.49	0.55	0.50
	(0.39)	(0.36)	(0.36)

The result is consistent with both the expectation- and commitment-based theory for promise keeping. People may keep their words either because they do not want to hurt the others (as they are guilt averse, dependently from second-order beliefs) or because they do not want to break their own promise (hurting themselves, independently of second-order beliefs).

As shown by Vanberg (2008), the two explanations for promise keeping can be evaluated by comparing the roll rates of no switched and switched dictators who made or observed a promise. From Table 2, we find that the roll rate of the no-switched dictators (59%) is significantly greater than the roll rate of the switched ones (39%) ($Z = 3.01$, $p = 0.00$). However, differenlty from Vanberg (2008), Table 3 tells us that expectations of no-switched dictators (0.77) are different than the switched ones (0.68) ($Z = 2.02$, $p = 0.04$)—as predicted by Proposition 1. Therefore, no evidence in favor of either theory emmerges from the data. Our result is summarized by the following proposition.

Proposition 5 *Reading own messages is a different task than interpreting messages written by others. Comparing average rolls of switched and not switched subjects cannot provide unequivocal evidence in favour of a specific theory for promise keeping.*

Table 3 shows a correlation between dictators' second-order beliefs and promises. This correlation is weaker when promises are made by others. This provides evidence of our claim that switched and not switched subjects face a different task in interpreting messages. Therefore, simple comparison between averages cannot be used to discriminate between the two theories. Our results differ from Vanberg (2008), who finds that second-order beliefs

[14] In brackets we report the standard deviations.

are the same under the switched and not switched conditions. The reason is that we use and test a different and possibly unflawed promise definition. We will come back to the two theories later, when we estimate panel models based on individual-level data, instead of session averages.

When dictators read communication of swiched partners, they may either find that it is a promise or not. If these dictators are rational, their second-order beliefs must be equal, conditional on the kind of message they read. However, we find a bias. Table 4 shows that those how made a promise have higher second-order beliefs when they read a promise made by another. Those who made a promise and read a promise expect 0.68 on averge. Those who did not make a promise and read a promise expect 0.55. Since they have the same information set and the same task, a difference in expectations can only be explained by a cognitive bias, as stated by Proposition 2.

We can test if the difference in the second-order beliefs emerges before or after the switched dictators are able to read the communication of the new partners. If the difference emerges before, then there is a false consensus effect (and possibly there is also a confirmatory bias); if it emerges afterwards, then we can exclude the false consensus effect and there is a confirmatory bias. In T2, dictators form their expectations before reading the new partner messages, but after knowing if they are switched or not. In Table 5, we show the results. Second-order beliefs are different after dictators know that they are switched, before reading the communication of the new partner (see column (2)). Despite no differences in the information sets, expectations of dictators who promised to roll (0.66) are greater than those (0.45) of dictators who did not ($Z = 4.02$, $p = 0.00$). Then there is a false consensus effect.

Table 5 – Average Second-order beliefs (T2)[15]

	No Switch	Switch
Dictator		
	(1)	(2)
promise	0.81	0.66
	(0.28)	(0.32)
no promise	0.55	0.45
	(0.39)	(0.34)

Proposition 6 *There is a false consensus effect in belief formation and the existence of a confirmatory bias cannot be excluded.*

[15] In brackets we report the standard deviations.

Table 5 – Average Second-order beliefs (T2)[16]

	No Switch	Switch
Dictator		
	(1)	(2)
promise	0.81	0.66
	(0.28)	(0.32)
no promise	0.55	0.45
	(0.39)	(0.34)

We now test the two theories for promise keeping by panel models. Here we use individual-level data. Results are reported in Table 6. The dependent variable is always the probability that the dictator chooses roll. Columns 1-4 are random intercept logit and probit models. Column 5 is a conditional logit. Starting from a general model (column 1), we eliminate insignificant interaction effects, deriving a relatively simple model (rilog3), which we also estimate using a probit and a conditional logit (columns 4 and 5).

[16] In brackets we report the standard deviations.

Table 6 – Estimates of panel regressions[17]

	rilog1	rilog2	rilog3	probit	clogit
Dictator agreed to roll	-0.32	-0.312	0.082	0.054	-0.044
	(0.84)	(0.82)	(0.51)	(0.30)	(0.47)
Partner was switched	-0.08	-0.15	0.21	0.12	0.058
	(0.59)	(0.60)	(0.28)	(0.17)	(0.26)
Second-order belief	1.49**	1.48**	1.74***	1.01***	0.79*
	(0.61)	(0.60)	(0.48)	(0.28)	(0.41)
Dictator agreed to roll ×	1.11	0.97			
partner was switched	(1.14)	(1.04)			
Dictator agreed to roll ×	2.44**	2.38**	1.87**	1.09**	1.76**
second-order belief	(1.10)	(1.07)	(0.75)	(0.43)	(0.72)
Partner switched ×	-1.57	-0.48			
partner agreed to roll	(1.10)	(0.39)			
Partner switched ×	0.70	0.639			
second-order belief	(0.83)	(0.84)			
Dictator agreed to roll	0.82				
x partner switched ×	(0.84)				
partner agreed to roll					
Partner switched ×	0.76				
second-order belief ×	(1.16)				
partner agreed to roll					
Dictator agreed to roll ×	-3.32**	-2.65*	-1.73***	-1.00***	-1.37***
partner was switched ×	(1.60)	(1.35)	(0.52)	(0.30)	(0.50)
second-order belief					
Round	-0.04	-0.04	-0.04	-0.03	-0.06
	(0.04)	(0.04)	(0.04)	(0.02)	(0.04)
Constant	-1.94***	-1.933***	-2.087***	-1.20***	
	(0.52)	(0.51)	(0.46)	(0.27)	
Log likelihood	-456.98	-457.67	-458.75	-459.91	-171.45

The effects of promises and partner switches are not significant once we control for second-order beliefs. This suggests that promises do not affect behavior *per se*. However, once we control for second-order beliefs, a non-switched dictator who promises is more likely to roll than a switched dictator who observes a promise made by others.

We believe, and our results show, that the reason why people keep promises

[17] We estimete models using GLLAMM (Stata). Standard errors take into account clustering by session. By *, **, and *** we indicate significance at the 10%, 5%, and 1% levels, respectively. The CLOGIT panel regression is based on 519 observations, all the others estimates are based on 768 observations.

is somehow in the intersection of the two theories. On the one hand, as predicted by commitment-based theory, people keep only their own promises and not promises made by others. On the other hand, people keep their promises only if promises raised partner's expectations, as predicted by guilt aversion theory.[18] Results are summarized by the following proposition.

Proposition 7 *People only keep their own promises and they do it when, because of these promises, expectations are higher. Therefore, promise keeping is due to guilt aversion.*

The table also shows that the coefficient on second-order beliefs is highly significant in all regressions as one can expected because a direct causal impact of second-order beliefs on behavior, as suggested by the theory of guilt aversion or because there exists e.g. a false consensus/confirmatory bias, which imply a correlation of second-order beliefs with unobserved factors. The latter interpretation is in line with our finding reported in the previous tables.

5 Conclusions

In this paper we tried to understand why people keep promises. We compared two exisitng theories. The first argues that people keep promises because they feel a sort of moral obligation. Thus, people keep only their own promises indepenently of the parter's expecations. The second one states that people keep promises only if the partner has high expectations since otherwise they feel guilty. This occurs independently of whom made the promise.

The truth is somehow in the intersection of these two theories. Making a promise is not sufficient to make people feel a moral obligation. People can sometimes "forget" their promise. It is necessary that the one who promises also believes that she raised the partner's expectations. Only in this case, she feels guilty and honors her obligation.

Our results come out from some crucial refinements of Vanberg's (2008) design, which capture important aspects of bilateral communication and expectation formation processes. A crucial aspect for testing promise keeping is in fact understanding how communication yields to a promise and how the latter shapes expectations. We showed as flawed promise definitions may lead to biased interpretation of the experimental data. We assumed that a promise in bilateral communication is the result of proposals and counter-proposals, which leads to an agreement. We verified that our definition is

[18] Our view is close to the idea of role-dependent guilt aversion, which has been formalized in psychological games under incomplete information by Attansi *et al.* (2014).

consistent with the data, whereas Vanberg's alternative definition based on unilateral proposal is not. Regarding the belief formation process, we assumed that it is complex and potentially affected by cognitive biases. We formalized it as a signal problem, where reading own "communication" is a different task compared to interpret messages written by others. Empirical results were in line with our interpretation We found evidence for cognitive biases and substantial differences in interpreting promises made by other people compared to own promises.

References

Attanasi, G., P. Battigalli and E. Manzoni (2014), "Disclosure of belief-dependent preferences in a trust game," paper presented a the workshop on *Communication, reciprocity and beliefs*, Sapienza University of Rome (April).

Battigalli, P. and M. Dufwenberg (2007), "Guilt in games," *American Economic Review*, 97: 170-176.

Battigalli, P. and M. Dufwenberg (2009), "Dynamic psychological games," *Journal of Economic Theory*, 144: 1-35.

Ben-Ner, A. and L. Putterman (2009), "Trust, communication and contracts: An experiment," *Journal of Economic Behavior and Organization*, 70: 106-121.

Ben-Ner, A., L. Putterman and T. Ren (2011), "Lavish returns on cheap talk: Non-binding communication in a trust experiment," *Journal of Socio-Economics*, 40: 1-13.

Charness, G. and M. Dufwenberg (2006), "Promises and partnership," *Econometrica*, 74: 1579-1601.

Charness, G. and M. Dufwenberg (2010), "Bare promises: An experiment," *Economics Letters*, 107: 281-283.

Charness, G. and M. Dufwenberg (2011), "Participation," *American Economic Review*, 101: 1211-1237.

Ellingsen, T. and M. Johannesson (2004), "Promises, threats and fairness," *The Economic Journal*, 114: 397-420.

Fischbacher, U. (2007), "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10: 171-178.

Geanakoplos, J., D. Pearce and E. Stacchetti (1989) "Psychological games and sequential rationality," *Games and Economic Behavior*, 1: 60-79.

Rabin, M. and J.L. Scharg (1999), "First impressions matter: A model of confirmatory bias," *Quarterly Journal of Economics,* 114: 37-82.

Vanberg, C. (2008), "Why do people keep their promises? An experimental test of two explanations," *Econometrica*, 76: 1467-1480.