PROMISES AND EXPECTATIONS

By

Florian Ederer and Alexander Stremitzer

December 2013
Revised October 2017

COWLES FOUNDATION DISCUSSION PAPER NO. 1931R

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Promises and Expectations*

Florian Ederer† Yale University

Alexander Stremitzer‡ UCLA

October 12, 2017

Abstract

We investigate why people keep their promises in the absence of external enforcement mechanisms and reputational effects. In a controlled laboratory experiment we show that exogenous variation of second-order expectations (promisors' expectations about promisees' expectations) leads to a significant change in promisor behavior. We provide evidence that a promisor's aversion to disappointing a promisee's expectation leads her to behave more generously. We propose and estimate a simple model of conditional guilt aversion that is supported by our results and nests the findings of previous contributions as special cases.

Keywords: promises, expectations, beliefs, contracts, guilt aversion

JEL Classification: A13, C91, D03, C72, D64, D80, K12.

*We wish to thank the associate editor and two anonymous referees for comments and suggestions that greatly improved the paper. We are also grateful to Jason Abaluck, Jennifer Arlen, Pierpaolo Battigalli, Andreas Blume, Arthur Campbell, Gary Charness, Martin Dufwenberg, Christoph Engel, Florian Englmaier, Constança Esteves-Sorenson, Christine Exley, Robert Gibbons, Holger Herz, Lisa Kahn, Navin Kartik, Camelia Kuhnen, Rosario Macera Parra, Bentley MacLeod, Benjamin Nyblade, Jens Prüfer, Andreas Roider, Frédéric Schneider, Marta Serra Garcia, Seana Shiffrin, Joel Sobel, Rebecca Stone, Steven Tadelis, Noam Yuchtman, Kathryn Zeiler and seminar audiences at Columbia, the Max Planck Institute for Research on Collective Goods, Northwestern, NYU, MIT, Regensburg, Toronto, UCLA, Yale, Zurich, ALEA, the NBER Summer Institute, the NBER Organizational Economics Meeting, the AEA 2015 meetings, SITE 2015, and the UCSD Deception Conference for helpful comments and suggestions as well as Estela Hopenhayn and the California Social Science Experimental Laboratory (CASSEL) at UCLA for helping us to conduct the experiments. We are particularly grateful to Mark Greenberg whose insightful comments influenced the design of our experiment. Jessie Cammack, James Davis, Sean Maddocks, and Jaimini Parekh provided excellent research assistance. We acknowledge financial support from the UCLA Faculty Research Grant Program.

†Yale School of Management, 165 Whitney Avenue, New Haven, CT 06511, florian.ederer@yale.edu.

‡UCLA School of Law, 385 Charles E. Young Drive, 1242 Law Building, Los Angeles, CA 90095, stremitzer@law.ucla.edu.

1 Introduction

To facilitate production and exchange over time, parties often make promises in order to commit to a particular course of action. There are three main reasons why a party would honor such an obligation (Dixit 2009). The first is the existence of a third party enforcement mechanism, as studied in the formal contracting literature beginning with Mirrlees (1976) and Holmström (1979). A second reason for honoring an obligation is the reputational incentive that arises when a party is concerned that reneging on a promise might hurt future payoffs, as studied extensively in the literature on relational contracting (Macaulay 1963, Klein and Leffler 1981, Bull 1987, Kreps 1990, MacLeod and Malcomson 1989, Levin 2003). A third reason, and the focus of the present paper, is the moral force of promise-keeping. A string of recent studies offers experimental evidence that promises, even if they come in the form of mere cheap talk, considerably enhance subsequent levels of cooperation in experimental trust and dictator games (Ellingsen and Johannesson 2004, Charness and Dufwenberg 2006, Vanberg 2008, Charness and Dufwenberg 2011). In this paper, we investigate the third channel, specifically how expectations about future payoffs influence promise-keeping. By exogenously varying expectations, we provide evidence that a promisor's aversion to disappointing a promisee's expectations leads her to behave more generously.

While the practical relevance of the moral force of promise-keeping is undisputed, there is a vigorous debate in economics, social psychology, philosophy, and law about why people keep (or should keep) their promises in the absence of explicit contractual and reputational concerns.[1] A clear understanding of what drives a person to keep her promise is essential to harnessing the beneficial effects of promises in institutional design, whether it be in the design of legal policy, regulatory regimes, contracts, or organizations.

Two leading explanations for the moral force of promise-keeping have been proposed. Proponents of the *expectation-based* theory argue that promisors (senders of promises) keep their word in order to avoid guilt incurred by failing to meet the expectations created in promisees (receivers of promises). A promisor would therefore be more likely to keep her promise if she believed that the promisee expected her to keep her promise.[2] In contrast,

[1] Notable contributions to the broader literature on promise-keeping in political sciences and social psychology include Ostrom, Walker, and Gardner (1992), Kerr and Kaufman-Gilliland (1994), Sally (1995), and Bicchieri and Lev-On (2007). In legal philosophy, classic references include Fried (1981), Atiyah (1983), and Scanlon (1998). For a recent contribution containing a survey of the previous literature, see Shiffrin (2008).

[2] Charness and Dufwenberg (2006) provide experimental evidence consistent with expectation-based the-

the *commitment-based* theory claims that promisors have a preference for keeping their word independent of the expectations of promisees. A promisor would therefore suffer a cost from behaving in a way inconsistent with what she has promised.[3] The factors emphasized by these explanations are not mutually exclusive: a promisor may both keep her promise in order to avoid feeling guilt and to avoid suffering an additional cost independent from feeling that guilt. However, previous experimental research has either failed to disentangle these two theories or has only documented unambiguous support for the commitment-based explanation. The evidence on the role of expectations has been inconclusive. On the one hand, Dufwenberg and Gneezy (2000), Guerra and Zizzo (2004), Reuben, Sapienza, and Zingales (2009), Bellemare, Sebald, and Strobel (2011), Regner and Harth (2014), and Khalmetski, Ockenfels, and Werner (2015) present evidence that the recipient's expectations influence the other player's behavior in a variety of dictator, trust, and lost wallet games. On the other hand, Vanberg (2008) and Ellingsen, Johannesson, Tjøtta, and Torsvik (2010) document that the recipient's expectations do not matter. However, all of these studies derived their findings in settings where there was no direct promissory link between the two parties.[4] Instead, they tested whether a decision maker (the "dictator") is influenced in her decision by the recipient's expectations. They therefore did not directly test whether promisees' expectations influenced promise-keeping.[5]

Using a novel design which exogenously varies expectations while preserving promissory links, this paper is the first to test the expectation-based theory of promise-keeping. We find that expectations matter and that this result is primarily driven by dictators' decisions when they had previously made a promise. Consistent with these results, we argue that the inconclusive evidence for the expectation-based account in the prior literature might be due to the fact that a recipient's expectations do not matter, or at least matter much less to a dictator if she has not given a promise. One reason for this might be that a

ories, but do not exogenously vary second-order expectations.

[3]Experimental evidence for the commitment-based explanation for promise keeping can be found in the contributions of Braver (1995), Ostrom, Walker, and Gardner (1992), Ellingsen and Johannesson (2004), Vanberg (2008), and Ismayilov and Potters (2012).

[4]For example, in Vanberg (2008), recipient expectations are exogenously varied through third-party promises, but this variation comes at the expense of a broken promissory link between the two parties. In Ellingsen et al (2010), no promises are ever made to begin with.

[5]These papers still contribute to our understanding of promises because promises are one of many ways to create expectations in recipients. However, they do not study whether there is something special about expectations created and supported by a direct promise.

promise creates a sense of responsibility in the promisor, perhaps because the promisor thinks that her act of promising *caused* the promisee's expectations, or perhaps because a promise establishes a personal connection that increases the salience of the promisee's expectations. We propose a theory of conditional guilt aversion in which a promisor is influenced by her promisee's expectations but only if those expectations are supported by the promise made by the promisor. This theory is consistent with our experimental results and nests the findings of previous contributions as special cases.

One implication of our results is that promisors can create commitment by explicitly encouraging and inviting the creation of expectations (e.g., through advertisement, or through explicit contractual terms displacing background legal defaults) and that, ex post, promisees should make their expectations salient in order to encourage promisors to keep their promises, especially in situations where other incentives to perform a contract are muted. For example, Eigen (2012) argues that an effective way to assure compliance is to remind a contracting party that he has made a promise when entering into a contract. Common law also tracks our account of conditional guilt aversion. First, it recognizes the centrality of promises: every legally enforceable contract requires the existence of a promise as one of its elements (Restatement 2d of Contracts §1). Second, under the doctrine of promissory estoppel, common law does not recognize detrimental reliance by a disappointed party as a basis for a claim unless this reliance was induced by the other party's promise (Restatement 2d of Contracts §90).[6]

In particular, we use a trust game where a dictator (trustor, she) can make a free-form promise to a recipient (trustee, he) and the recipient can decide whether to trust the dictator and to remain in the game. Our main innovation is to introduce a move of nature after this opt-in decision which determines the probability that it will be technically possible for the dictator to keep her promise. Both parties learn at this point whether the game is played with a "reliable random device" under which there is a high probability that the dictator will be able to keep her promise or whether the game is played with an "unreliable random device," under which there is a low probability that the dictator will be able to keep her promise. In the next step, another move of nature determines whether the dictator is able to perform or not. While both parties know with which random device the game is played,

[6] See Stone and Stremitzer (2016) for a follow-up experiment testing the effect of detrimental reliance by the promisee on promise-keeping and for further discussion of the doctrine of promissory estoppel.

only the dictator but not the recipient learns whether or not the dictator is able to perform. Therefore, a dictator who knows she is able to perform may face two kinds of recipients: either the recipient has high expectations, as he has learned that the game is played with the reliable random device, or he has low expectations, as he has learned that the game is played with the unreliable random device. This design allows us to compare promise-keeping rates among dictators who are both able to keep their promises but hold different second-order beliefs (beliefs about how much the receiver expects to receive), depending on whether the history of the game leading up to the dictator's decision reveals that it was likely ("reliable random device") or unlikely ("unreliable random device") that the dictator would be able to perform.

Using a within-subject design that allows us to observe dictators under both reliability settings, we show that the exogenous variation of the reliability of the random device with which the game is played directly affects the recipient's first-order and the dictator's second-order expectations and that these significantly change the dictator's decision to keep her promise. Our findings provide clean evidence for an expectation-based explanation of promise-keeping: while the commitment created by promises between the two parties remains constant, second-order expectations increase due to the increase of the reliability of the random device, which in turn induces an increase in the promisor's performance rate.

Finally, with a simple structural model we recover subject-specific susceptibilities to guilt aversion and characterize their distribution in the subject population. While slightly less than half of our subjects are unaffected by this behavioral trait, the remaining proportion exhibits some degree of guilt aversion and there is significant variation in how guilt-averse these subjects are.

The remainder of the paper is organized as follows: Section 2 presents the design of the experiment and the experimental procedures. In Section 3 we report our results. In Section 4 we present a simple model of promises and conditional guilt aversion and use it to estimate guilt aversion in the subject population. Section 5 concludes. In the appendix we provide additional regression results, instructions for the subjects participating in our experiment, and formal proofs for our theoretical predictions.

2 Experimental Design and Procedure

Our experiment is designed to investigate the role of expectations in promise-keeping. We hypothesize that a dictator is more likely to keep her promise if she believes that the recipient expects her to keep her promise. Underlying our hypothesis is the idea that a dictator will be concerned about disappointing the recipient's expectations created by the dictator's promise.

Previous experiments were not designed to investigate this question, and hence they either confounded expectation- and commitment-based explanations or they used the expectations created by other promisors as a means of varying the level of promisees' expectations. Unlike in Vanberg (2008), in which the promisee's second-order expectations depended on additional promises made by a third party, in our experiment the promissory link between the promisor and the promisee is unbroken. Rather, the magnitude of the dictator's second-order expectations is exogenously varied by the type of random device that is selected.

2.1 Experimental Design

In our experiment, subjects are randomly matched in pairs in each period and play the experimental trust game depicted in Figure 1. The dictator sends the recipient a free-form message. The recipient can then decide to opt in or opt out. Finally, the dictator decides how much to contribute to the recipient.[7]

The main feature of our design is that, after the recipient decides to opt in, nature selects whether the subjects play the game with a *reliable* or an *unreliable* random device. This device determines how likely it is that the dictator will be able to choose some positive level of performance (i.e., any action other than *Don't Perform*). If the device is reliable, the dictator will have a 5/6 chance to be able to choose among five possible actions (*Perform, 3/4 Perform, 1/2 Perform, 1/4 Perform, Don't Perform*), four of which will deliver a positive payoff to the recipient. If the random device is unreliable, the dictator will have only a 1/6 chance to be able to choose anything other than *Don't Perform*. For example, if performance is impossible, the dictator receives \$14 and the recipient receives \$0. If performance is possible and the dictator chooses *Perform*, she receives \$10 and the recipient receives \$12.

[7] This free-form message approach, which allows the dictator to send any message to the recipient (with the exception of identifying information such as name, age, race, gender) follows previous research by Charness and Dufwenberg (2006) and Vanberg (2008). In contrast, Charness and Dufwenberg (2010) use pre-coded messages and find only small effects of such "bare" promises.

Figure 1 depicts the remaining payoffs for the two players. If the dictator chooses *Don't Perform,* the parties receive the same payoffs $(14, 0)$ as if performance had not been possible.

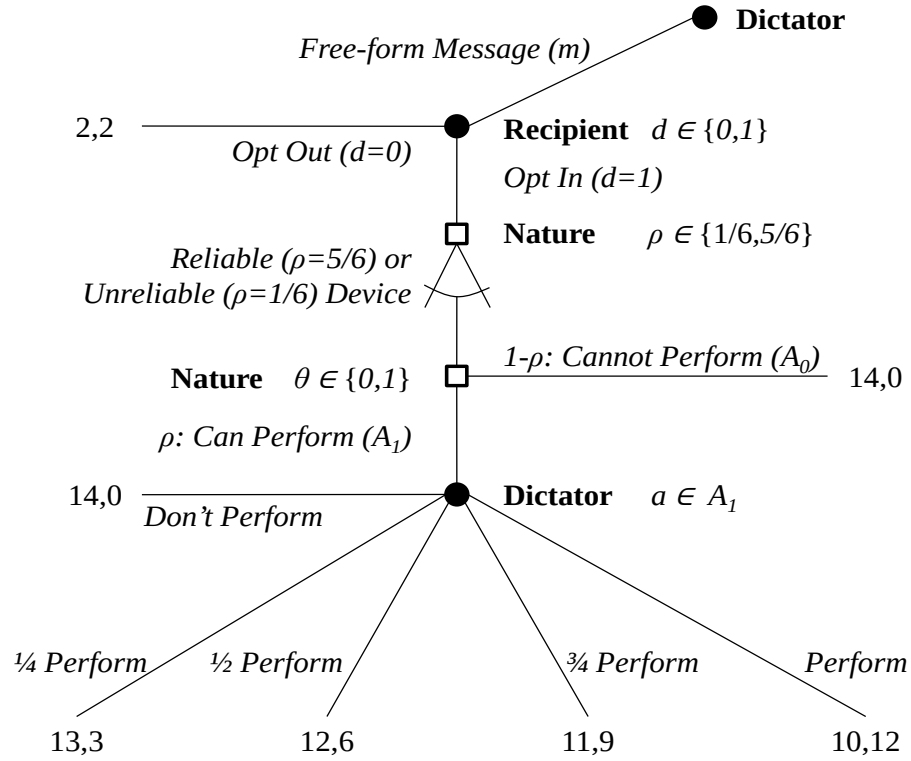

Figure 1: Dictator game with opt-out choice and reliable/unreliable device

Formally, the random device determines how likely it is that the dictator will find herself in one of two states of the world, $\theta \in \{0, 1\}$, with associated action space A_θ. This action space depends on whether the dictator is only able to choose *Don't Perform*, $\theta = 0$, or is able to choose other actions in addition, $\theta = 1$:

$$a \in A_\theta = \begin{cases} A_0 = \{0\} & \text{if } \theta = 0 \\ A_1 = \{0, .25, .5, .75, 1\} & \text{if } \theta = 1 \end{cases}.$$

Figure 1 illustrates that the dictator's monetary payoff can be written as a function of her decision a as $\pi_D(a) = 14 - 4a$. Similarly, the recipient's monetary payoff can be written as $\pi_R(a) = 12a$.

The timing of the game is as follows. At the beginning of each period $(t = 0)$, subjects are randomly paired, and nature randomly determines the identity of the dictator and the recipient in each pair. At $t = 1$ the dictator can send a free-form message m. In our

experimental analysis, these messages are coded as no message, empty talk, or promise, $\mu(m) \in \{\emptyset, 0, 1\}$. At $t = 2$, the recipient decides to opt out or to stay in, $d \in \{0, 1\}$. If $d = 0$, the game ends and payoffs $(2, 2)$ for the dictator and the recipient are realized. If $d = 1$, the game continues to $t = 3$. At $t = 3$, nature determines the type of the random device ρ with which the game is played. The random device can be either *reliable* ($\rho = 5/6$) or *unreliable* ($\rho = 1/6$), where ρ denotes the probability that the dictator will be able to choose from action space A_1. Both parties learn the type ρ of the random device. At $t = 4$, the dictator–but not the recipient–learns the state of the world θ and she makes the decision $a \in A_\theta$. Thus if $\theta = 1$ the dictator knows that she can perform and, when making her choice, she faces a recipient who plausibly expects a higher payoff under the reliable scenario than under the unreliable scenario. At $t = 5$, both players learn their payoffs, and the recipient learns the state θ.

At $t = 3$, we also elicit the recipient's and the dictator's beliefs. The recipient is asked which action $a \in A_1$ the dictator will choose if she is able to perform at $t = 4$. Because now the recipient knows with which random device ρ the game will be played, these beliefs might depend on the history of the game (realizations of μ and ρ). We therefore denote the first-order belief of the recipient by $\tau_R(\mu, \rho) \in [0, 1]$. In turn, the dictator has a belief (probability measure) regarding τ_R. Let $\tau_D(\mu, \rho) \in [0, 1]$ denote the dictator's mean second-order belief about the recipient's belief $\tau_R(\mu, \rho)$.[8] As we discuss in our detailed description of the experimental procedure, this elicitation of beliefs was incentivized. It is important to note that at $t = 4$ (i.e., at the time when the dictator makes her decision) the dictator—but not the recipient—knows the state of θ (i.e., whether the dictator can perform at all).

This game is largely identical to the trust/dictator game in Charness and Dufwenberg (2006), with a few differences. First, there is a richer action space to allow for more variation in the contribution rates of dictators. Second, we use a within-subject design that asks dictators to choose actions for both the reliable and unreliable device before either the dictator or the recipient learns with which of the two devices the game will be played.[9] Third,

[8] Similar to the design in Vanberg (2008) the elicitation of the beliefs in our experiment restricts the set of first- and second-order beliefs to a set of five elements. In our case, these beliefs also mirror the action set A_1.

[9] Charness and Dufwenberg (2006) used the strategy method for the recipient's initial opt-in decision. There is an extensive literature exploring whether or not there are systematic differences between within- and between-subject designs (Brandts and Charness 2000, Casari and Cason 2009, Charness, Gneezy, and Kuhn 2012). One advantage of the within-subject design is that it ensures observation of the behavior of

our design does not grant the dictator deniability of her actions vis-à-vis the recipient because the recipient learns the state θ at the end of the game.[10] Finally, and most importantly, we introduce a random device which determines the probability ρ with which the dictator will be able to choose some positive level of performance, and we elicit beliefs $\tau_R(\mu, \rho)$ and $\tau_D(\mu, \rho)$ after the recipient's opt-in but before the dictator's performance decision. The main purpose of this design innovation is to exogenously vary (and measure) the dictator's and the recipient's expectations without breaking any promissory link that exists between a dictator and a receiver.

If the random device is reliable, then there is a probability of $\rho = 5/6$ that the dictator will be able to choose *Perform*. If, on the other hand, the random device is unreliable, there is only a probability of $\rho = 1/6$ that the dictator will be able to choose *Perform*. Thus, recipients who are playing the game with an unreliable random device can plausibly expect lower monetary payoffs from the game. Because dictators are aware of this change in expectations (due to independent variation in the experiment), their second-order expectations are also exogenously changed. It is important to note that our manipulations cannot affect the commitment *per se* because at the time the promise is made, the dictator only knows that the game is potentially played with the reliable or the unreliable random device, but she does not know which of the two scenarios will subsequently be realized. Similarly, at the time the dictator makes her decision at $t = 4$, she—but not the recipient—knows whether she is able to perform independently of the random device used in the game. At this point, only the history of the game differs. If higher second-order expectations lead to higher contribution rates by the dictators who promised to perform, this would constitute evidence for the expectation-based explanation of promise-keeping.

Our design allows us to test the expectation-based explanation of promise-keeping. First, because dictators experience more guilt under the reliable ($\rho = 5/6$) than the unreliable

each subject for exactly the same free-form message. We deem this crucial in a design which tries to vary expectations while holding commitment constant. There are, however, shortcomings of such a within-subject design because it introduces the possibility that participants may feel like they should differentiate their answers. Mischkowski, Stone, and Stremitzer (2016) report similar results to ours using a between-subject design.

[10] If the dictator chooses not to perform although she is able to, the recipient will know that the dictator did not perform. Eliminating this deniability is important because it rules out "guilt from shame" (Battigalli and Dufwenberg 2007, Tadelis 2011) as a competing explanation: a dictator would otherwise benefit from a higher level of deniability in the unreliable ($\rho = 1/6$) as opposed to the reliable ($\rho = 5/6$) device. Charness and Dufwenberg (2006) grant the dictator deniability, but because they do not use different random devices they do not simultaneously vary simple guilt and guilt from shame.

($\rho = 1/6$) random device, we expect the dictator's action a to be higher under the reliable device. This is the central hypothesis we test in this paper.

Hypothesis 1 *The dictator's action choice, a, is higher for the reliable device than for the unreliable device.* ***(H1)***

Testing this hypothesis relies on the exogenous variation induced by the two random devices. It should induce differential behavior if guilt aversion plays a role in the dictator's action choices. However, as discussed above, previous research found ambiguous evidence of guilt aversion when there is no promissory link between the two parties. Thus, it is possible that this exogenous variation will only have bite if the dictator has made a promise ($\mu = 1$). If there is no promise ($\mu = 0$) or no message ($\mu = \emptyset$) the dictator will likely feel less or no guilt at all, with the result that there will be no difference between the two devices. Note though that our experiment is not specifically designed to test this second difference in behavior across message categories because we do not have a second source of exogenous variation allocating different messages to subject pairs.

Second, because the type of random device creates different performance choices (see **H1**), we also expect first- and second-order beliefs about the action chosen by the dictator to differ. In particular, because performance is expected to be higher for the reliable random device, first-order beliefs of recipients should adjust accordingly in the two settings. This leads us to our second hypothesis.

Hypothesis 2 *First-order beliefs, $\tau_R(\mu, \rho)$, and second-order beliefs, $\tau_D(\mu, \rho)$, about the dictator's action choice, a, are higher for the reliable device than for the unreliable device.* ***(H2)***

In contrast to the central hypothesis about contribution actions **(H1)** our second hypothesis regarding *beliefs* **(H2)** is more restrictive because it requires beliefs to adjust to the changes in actions postulated in **H1**. The reason for this is as follows. Our design exogenously varies the recipient's first-order *expectations* $12\tau_R(\rho)\rho$ and the dictator's second-order *expectations* $12\tau_D(\rho)\rho$ by letting subjects choose under $\rho = 5/6$ and $\rho = 1/6$ to test the causal impact of expectations on performance rates a (**H1**). For this change in performance to occur, subject beliefs $\tau_R(\rho)$ and $\tau_D(\rho)$ need not change at all between $\rho = 5/6$ and $\rho = 1/6$ because the exogenous shock is already achieved through the multiplicative impact of ρ on

expectations. However, for the secondary prediction about the change in beliefs (**H2**) it is necessary that performance rates actually differ between the two devices. In other words, in order for beliefs to differ between the reliable and unreliable device we need actions to differ (**H2**), but for actions to differ between reliable and unreliable device (**H1**) we do not need beliefs to differ because expectations are already being shocked exogenously. We now test these hypotheses in our data.

2.2 Experimental Procedure

We conducted 20 experimental sessions with a total of 280 student subjects at the California Social Science Experimental Laboratory (CASSEL). The CASSEL subject pool consists exclusively of UCLA undergraduate students. Subjects were assigned to visually isolated computer terminals. Beside each terminal they found paper instructions, which are reproduced in Appendix B. Questions were answered individually at the subjects' seats.

Each session consisted of 2 unpaid practice rounds followed by 8 paying rounds. In each round, subjects interacted with another randomly chosen participant. Under no circumstances did any participant interact with any other participant two times in the paying rounds. We achieved this by creating matching groups of exactly 10 subjects. At the end of the experiment, one of the 8 paying rounds was randomly chosen for payment. Each round was equally likely to be selected. The amount paid out at the end of the experiment depended on the decisions made in that round. In each period we also elicited first- and second-order beliefs of subjects about the behavior of other subjects. This elicitation of beliefs was incentivized and to prevent hedging subjects were paid for all rounds except the one chosen for payment of the decision. The subjects received a fixed fee of $10 for arriving on time. The experiment was programmed and conducted with the software z-Tree (Fischbacher 2007).

First, each subject was randomly matched with an interaction partner, and one participant in each pair was randomly assigned to the role of dictator or the role of recipient.[11] The pairings and the players' roles were randomly assigned anew in each round. A subject was always equally likely to be assigned to either role, regardless of the previous messages or actions in the game.

[11] In the instructions, we neutrally refer to the role of the dictator and the role of the recipient as "Role A" and "Role B," respectively.

Second, each dictator could send free-form messages to her recipient. The dictator could send any number of (unidirectional) messages within a time frame of 90 seconds.[12] Each message could contain up to 256 characters. Subjects were not allowed to reveal their names or any other identifying features such as race, gender, hair color, or seat number. In every other respect, subjects were free to send any message they liked.

Third, after receiving the dictator's message, each recipient could decide whether to opt in or out. If the recipient chose to opt out, each player received \$2. If the recipient chose to opt in, the game continued. At this point, neither player knew whether the random device determining whether the dictator would be able to perform was *Random Device 5/6* ("reliable random device", probability of 5/6 that the dictator would be able to choose some action other than *Don't Perform*) or *Random Device 1/6* ("unreliable random device", probability of 1/6 that the dictator would be able to choose some action other than *Don't Perform*). However, both parties knew that each scenario could occur with an equal probability of 50%.

Fourth, the recipient guessed which choice the dictator would likely make if she could choose to perform, and the dictator guessed which payoff the recipient expected to receive. Specifically, recipients and dictators were asked to choose from a five point scale. While the recipient's guessing payoff depended on the contribution decision of the dictator, the dictator's guessing payoff depended on the belief chosen by the recipient. Both payoffs were higher the closer they were to the actual contribution and belief, respectively.[13] Consistent with our use of the strategy method, we asked the players to make their guesses for both reliability scenarios. We asked them to assume that both parties knew the game was played with the reliable or the unreliable random device, respectively, and record their choices for each scenario. Note that if a recipient thought that the dictator intended to choose *Perform* (allocating \$12), a recipient's expected payoff depended on the reliability scenario. The expected payoff was \$2 if the game was played with Random Device 1/6 ($12 \times 1/6 = 2$) and \$10 if the game was played with Random Device 5/6 ($12 \times 5/6 = 10$). Asking the dictator to make her guess in terms of the recipient's expected payoff allowed us to make those expectations particularly salient. These elicitations yielded five point scales between 0 (performance very unlikely) and 1 (performance very likely) for first- and second-order

[12] Note that the 90-second time frame was not enforced. It just served as an informal pacemaker, but all dictators concluded their communication before this time frame.

[13] Except for offering five (rather than two) potential choices for the contribution decision, our belief elicitation method is identical to that used in Vanberg (2008).

beliefs.

Fifth, the dictator was asked to assume that she was able to perform, and to make her contribution decision as if the game leading up to that point had been played with Random Device 5/6 or Random Device 1/6. Figure 1 depicts the players' payoffs under the different possible contribution decisions.

Sixth, the computer randomly determined whether the game was played with the reliable or unreliable random device and drew an equally likely integer between 1 and 6 for each pair using z-Tree's random number generator. If the random device was reliable it was possible for the dictator to perform for the numbers 2, 3, 4, 5, and 6. If the random device was unreliable, the dictator was able to perform for the number 1.

Finally, at the end of each round, both dictators and recipients learned with which random device the game had been played, whether it had been possible for the dictators to perform, and what payoffs both participants had earned.

3 Results

The data comprise 20 experimental sessions involving a total of 280 subjects with a total of 28 matching groups of 10 subjects. Each session lasted for 8 rounds. The average number of dictator decisions made by each subject is 4. As we used the strategy method to elicit first- and second-order beliefs and contribution choices for both the reliable and the unreliable device, this within-subject design gives us a total of 1,120 decisions made under each reliability scenario. However, each matching group constitutes only one independent observation. Non-parametric tests are therefore based on matching group averages of the relevant variables. For comparisons between the random devices, we match observations that allow us to use two-tailed Wilcoxon signed-rank tests. For unmatched comparisons between message categories we use two-tailed Mann-Whitney rank-sum (MW) tests.[14]

[14]For all the instances where the MW rank-sum test is used, we also used the Fligner-Policello (FP) test. This test is a robust rank test for unmatched data which does not require that the two populations that are to be compared have equal variances. See Section 4.4 of Hollander, Wolfe, and Chicken (2013) for a complete description of the FP test. In our analysis the p-values of each MW rank-sum test and its corresponding FP test are almost identical and we therefore omit reporting the results for the latter test.

3.1 Performance Rates

We first examine how the experimental variation in beliefs affected subject performance rates. The average contribution (performance rate) the dictator gave to the recipient was \$5.76 (0.48) for the reliable device ($\rho = 5/6$) and \$5.28 (0.44) for the unreliable device ($\rho = 1/6$), conditional on performance being feasible. While this difference in contributions is small in magnitude it is statistically significant (Wilcoxon signed-rank, p-value < 0.01) and consistent with our central hypothesis **H1**. The dictator chose higher performance rates under the reliable device, when she was likely to be able to perform, than under the unreliable device, when the ability to perform was unlikely. Furthermore, as we show below, the magnitude of the contribution differentials is in line with the small, but significant differentials in second-order beliefs between the two random device settings, lending further empirical support to the expectations-based explanation for promise-keeping. However, this difference in behavior between the two random devices is driven by dictators' decisions when they had made a prior promise.

To investigate the role of promises, we asked a student assistant to code dictators' messages according to whether they contained a promise stating that the subject would choose any action other than *Don't Perform*. Following Charness and Dufwenberg (2006), this classification yielded three categories: "no message" ($\mu = \emptyset$) containing no text at all; "empty talk" ($\mu = 0$) messages (e.g., "Hey I just met you/and this is crazy/but here's my message/so money maybe?"); and "promise" ($\mu = 1$) messages (e.g., "im going to choose 3/4th perform so please dont opt out"). After accounting for all opt-out decisions (see Section 3.4), there remain 383, 300, and 268 individual observations, and 28, 27, and 28 matching group observations in the three message categories (promise, empty talk, no message) for both the reliable device and the unreliable device. We use these observations for our remaining analysis.[15]

When the dictator made a promise ($\mu = 1$), the average contribution (performance rate) she made to the recipient was \$7.08 (0.59) for the reliable ($\rho = 5/6$) and \$6.48 (0.54) for the unreliable device ($\rho = 1/6$), conditional on performance being feasible. This difference is highly statistically significant (Wilcoxon signed-rank, p-value < 0.01) and again consistent with our central hypothesis **H1**. Dictators contributed significantly more under the reliable

[15] Our freeform message experimental design also allows the dictator to make conditional promises. However, no messages containing conditional promises were sent.

device because in that scenario the random device exogenously induced higher second-order expectations and hence more guilt when falling short of those heightened expectations.[16]

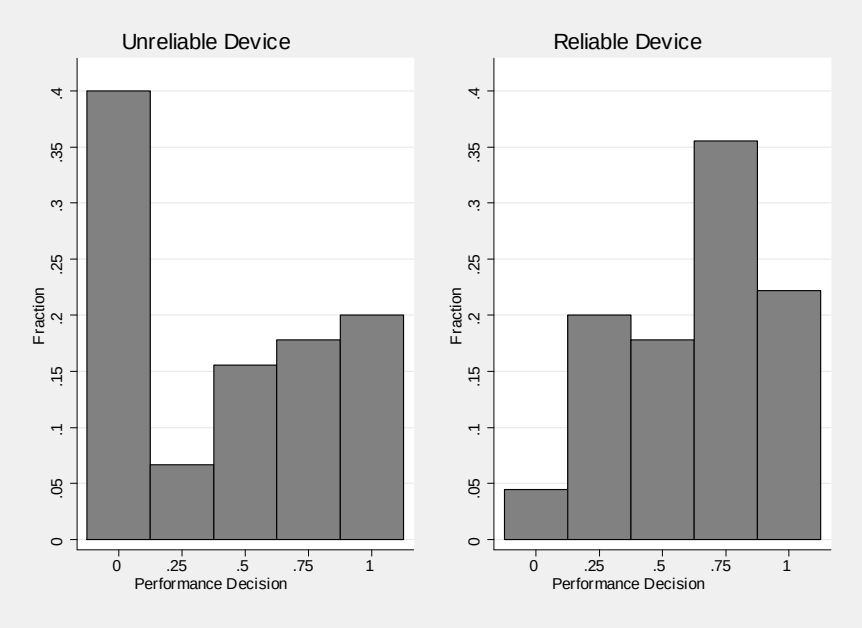

Figure 2: Fraction of performance decisions of dictators (*Don't Perform*, *1/4 Perform*, *1/2 Perform*, *3/4 Perform*, *Perform*) who sent a message containing a promise and made different decisions for the unreliable and the reliable device

Since we are employing the strategy method (and thus a within-subject design), it is particularly instructive to examine the behavior of those dictators who made different contribution decisions in the unreliable and the reliable device settings. Figure 2 shows the contribution decisions of dictators who promised to contribute and chose different contribution decisions depending on whether the device was reliable or unreliable.[17] A much lower proportion of dictators chose *Don't Perform* for the reliable than for the unreliable device; hence, more of them ended up choosing higher performance rates. While 40% of dictators chose *Don't Perform* under the unreliable device, less than 5% chose the same action under the reliable device. As a result, a much larger proportion of dictators chose to contribute a positive amount under the reliable device, with the *3/4 Perform* action seeing the largest increase.[18]

[16] In Section 3.3 we show that in addition to generating higher second-order payoff expectations by design, the reliable random device also induced higher second-order beliefs regarding actions.

[17] Note that all of our statistical tests are based on the full sample of dictators and not just those dictators who made different decisions for the reliable/unreliable device.

[18] Figure 2 also shows that many subjects make interior choices of a, suggesting that guilt aversion is not

It is crucial to note that the statistically significant difference in dictator contributions between the two settings disappeared when the dictator engaged in empty talk ($\mu = 0$, Wilcoxon signed-rank, p-value 0.27) or sent no message at all ($\mu = \emptyset$, p-value 0.14). Thus, in the absence of an explicit promise, higher reliability did not lead to significantly higher performance. These findings do not support **H1**, which hypothesized that contributions should differ between the two reliability settings even if the dictator did not promise. Our results suggest that the receiver's expectations only influence the dictator's contribution decisions when the dictator committed herself to a promise; they play little or no role when the dictator made no promise.

These findings also suggest a particular structure for the role of expectations in promise-keeping: no matter how different the dictator's second-order expectations under the two reliability scenarios, the dictator will respond to them only if she initally made a promise. However, our design is not ideally suited to directly test that expectations do not matter for those dictators who did not make a promise, as the absence of an effect could be due to selection: dictators who are more likely to promise are also more likely to be affected by differences in recipients' expectations. Still, some support for our conjecture may be derived from the fact that Vanberg (2008) found no effect of recipients' expectations on dictators' decisions to perform when there was no promissory link, while we identify a positive effect on performance when such a direct promissory link exists.

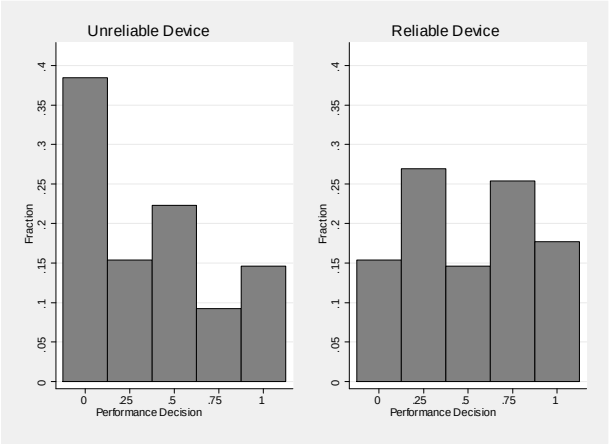

Figure 3a: Same as Figure 2 for empty talk

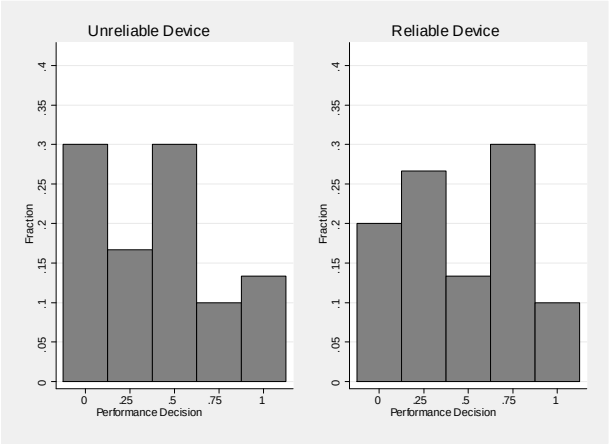

Figure 3b: Same as Figure 2 for no message

linear ($k = 1$), but that it is instead convex ($k > 1$) in the difference between beliefs and actions. We explore the magnitude of guilt aversion in greater detail in Section 4.2.

Figure 3 depicts the same data as Figure 2 for those dictators who changed their contribution decision between the two settings, but its two panels instead focus on the message categories of empty talk and no message, respectively. While there is a small change in behavior (tending towards more generous contribution rates) from the unreliable to the reliable device for dictators who sent an empty talk message (Figure 3a), there is practically no change for dictators who sent no message (Figure 3b). This slightly positive (but statistically insignificant) shift in dictator contribution rates is larger for empty talk messages than for no messages. In a related experiment investigating commitment-based explanations of promise-keeping, Ismayilov and Potters (2012) find that both trustees who make a promise *and* those who do not are more likely to be trustworthy if their message is delivered to the trustor. Their findings as well as ours suggest that any form of communication increases trustworthiness irrespective of the content of the message. This tentatively suggests that the guilt aversion effect is larger when there is some communication rather than none at all. However, as noted above, in both cases (and in contrast to when the dictator made a promise) there is no significant difference in the mean of the contribution decisions.

The highly significant statistical difference in performance rates between the reliable and unreliable device for promises (Wilcoxon signed-rank, p-value < 0.01) and the lack of statistical significance for empty talk (p-value 0.27) and no message (p-value 0.14) are not caused by large differences in the number of individual observations (and hence statistical power) across the three categories. This is because there are 383 (promise), 300 (empty talk), and 268 (no message) individual observations, and 28, 27, and 28 matching group observations, respectively, in the three message categories for both the reliable and the unreliable device. Furthermore, in the first column of Table 1 we report point estimates and standard errors adjusted for matching group clusters from a regression analysis that controls for subject fixed effects. The dependent variable is the performance rate a and the independent variables are the message categories interacted with the reliability of the random device such that "Promise" is the omitted category. As in our non-parametric tests, the performance rate is statistically significantly different between the two random devices for promises as shown by the positive and significant coefficient 0.049 on "Promise $\times$ Reliable". This difference in performance between the reliable and unreliable device is lower (though still positive) and not statistically significant for the other two message categories as shown by the coefficients 0.025 on "Empty Talk $\times$ Reliable" and 0.022 "No Message

$\times$ Reliable". Furthermore, the difference in performance between the two random devices when the dictator has made a promise is significantly larger than the difference when the dictator has sent no message (0.049 vs 0.022, p-value 0.04) as well as larger (though not significantly larger) than the difference under empty talk (0.049 vs 0.025, p-value 0.12). Thus, as before, the regression analysis suggests that the guilt aversion effect is largest when there is a promise, and that it is much less pronounced when the dictator uses empty talk or sends no message. However, while the reliability mechanism is randomly assigned, the message category is not. Hence, these are conditional treatment effects.

Finally, in accordance with previous papers, average performance rates are higher if the dictator made a promise (0.56) than if there was just empty talk (0.38) or no message (0.38) (MW rank-sum, p-values < 0.01). This translates into a contribution difference of \$2.16, about 3 times as large as the \$0.60 difference in contributions between the two random devices. The difference in contributions between the reliable/unreliable devices is purely due to differences in second-order expectations. In contrast, the \$2.16 difference between dictators who made a promise and dictators who sent empty talk or no promise, is a combination of several effects. It contains a selection effect (subjects who promise are different from those who do not promise), the commitment effect (subjects feel compelled to contribute just because of the promise *per se*) and the expectation effect. Furthermore, in contrast to face-to-face interactions between people who know each other well and may care strongly about not disappointing the other party's expectations, we would not expect the expectation effect to be very large in magnitude in our anonymous, low-interaction experimental laboratory setting. However, we show that even in this setting, which is very similar to consumer-to-consumer e-commerce transactions, expectations matter.

3.2 First-Order Beliefs and Expectations

Having documented evidence for the expectation-based explanation by analyzing the differences in contribution rates across treatments, we now investigate whether the secondary hypothesis regarding beliefs (**H2**) is also borne out in our data. Recipients were asked to guess the dictator's decision on a five-point scale between 0 and 1. When the random device was reliable, the recipients had a mean first-order belief $\tau_R(\mu, 5/6)$ of 0.28; when the device was unreliable, recipients had a lower mean first-order belief $\tau_R(\mu, 1/6)$ of 0.24. Although the sign of this difference is in accordance with **H2** it is not statistically significant (p-value

0.15).

The same pattern holds for comparisons within the three message categories. Under the reliable device the average first-order beliefs were 0.23, 0.26, and 0.34 for no message, empty talk and promise, respectively. Under the unreliable device they were lower across the board at 0.22, 0.21, and 0.29. However, these differences in first-order beliefs between the two device settings are not statistically significant (Wilcoxon signed-rank, p-values 0.17, 0.95, 0.45). Even when the dictator sent a promise, the difference in the recipient's first-order beliefs is not statistically significant between a reliable and an unreliable device. We obtain similar results from a subject fixed effects regression (column 2 of Table 1). There is a positive difference in first-order beliefs between the reliable and unreliable random devices that is largest and statistically significant at the 10% level when there is a promise (coefficient of 0.052). This difference is smaller and statistically insignificant under empty talk (0.015), and even smaller and insignificant under no message (0.010). Furthermore, the belief difference under a promise is significantly larger than under no message (0.052 vs 0.010, p-value 0.08), but not than under empty talk (0.052 vs 0.015, p-value 0.21). Thus, our experimental results on first-order beliefs provide some evidence for **H2**, with the strongest evidence coming from elicited beliefs when there is a promise and less so otherwise.

It is important to remember that we elicited conditional *beliefs* to allow for easy comparability across the two reliability devices. Recipients were asked how much they thought the dictator would contribute if performance were feasible, as the recipients did not actually learn whether performance was feasible for the dictator until after the end of each round. Therefore, the relevant first-order *expectations* at the time the dictator forms her second-order expectations and makes her performance decision are given by the unconditional first-order expectations, which are $12\tau_R(\rho)\rho$. Thus, in order to obtain the first-order expectations in terms of expected payoffs, the elicited conditional first-order beliefs have to be multiplied by $5/6 \times 12 = 10$ for the reliable and by $1/6 \times 12 = 2$ for the unreliable device. We find that these unconditional first-order beliefs are substantially higher in the reliable (\$2.60, \$2.30, and \$3.40) than in the unreliable scenario (\$0.42, \$0.44, and \$0.58). These differences are all statistically significant (Wilcoxon signed-rank, p-values < 0.01).[19]

[19]This stark difference is a (mechanical) feature of our experimental design. In contrast to previous contributions, we do not vary second-order expectations through the endogenous variation of first-order beliefs. Instead, we directly and exogenously change second-order expectations through the different random device scenarios and the timing of when the dictator and the recipient learn about which random device was

Finally, for both outcome realizations of the random device, receiving a promise significantly raised the recipients' expectations relative to receiving no message or empty talk about how much they would receive from the dictator, moving first-order expectations from the lower values of 0.23 and 0.26 to 0.34 and from 0.22 and 0.21 to 0.29, respectively (MW rank-sum, p-values 0.007, 0.005, 0.022, 0.006). This pattern mirrors the results of Charness and Dufwenberg (2006) and Vanberg (2008).

3.3 Second-Order Beliefs and Expectations

We next investigate how second-order beliefs $\tau_D(\mu, \rho)$ (i.e., a dictator's belief $\tau_D(\mu, \rho)$ about the belief $\tau_R(\mu, \rho)$ that the recipient has about the dictator's performance decision) vary with the reliability of the random device. When the random device was reliable, the average second-order belief was equal to 0.64. In contrast, when the random device was unreliable, the same second-order belief fell to 0.53. This difference in second-order beliefs between the reliable and the unreliable random device is statistically significant (Wilcoxon signed-rank, p-value < 0.01) and is in accordance with **H2**. That is to say, the dictator's belief about the contribution decision the recipient expected the dictator to make was significantly higher when the random device was reliable than when it was unreliable. As explained in detail in Section 2 we hypothesize that this difference in second-order beliefs is an equilibrium response to the random device's exogenous shocks to first- and second-order expectations.

These results on second-order beliefs hold for all message categories. Second-order beliefs were 0.59 when the dictator sent no message, 0.59 for an empty talk message, and 0.72 when a promise was given under the reliable random device; they fell to 0.48, 0.43, and 0.64, respectively, under the unreliable counterpart. All of these differences are statistically significant (Wilcoxon signed-rank, p-values < 0.01). This was the case even when the dictator did not send any message or sent an empty talk message. Once again, we obtain similar results using a subject fixed effects regression analysis reported in column 3 of Table 1. The difference in second-order beliefs between the two random devices is positive and significant when there is a promise. Yet, this is also true under empty talk and no message. However, as shown in Section 3.1, contribution rates were not significantly higher when expectations were not supported by a promise. This again suggests a conditional structure of guilt aversion. Although the second order beliefs vary for all message categories, they only seem to matter

chosen.

for contribution decisions if there was a promise.

The second-order expectations given by $E[\pi_R|\tau_D, \rho] = 12\tau_D(\rho)\rho$ that correspond to these second-order beliefs influence the level of guilt experienced by the dictators. These second-order expectations are equal to \$5.88, \$5.91, and \$7.18 for the reliable device and significantly lower (Wilcoxon signed-rank, p-values < 0.01) for the unreliable device where they are equal to \$0.96, \$0.87, and \$1.28. This large difference in second-order expectations is, of course, mainly exogenously (and mechanically) created by the use of the random device and it serves to separately identify the effect of guilt aversion on promise-keeping.

Finally, the second-order beliefs in both reliability settings are significantly higher for promises than for empty talk and no messages (MW rank-sum, p-values < 0.01), illustrating again the power of promises.

3.4 Promises and Opt-Out Decisions

Recipients chose to opt out at different rates depending on which message they received from the dictator with whom they were paired. While only 7.3% of recipients chose to opt out after receiving a promise, 21.6% opted out if they received no message at all, and 12.8% opted out after an empty talk message. These differential opt-out rates are consistent with similar findings on the role of communication and promises (Charness & Dufwenberg 2006, Vanberg 2008).

The differences in opt-out rates between recipients who received a promise and those who received an empty talk message, as well as between empty talk messages and no messages, are only significant at the 10% level (MW rank-sum, p-values 0.07, 0.07). In contrast, the difference in opt-out rates between participants who received a promise and those who did not receive any message is significant (MW rank-sum, p-value < 0.01). Similarly, the difference between participants who received a promise and (pooled) participants who did not receive a promise, either because they received an empty talk message or no message at all, is significant at the 1% level (MW rank-sum, p-value < 0.01). In the regression analysis of column 4 of Table 1, opt-out rates are significantly higher under empty talk and even higher under no message. The low opt-out rate of recipients who received a promise from their partnered dictator indicates that the recipients expected higher relative payoffs from staying in the game than from opting out compared to recipients who received no message at all or just an empty talk message. Furthermore, the higher opt-out rate for recipients who

received no message relative to recipients of an empty talk message suggests that some form of verbal engagement is better than none at all when it comes to inducing recipients to stay in the game.[20]

4 Discussion

Using exogenous variation in expectations our experimental results provide evidence for the important role of guilt aversion in promise-keeping. However, our results, particularly those for **H1** in Section 3.1, also suggest that the role of guilt aversion is limited to settings in which there exists a promise between the two parties. To formally explain this dichotomy in our experimental findings as well as in the previous literature, we present a simple model that builds on psychological game theory (Geanakoplos, Pearce and Stacchetti 1989, Battigalli and Dufwenberg 2007, 2009) and captures the documented effect of guilt aversion on promise-keeping. We then use this model to recover the distribution of guilt aversion in our subject population.

4.1 A Simple Model of Promises and Conditional Guilt Aversion

4.1.1 Setup

Battigalli and Dufwenberg (2007) propose two general theories of guilt aversion based on simple guilt and guilt from blame, respectively. The following model uses the former concept of simple guilt in which the dictator cares about the (monetary) extent to which she lets the recipient down.[21] In contrast to Battigalli and Dufwenberg (2007), where the degree to which a dictator experiences guilt is based on the recipient's expectation before his opt-in decision, we use the recipient's and dictator's expectations *after* the recipient chose to opt in because our experimental design shocks expectations after this stage. That is to say, what matters for the dictator when he makes his contribution decision is the current expectation of the recipient, not the recipient's expectation at the time of the recipient's choice to opt in or out.[22]

[20] This is in line with the aforementioned results of Potters and Ismayilov (2012) who also find that even some limited form of communication (i.e., empty talk) increases trustworthiness.

[21] Guilt could also be caused by expectations about the promisor keeping her promise, as opposed to expectations of the recipient's monetary outcome. The former concept of guilt need not be affected by (the experimental manipulation of) ρ.

[22] More generally, a philosopher might object to this terminology as it comes close to depicting guilt aversion as a primary moral motivation. It can only be a secondary one, activated by a person's belief that some

Define $\gamma_D \geq 0$ as a constant measuring the dictator's sensitivity to guilt from disappointing the recipient's expectations, which the dictator expects to be equal to $E[\pi_R|\tau_D, \rho] = \rho 12 \tau_D(\rho)$. In line with our experimental results that show differential responses under the two random devices when there is a promise and when there is none, we posit that the role of expectations in promise-keeping has a conditional structure: a recipient's expectations only play a role if the dictator has made a promise ($\mu = 1$). A recipient's expectations will not play a role if the dictator has not made a promise ($\mu = 0$) to the recipient.

The dictator's utility U_D when she chooses a at $t = 4$ can now be written in the following way:

$$\begin{aligned} U_D(a) &= \pi_D(a) - \frac{\mu\gamma_D}{k}\left(\max\left\{E[\pi_R|\tau_D(\mu,\rho),\rho] - \pi_R(a), 0\right\}\right)^k \\ &= 14 - 4a - 12^k \frac{\mu\gamma_D}{k}\left(\max\left\{\rho\tau_D(\mu,\rho) - a, 0\right\}\right)^k. \end{aligned} \tag{1}$$

The last term of the dictator's utility function captures the impact of guilt. This term only plays a role if the dictator sent a promise ($\mu = 1$) and if the dictator is susceptible to guilt aversion ($\gamma_D > 0$). Guilt from disappointing the recipient's perceived expectations $E[\pi_R|\tau_D, \rho]$ by choosing a low payoff $\pi_R(a)$ for the recipient has a negative effect on utility, but there is no gain from exceeding the recipient's expectations. The dictator can reduce the negative utility from guilt by increasing her action a up to the point where it matches the dictator's beliefs about the recipient's expectations. In contrast to Charness and Dufwenberg (2006) and Battigalli and Dufwenberg (2007, 2009), we allow guilt to be linear ($k = 1$) or convex ($k > 1$) in the difference between the dictator's expectations, $E[\pi_R|\tau_D, \rho]$, and the realized payoffs for the recipient, $\pi_R(a)$. For $k = 1$, our model nests the model of Charness and Dufwenberg (2006) as a special case that only admits corner solutions of a. For $k > 1$, interior solutions of a (i.e., $a \neq \{0, 1\}$) are also possible.[23]

4.1.2 Analysis

There are two benchmark cases in which expectations do not affect actions. First, a dictator who is motivated solely by her own monetary payoff and is not sensitive to guilt at all, $\gamma_D = 0$, would have a utility of $14 - 4a$ and would therefore maximize her payoff by choosing

other fact itself provides direct motivation to act. In this case, the prospect of disappointing expectations would be wrong and therefore one should not do it.

[23] For ease of presentation we focus on the latter case in the main text. In the appendix, we show that our results also hold for $k = 1$.

$a = 0$.[24] Second, in the settings considered by Vanberg (2008) and Ellingsen et al (2010), in which no direct promissory link between the dictator and the recipient exists and thus $\mu = 0$, beliefs about expectations $\tau_D(\rho)$ also do not matter.

Thus, our model requires two assumptions for second-order beliefs to play a role in promise-keeping. There must exist a promise between the two parties, $\mu = 1$, and the dictator must experience some guilt aversion, $\gamma_D > 0$. Second-order expectations will then generate different predictions about the contribution choice a for the reliable ($\rho = 5/6$) and the unreliable ($\rho = 1/6$) device. The dictator's utility is given by

$$U_D = 14 - 4a - 12^k \frac{\gamma_D}{k} \left(\max\left\{\rho\tau_D(\rho) - a, 0\right\}\right)^k \tag{2}$$

which yields different levels of guilt for the two different devices and where we write $\tau_D(\mu = 1, \rho) = \tau_D(\rho)$ for simplicity. It is straightforward to see that the impact of guilt is larger for $\rho = 5/6$ than for $\rho = 1/6$, thus leading to a higher equilibrium action a for two reasons. First, there is a difference in actions resulting purely from the exogenous variation in the reliability ρ of the device. Second, there is an additional (second-order) effect resulting from the impact of this exogenous variation on equilibrium beliefs τ_D. The first-order condition with respect to a for the dictator yields the following interior solution:

$$a^* = \rho\tau_D(\rho) - \left(\frac{4}{\gamma_D 12^k}\right)^{\frac{1}{k-1}} \tag{3}$$

The dictator's action a^* is increasing with the reliability of the random device, ρ, the dictator's second-order belief, τ_D, and her susceptibility to guilt aversion, γ_D.

To see the first effect of ρ on a, assume that second-order beliefs about actions are the same in both settings, $\tau_D(5/6) = \tau_D(1/6)$, and that, just for ease of exposition, $k = 1$. As can be seen from equation (2), the guilt experienced by the dictator when choosing $a = 0$ is $2\gamma_D\tau_D$ for $\rho = 1/6$, which is much smaller in magnitude relative to the guilt experienced for $\rho = 5/6$ where it is $10\gamma_D\tau_D$. This argument holds a fortiori for $\tau_D(5/6) > \tau_D(1/6)$ as is evident from the first-order condition for interior solutions of a from equation (3). Thus, in equilibrium, the dictator chooses higher levels of a for $\rho = 5/6$ than for $\rho = 1/6$.[25]

[24] Of course, there are many reasons other than guilt aversion, such as social preferences or norms (Rabin 1993, Fehr and Schmidt 1999, Bolton and Ockenfels 2000, Andreoni and Bernheim 2009), that would predict an equilibrium action a other than 0.

[25] See Appendix C for a more rigorous proof of the theoretical predictions taking the possibility of corner solutions into account.

Theoretical Result 1 *If there is a promise ($\mu = 1$), the dictator's contribution action a is higher for the reliable device than for the unreliable device. If there is no promise ($\mu = 0$), there is no difference in the dictator's contribution action.* ***(TR1)***

TR1 is a modified version of **H1**. In Section 3.1 we showed that contributions differed between the two reliability settings if the dictator promised, but not in the absence of a promise. This lexicographic theory of promise-keeping can explain both the difference in contribution actions between random devices when there is a promise as well as the lack of this difference without an established promissory link between the two parties. Furthermore, it also explains why Charness & Dufwenberg (2006) find evidence for guilt aversion while Vanberg (2008) and Ellingsen et al (2010) do not.

As a result of these different action choices, equilibrium first- and second-order beliefs also differ in our model. In particular, because actions are higher for $\rho = 5/6$, equilibrium first-order beliefs of recipients must adjust to the different actions that the dictator chooses in the two settings. Hence, first-order beliefs are higher, $\tau_R(5/6) \geq \tau_R(1/6)$ and, as a result, equilibrium second-order beliefs must be higher too, $\tau_D(5/6) \geq \tau_D(1/6)$ if $\gamma \geq 0$.

Theoretical Result 2 *If there is a promise ($\mu = 1$), first-order and second-order beliefs are higher for the reliable device than for the unreliable device. If there is no promise ($\mu = 0$), there is no difference in first-order and second-order beliefs.* ***(TR2)***

Analogously, **TR2** is a modified version of **H2**. However, while we found strong support for **TR1** in our experimental data, the evidence for **TR2** is mixed. As shown in Section 3.2 first-order beliefs only significantly differ when there is a promise, but not otherwise. At the same time, as shown in Section 3.3, second-order beliefs differ significantly between random devices for all message categories. Given this mixed evidence on first- and second-order beliefs, in our ensuing estimation of the distribution of guilt aversion we only use **TR1** which results from the dictator's optimal contribution decision a^*, and simply treat the elicited beliefs τ_R and τ_D as given without requiring equilibrium consistency of beliefs.

4.2 Distribution of γ_D

Using this simple model, in particular **TR1** about the dictator's optimal contribution action, we can directly recover each dictator's susceptibility to guilt aversion, γ_D. Because many

dictators choose strictly positive levels of performance a even when there is no promise ($\mu = 0$), we augment our previous model by an additional term that captures altruism. The dictator's utility function is then given by

$$U_D = \pi_D(a) - \mu \frac{\gamma_D}{k} \left(\max\left\{E\left[\pi_R | \tau_D, \rho\right] - \pi_R(a), 0\right\}\right)^k - \frac{\delta_D}{r} \left(12 - \pi_R(a)\right)^r$$

where the dictator suffers a convex disutility if the receiver's payoff $\pi_R(a)$ falls short of his maximum possible payoff of 12.[26] If $\delta_D = 0$, the dictator is not driven by altruism, but as δ_D increases she cares more about the payoff obtained by the receiver. The values of k and r influence the convexity of the guilt aversion and the altruism terms, and are assumed to be known by the dictator and the recipient. Using identifying variation for subjects who are observed in the data choosing a under both a promise ($\mu = 1$) and no promise ($\mu = 0$), the unknown subject-specific altruism and guilt aversion parameters, δ_D and γ_D, are exactly identified from the two first-order conditions with respect to a under $\mu = 0$ and $\mu = 1$:

$$\begin{aligned} \frac{\partial U_D}{\partial a} &= -4 + 12\delta_D \left(12 - 12a\right)^{r-1} \\ \frac{\partial U_D}{\partial a} &= -4 + 12\gamma_D \left(\max\left\{12\rho\tau_D - 12a, 0\right\}\right)^{k-1} + 12\delta_D \left(12 - 12a\right)^{r-1} \end{aligned}$$

Taking corner solutions at $a = \{0, 1\}$ into account we use dictator-specific averages of a for given μ and ρ to solve for δ_D and γ_D.

The two panels of Figure 4 show the distribution of the altruism and guilt aversion parameters in the dictator population for quadratic altruism, $r = 2$, and quadratic guilt aversion, $k = 2$.[27] Given our assumptions, we are only able to identify the distribution of γ_D for 104 dictators who are observed under both $\mu = \{0, \emptyset\}$ and $\mu = 1$. As suggested by our reduced form analysis that documents a significant positive shift in performance from unreliable to reliable device when $\mu = 1$, the distribution of γ_D (Figure 4b) shows that more than half of the dictators exhibit guilt aversion, $\gamma_D > 0$, while the remaining slightly smaller proportion of just under 50% is unaffected by this behavioral trait, $\gamma_D = 0$.

[26] By using altruism we chose a very simple form of other-regarding preferences. This is because more general specifications could potentially lead to higher-order beliefs also playing a role. For example, dictators could be making contributions according to what is expected of them independent of their initial statement.

[27] Note that while convexity in the guilt aversion term (i.e., $k > 1$) is required to explain any subject choices that are not corner solutions, we chose these particular parameters for their simplicity and their fit with our data. Our results are largely unchanged under different parameter assumptions.

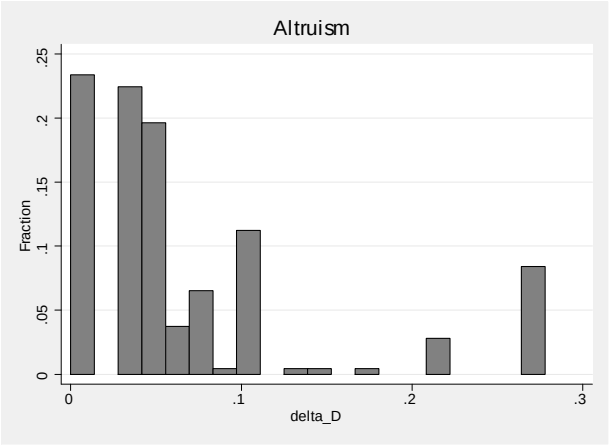

Figure 4a: Altruism parameter δ_D for $r = k = 2$

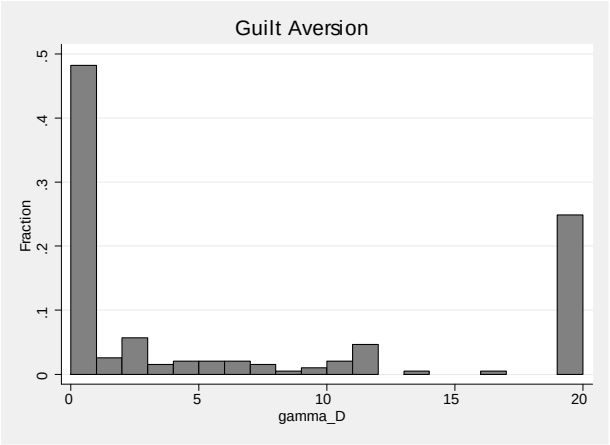

Figure 4b: Guilt aversion parameter γ_D for $r = k = 2$

The dictators with a positive γ_D fall into two broad categories as can be seen from Figure 4b. First, there is a mass of about 25% of all dictators (at $\gamma_D \approx 20$) where γ_D is so large that in equilibrium the dictators raise a sufficiently high such that $a \approx \rho\tau_D$. In this way, they reduce their own monetary payoff in order to completely avoid any loss from guilt aversion which they would otherwise suffer if they chose a lower performance a. Of course, the true γ_D of these dictators might be even higher than 20, but these subjects are already at a corner solution in our data. Second, for roughly 25% of dictators, γ_D lies between the two mass points of 0 and 20, and so these dictators trade off some monetary gains against losses from guilt aversion. They do not, however, raise a high enough to completely eliminate guilt in equilibrium.

5 Conclusions

Many psychological and economic experiments have shown that promises greatly enhance cooperative behavior in experimental games. In this paper we provided experimental evidence for the expectation-based explanation of promise-keeping. Previous experiments either could not distinguish between commitment-based and expectation-based explanations because treatment-induced changes in the alternative causal factors (promises and second-order beliefs) had occurred simultaneously, or the experiments focused on settings in which there was no promissory link between the dictator and the recipient.

In contrast, we designed our experiment to achieve independent variation in second-order expectations in an environment where these were supported by a direct promissory link between the dictator and the recipient, and thus by the existence of a sufficiently high level of commitment. Changes in the probability with which a dictator would be able to contribute directly impacted recipients' first-order and dictators' second-order expectations, which in turn significantly changed behavior.

In light of our own findings that recipients' expectations matter if supported by a promise, as well as previous findings which provide mixed evidence that expectations matter *per se* (i.e., in the absence of a promise), we proposed a theory of conditional guilt aversion in which we assume that the dictator's sensitivity to the recipient's expectations is switched on, or at least heightened, by the fact that she has given a promise and therefore presumably feels more responsible for the recipient's expectations.

As we noted previously, our design was not suited to directly test our theory of conditional guilt aversion. Specifically, our finding that expectations do not matter in the absence of a promise might also be due to different types of individuals choosing to make promises or not. However, our ability to explain the experimental results through the lens of this theory casts doubt on whether it is possible to conclude from Vanberg (2008) that promising *per se* has an effect on performance rates. To see this, imagine that the only motivation for keeping a promise is that the promisor does not want to disappoint the promisee's expectations. Further assume that, consistent with our theory of conditional guilt aversion, the sensitivity to the recipient's expectations is only switched on by a promise. Then, under the expectation-based theory, we would predict that a dictator who has made a promise is more likely to perform than a dictator who has not made a promise simply because her sensitivity to expectations will only be switched on if she made a promise. Note that for this to happen we need *not* assume any independent preference for promise-keeping. The whole effect could work exclusively through the desire not to disappoint expectations so long as they are supported by a promise. In order to show that there is an independent preference for promise-keeping we would have to design an experiment that varies the dictator's promissory commitment while keeping the recipient's expectations at zero.

Our results also suggest an interesting avenue for future work in this area. Because Vanberg (2008) and our design use exogenous variation in either promises or beliefs but not both, one would wish to disentangle the different effects in a design that provides exogenous

variation of both promises *and* beliefs. Recent contributions by Di Bartolomeo, Dufwenberg, Papa, and Passarelli (2017) and by Mischkowski, Stone, and Stremitzer (2016) who, in addition, provide evidence supporting our theory of conditional guilt aversion, are promising endeavors to fill this gap.

References

Andreoni, J., and B. D. Bernheim (2009): "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica*, 77(5), 1607–1636.

Atiyah, P. (1983): *Promises, Morals, and Law*. Clarendon Press.

Battigalli, P., and M. Dufwenberg (2007): "Guilt in Games," *American Economic Review, Papers and Proceedings*, 97(2), 170–176.

——— (2009): "Dynamic Psychological Games," *Journal of Economic Theory*, 144(1), 1–35.

Bellemare, C., A. Sebald, and M. Strobel (2011): "Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models," *Journal of Applied Econometrics*, 26(3), 437–453.

Bicchieri, C., and A. Lev-On (2007): "Computer-Mediated Communication and Cooperation in Social Dilemmas: An Experimental Analysis," *Politics, Philosophy, and Economics*, 6(2), 139–168.

Bolton, G. E., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90(1), 166–193.

Brandts, J., and G. Charness (2000): "Hot vs. Cold: Sequential Responses and Preference Stability in Experimental Games," *Experimental Economics*, 2(3), 227–238.

Braver, S. L. (1995): "Social Contracts and the Provision of Public Goods," in *Social Dilemmas: Perspectives on Individuals and Groups*, ed. by D. Schroeder. Praeger, New York.

Bull, C. (1987): "The Existence of Self-Enforcing Implicit Contracts," *Quarterly Journal of Economics*, 102(1), 147–159.

Casari, M., and T. Cason (2009): "The strategy method lowers measured trustworthy behavior," *Economics Letters*, 103(3), 157–159.

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

——— (2010): "Bare promises: An experiment," *Economics Letters*, 107(2), 281–283.

——— (2011): "Participation," *American Economic Review*, 101(4), 1211–1237.

Charness, G., U. Gneezy, and M. A. Kuhn (2012): "Experimental methods: Between-subject and within-subject design," *Journal of Economic Behavior & Organization*, 81(1), 1–8.

Di Bartolomeo, G., M. Dufwenberg, S. Papa, and F. Passarelli (2017): "Promises, Expectations, and Causation," *Sapienza University of Rome Working Paper*.

Dixit, A. (2009): "Governance Institutions and Economic Activity," *American Economic Review*, 99, 5–24.

Dufwenberg, M., and U. Gneezy (2000): "Measuring Beliefs in an Experimental Lost Wallet Game," *Games and Economics Behavior*, 30(2), 163–182.

Eigen, Z. (2012): "When and Why Individuals Obey Contracts: Experimental Evidence of Consent, Compliance, Promise, and Performance," *Journal of Legal Studies*, 41, 67–93.

Ellingsen, T., and M. Johannesson (2004): "Promises, Threats and Fairness," *Economic Journal*, 114(495), 397–420.

Ellingsen, T., M. Johannesson, S. Tjøtta, and G. Torsvik (2010): "Testing Guilt Aversion," *Games and Economic Behavior*, 68(1), 95–107.

Fehr, E., and K. M. Schmidt (1999): "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 114(3), 817–868.

Fischbacher, U. (2007): "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments," *Experimental Economics*, 10, 171–178.

Fried, C. (1981): *Contract as Promise*. Cambridge MA, Harvard University Press.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1(1), 60–79.

Guerra, G. A., and D. J. Zizzo (2004): "Trust Responsiveness and Beliefs," *Journal of Economic Behavior and Organization*, 55, 25–30.

Hollander, M., D. Wolfe, and E. Chicken (2013): *Nonparametric Statistical Methods*, Wiley Series in Probability and Statistics. Wiley.

Holmström, B. (1979): "Moral Hazard and Observability," *Bell Journal of Economics*, 10, 74–91.

Ismayilov, H., and J. Potters (2012): "Promises as Commitments," *Tilburg University, Center for Economic Research, Discussion Paper*, 2012-064.

Kerr, N. L., and C. M. Kaufman-Gilliland (1994): "Communication, Commitment and Cooperation in Social Dilemma," *Journal of Personality and Social Psychology*, 66(3), 513–529.

KHALMETSKI, K., A. OCKENFELS, AND P. WERNER (2015): "Surprising Gifts: Theory and Laboratory Evidence," *Journal of Economic Theory*, 159, 163–208.

KLEIN, B., AND K. B. LEFFLER (1981): "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, 89(4), 615–641.

KREPS, D. M. (1990): "Corporate Culture and Economic Theory," in *Perspectives on Positive Political Economy*, ed. by J. E. Alt, and K. A. Shepsle, pp. 90–143. Cambridge University Press.

LEVIN, J. (2003): "Relational Incentive Contracts," *American Economic Review*, 93(3), 835–857.

MACAULAY, S. (1963): "Non-Contractual Relations in Business: A Preliminary Study," *American Sociological Review*, 28, 55–69.

MACLEOD, B., AND J. MALCOMSON (1989): "Implicit Contracts, Incentive Compatibility, and Involuntary Unemployment," *Econometrica*, 57, 447–480.

MIRRLEES, J. (1976): "The Optimal Structure of Incentives and Authority Within an Organization," *Bell Journal of Economics*, 7, 105–131.

MISCHKOWSKI, D., R. STONE, AND A. STREMITZER (2016): "Promises, Expectations, and Social Cooperation," *Harvard Law School John M. Olin Center Discussion Paper No. 887*.

OSTROM, E., J. WALKER, AND R. GARDNER (1992): "Covenants With and Without a Sword: Self-Governance Is Possible," *American Political Science Review*, 86(2), 404–417.

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

REGNER, T., AND N. S. HARTH (2014): "Testing belief-dependent models," *Jena Working Paper*.

REUBEN, E., P. SAPIENZA, AND L. ZINGALES (2009): "Is Mistrust Self-Fulfilling," *Economics Letters*, 104, 89–91.

SALLY, D. (1995): "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992," *Rationality and Society*, 7(1), 58–92.

SCANLON, T. (1998): *What We Owe to Each Other*. Cambridge MA, Harvard University Press.

SHIFFRIN, S. V. (2008): "Promising, Intimate Relationships, and Conventionalism," *Philosophical Review*, 117, 481–524.

STONE, R., AND A. STREMITZER (2016): "Promises, Reliance, and Psychological Lock-in," *UCLA School of Law, Law-Econ Research Paper No. 15-17*.

TADELIS, S. (2011): "The Power of Shame and the Rationality of Trust," *UC Berkeley Haas Working Paper*.

VANBERG, C. (2008): "Why Do People Keep Their Promises? An Experimental Test of Two Explanations," *Econometrica*, 76, 467–1480.

A Tables

	Dependent Variable			
	(1)	(2)	(3)	(4)
	Performance a	1st-order belief τ_R	2nd-order belief τ_D	Opt-out
Promise × Reliable	0.049**	0.052*	0.076***	
	(0.018)	(0.029)	(0.021)	
Empty Talk	-0.077**	-0.089***	-0.122***	0.103**
	(0.034)	(0.027)	(0.036)	(0.040)
Empty Talk × Reliable	0.025	0.015	0.068**	
	(0.033)	(0.021)	(0.026)	
No Message	-0.092***	-0.068**	-0.108***	0.139***
	(0.022)	(0.030)	(0.029)	(0.043)
No Message × Reliable	0.022	0.010	0.056**	
	(0.020)	(0.028)	(0.022)	
Subject FE	Yes	Yes	Yes	Yes
Observations	1902	1902	1902	1120
Clusters	28	28	28	28

Table 1: Regressions for performance, 1st-order beliefs, 2nd-order beliefs, and opt-out rates

This table presents subject fixed effects regressions with the performance rate (column 1), the 1st-order belief (column 2), the 2nd-order belief (column 3), and the opt-out rate as the dependent variable. In columns (1)-(3) the message categories and the message categories interacted with the reliability of the random device are the independent variables where "Reliable" refers to the reliable ($\rho = 5/6$) random device. In column (4) only the message categories are the independent variables because the opt-out decision occurs before the random device shock. In all columns (1)-(4) "Promise" is the omitted category. Standard errors clustered at the matching group are reported in parentheses. Corresponding p-values are denoted by stars: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B Instructions

Thank you for participating in this experiment. The purpose of this experiment is to study how people make decisions in a particular situation. In case you should have questions at any time, please raise your hand. Please do not speak to other participants during the experiment. You will receive $10 for arriving on time. Depending on the decisions made and the decisions of other participants, you may receive an additional amount (as described below). At the end of the experiment, the entire amount will be paid to you individually and privately in cash.

This session consists of 2 practice rounds and 8 paying rounds with money prizes. In each round, you will interact with another randomly chosen participant. Under no circumstances will you interact with the same participant twice. No participant will learn the identity of the persons with whom he or she has interacted during the experiment.

At the end of the experiment, one of the 8 paying rounds will be randomly chosen for payment (every round is equally likely). The amount that you will receive at the end of the experiment will depend on the decisions made in that round.

Each round consists of 7 steps, which are described below.

Overview

There are two players; Player A and Player B. Initially, A can send a chat message to B over the computer, and B can decide whether he wants to opt out of the game, leading to payoffs of \$2 for each player. If B does not opt out, a random device will determine whether it will be possible for A to perform, that is, allocate money to B. If it is impossible to perform, Player A gets a payoff of \$14 and B gets a payoff of 0. If it is possible for A to perform, he can make one of 5 choices:

- *Don't Perform*: A keeps \$14 for himself and allocates \$0 to B.
- *1/4 Perform*: A keeps \$13 for himself and allocates \$3 to B.
- *1/2 Perform*: A keeps \$12 for himself and allocates \$6 to B.
- *3/4 Perform*: A keeps \$11 for himself and allocates \$9 to B.
- *Perform*: A keeps \$10 for himself and allocates \$12 to B.

There are two types of Random Device

- Random Device 5/6: A is able to choose something other than *Don't Perform* with probability 5/6.
- Random Device 1/6: A is able to choose something other than *Don't Perform* with probability 1/6.

The players learn about the type of the random device after B has made his opt-out decision.

Step 1: Role assignment. At the beginning of each round, you will be anonymously and randomly matched with another participant. Each member of the pair will then be randomly assigned Role A or Role B with equal probability (50%).

Step 2: Communication. During the communication phase, Player A can send a chat message to Player B. Important: You are not allowed to reveal your identity to the other participant. (That is, you may not reveal your name or any other identifying feature such as race, gender, hair color, or seat number.) In every other respect, you are free to send any message you like. Please continue to remain quiet while communicating with the other participant. Participants who violate these rules (experimenter discretion) will be excluded from the experiment and all payments.

Step 3: Opt-out decision. Player B can decide whether to opt out. If B chooses to opt out, each player receives \$2. If B chooses not to opt out, the game continues. Information: Neither player knows, whether the Random Device determining if A will be able to choose

Perform is Random Device 5/6 (probability that A can choose something other than *Don't Perform* is 5/6) or Random Device 1/6 (probability that A can choose something other than *Don't Perform* is 1/6). However, both parties know that each scenario occurs with equal probability (50%).

Step 4: Nature of the Random Device revealed. The players learn whether they play with Random Device 5/6 or Random Device 1/6.

Step 5: Guessing. Player B guesses which choice Player A is likely to make in Step 7. A guesses which payoff B expects to gain. Note that if B thinks that A intends to choose *Perform*—allocating \$12—B's expected payoff depends on what B has learned about the Random Device: The expected payoff is \$2 if the game is played with Random Device 1/6 ($12 \times 1/6 = 2$) and \$10 if the game is played with Random Device 5/6 ($12 \times 5/6 = 10$).

Step 6: Player A learns whether he will be able to perform. If only *Don't Perform* is possible, the game ends. If A is able to perform, the game continues to Step 7.

Step 7: Decision phase. A decides whether to choose *Don't Perform* (keep \$14 and send \$0 to B), or whether to choose *Perform* (keep \$10 and send \$12 to B) or any of the options in between. The payoffs are

	A	B
A chooses *Don't Perform*	\$14	\$0
A chooses *1/4 Perform*	\$13	\$3
A chooses *1/2 Perform*	\$12	\$6
A chooses *3/4 Perform*	\$11	\$9
A chooses *Perform*	\$10	\$12
B chooses "Opt Out"	\$2	\$2
Performance not possible	\$14	0

Information at the end of a round. Players learn their own payoff, which random device was chosen, and the players learn whether player A was able to perform.

Conditional Choice. You will be asked to make the guess in Step 5 and the decision in Step 7 before Step 4 has actually been played. In other words, you will be asked to assume that A will be able to perform in Step 7, and then make the guess in Step 5 and the decision in Step 7 for two scenarios:

1. Random device 1/6 was chosen.

2. Random device 5/6 was chosen.

Subsequently, Steps 4 and 6 are played and A's recorded choice will be entered as A's decision in Step 7 (provided the game reaches this step). A's decision will influence payoffs as if A took the same decision in Step 7.

Bonus: Guessing. At certain points, you will have the additional possibility to earn a small amount by guessing the decisions of the other participant. Guessing will be paid in every round that is not chosen for payment of the decision. You will learn more about this during the experiment.

Do you have any questions?

C Proofs

We use Battigalli and Dufwenberg's (2007) general model of simple guilt to capture guilt aversion in our model. Applying their formulation to our game and notation yields the following utility function for the dictator

$$U_D = \pi_D(a) - \gamma_D \max\left\{E\left[\pi_R|\tau_D,\rho\right] - \pi_R\left(a\right), 0\right\}.$$

To this formulation we add the lexicographic structure of promise-keeping which is governed by μ in our model, and we allow for convex guilt to obtain

$$U_D = \pi_D\left(a\right) - \frac{\mu\gamma_D}{k}\left(\max\left\{E\left[\pi_R|\tau_D,\rho\right] - \pi_R\left(a\right), 0\right\}\right)^k$$

We will first prove **TR1** by showing that the dictator's equilibrium choice a^* after having given a promise is strictly increasing in ρ for sufficiently high guilt aversion and 0 otherwise. The prediction for $\mu = 0$ follows trivially from the dictator's utility function in expression (2). We distinguish two cases, $k = 1$ and $k > 1$.

Case $k = 1$. If $k = 1$, $U\left(a\right)$ is a linear function in a and $U'\left(a\right) = -4 + 12\gamma_D$. It follows that the equilibrium action a^* maximizing the dictator's utility is given by the following corner solutions:

$$a^* = \begin{cases} 0 & \text{if } \gamma_D \leq \frac{1}{3} \\ \rho\tau_D & \text{if } \gamma_D > \frac{1}{3}. \end{cases} \tag{4}$$

TR1 for $k = 1$ follows directly from (4).

Case $k > 1$. If $k > 1$, note that $U'\left(a\right) = -4 < 0$ on the interval $[\rho\tau_D, \infty)$. Therefore, the only candidate $\hat{a}_1$ for a maximizer of the dictator's utility function on the interval $[\rho\tau_D, \infty)$ is the corner solution $\rho\tau_D$, which is increasing in ρ and τ_D.

On the interval $[0, \rho\tau_D)$, note that the dictator's utility function is strictly concave as $U''\left(a\right) = -12^k\left(k-1\right)\gamma_D\left(\rho\tau_D - a\right)^{k-2} < 0$ for all $a \in [0, \rho\tau_D)$. First, assume that $U'\left(0\right) \leq 0$. Then, it follows from the concavity of the dictator's utility function that the $\hat{a}_2$ maximizing the dictator's utility function on the interval $[0, \rho\tau_D)$ is $\hat{a}_2 = 0$, which is independent of ρ. Second, assume that $U'\left(0\right) > 0$, which holds for sufficiently high guilt aversion, i.e.,

$$\gamma_D > \frac{4}{12^k\left(\rho\tau_D\right)^{k-1}}.$$

Then, assuming a maximizer $\hat{a}_3$ exists on $[0, \rho\tau_D)$, it must be an interior solution given by the following first-order condition:

$$\hat{a}_3 = \rho\tau_D - \left(\frac{4}{\gamma_D 12^k}\right)^{\frac{1}{k-1}}.$$

It can be seen that $\hat{a}_3$ increases in ρ, in τ_D, and γ_D, which proves **TR1**.

Beliefs and Second-order Effects on Performance So far we have implicitly assumed that the dictator's second-order beliefs are constant at τ_D. However, as the dictator's equilibrium choice a^* weakly increases in ρ, first- and second-order beliefs must adjust accordingly. Hence, τ_R and τ_D must be weakly increasing in ρ. This yields **TR2**.

As the dictator's equilibrium action is increasing in τ_D, this adjustment of beliefs leads to a (second-order) effect, reinforcing **TR1**.

Opt-out Decision A (risk-neutral) recipient will opt in if

$$E[\pi_R] = 12E[\rho\tau_R(\rho)] > 2.$$

If $\mu = 0$, the dictator will choose $a = 0$ and the recipient's beliefs will adjust accordingly. Hence, $\tau_R(\cdot)$ and therefore also $E[\pi_R] = E[\rho\tau_R(\rho)]$ will be higher for $\mu = 1$ than for $\mu = 0$. As a result, opt-out rates will be lower for $\mu = 1$ than for $\mu = 0$.

Example: Assume that $\tau_R(\cdot) = 0$ for $\mu = 0$, and $\tau_R(\cdot) = 1$ for $\mu = 1$. Then we would have $0 > 2$ for $\mu = 0$ and $12\binom{1/2}{1/2}\binom{1/6}{5/6} = 6 > 2$ for $\mu = 1$.