Promise keeping and reliance damage*

Arjun Sengupta[b,a], Christoph Vanberg[a]

[a] *Department of Economics, Heidelberg University*
[b] *Nottingham University Business School, University of Nottingham*

Abstract

This paper experimentally investigates the hypothesis that promise-keeping behavior is affected by the "reliance damage" that a counterpart would suffer as a result of a breach. Reliance damage is defined as the difference between the counterfactual benefit that the counterpart would have obtained had they not relied on the promise and that which they would obtain following a breach. We discuss two motivational mechanisms that could drive such an effect. One is that people intrinsically dislike causing reliance damage per se. The other is that people dislike causing regret in another person. We experimentally test these ideas in the context of an experimental trust game. Our evidence is consistent with the hypothesis that promise keeping is affected by reliance damage, and that the underlying mechanism involves a desire not to cause regret in others.

1. Introduction

Formal contracts which govern important economic transactions are often incomplete and leave room for contracting parties to behave opportunistically (Hart and Moore, 1988). As a result, factors supporting trust and trustworthiness may play an important role in promoting economic performance (Berg et al., 1995; Lorenz, 1999). The economic importance of trust and trustworthiness has been demonstrated by cross-country studies that show that measures of economic performance and development are positively correlated

*We thank Puja Bhattacharya and Kirby Nielsen and the seminar participants at the University of East Anglia and Heidelberg University for their valuable comments. We received helpful comments and suggestions from the associate editor and two anonymous referees.

with measures of trust and trustworthiness (Knack and Keefer, 1995, 1997; Zak and Knack, 2001; Whiteley, 2000). One particularly important factor that has been shown to promote trust and trustworthiness among strangers is the opportunity to forge non-binding informal agreements or exchange *promises*.

Given the economic significance of this phenomenon, it is natural for economists to study the factors explaining and promoting the effects of non-binding agreements and promises. Experimental economists have used simple laboratory experiments to investigate the underlying motivational mechanisms, predominantly in the context of trust or investment games (Berg et al., 1995).[1] Such games capture the strategic structure of many economic interactions in which promises are likely to play an important role. In these settings, a promise may help to persuade a potential partner to enter into a cooperative arrangement. Existing experimental evidence is consistent with the idea that promise-keeping is supported by a basic desire to fulfil *obligations* as well as an aversion to disappointing a counterpart's *expectations* (Ellingsen and Johannesson, 2004; Charness and Dufwenberg, 2006; Vanberg, 2008; Bhattacharya and Sengupta, 2016; Ederer and Stremitzer, 2017; Di Bartolomeo et al., 2019).

In this paper, we introduce and investigate an additional consideration which introspection suggests may play an important moderating role in promise keeping. We conjecture that a promisor (he) will compare the benefit that a counterpart (she) would obtain in case the promise is broken to what she *would have obtained* if she had not relied on his promise in the first place. We refer to this difference as "reliance damage".[2] Our main hypothesis is that a promisor will be more motivated to keep a promise the greater is

[1] In seminal work, Ostrom et al. (1992) studied the effects of non-binding verbal agreements in experimental common pool resource games.

[2] In many contexts, the concept of reliance damage will be exactly equivilant to that of an opportunity cost. In particular, this will be true whenever the payoff that a promisee gives up when relying on a promise is known to her when she makes the relevant choice. In contexts where this payoff is initially unknown (as will be the case in our experiment), a distinction must be made between the *ex ante expected* and the *ex post realized* value of the foregone opportunity. We thank a reviewer for pointing out that referring to the ex post realization of this value as an "opportunity cost" would constitute a potentially confusing abuse of the term. A prior version of this manuscript was entitled "Promises and opportunity cost" (Sengupta and Vanberg, 2020).

the reliance damage that would result from a breach.[3]

To illustrate, and to preview our experimental approach, consider the following scenario. Bob dreams of opening a pub but needs a partner in order to stem the necessary investments. His friend Ann has substantial savings that she had so far intended to invest in an MSCI World ETF. When Bob tells her of his plans, she tells him she is worried that he will not work hard enough to make the pub profitable for both of them. Bob promises that he will work hard, and Ann agrees to invest in the pub. A few months later, the pub is open and running well. After a particularly long and exhausting night behind the bar, Bob asks himself whether perhaps he should take it easy, even if it means less profit than promised for Ann.

In this situation, existing theories of promise-keeping suggest that Bob would ask himself questions like "how much money will Ann lose if I take it easy *as compared to if I work hard* like I promised?" and "how much money does Ann *expect* to earn, and how does *this* compare to what she will get if I take it easy?" In addition to this, we conjecture that Bob may ask "How much money *would Ann have earned* if she, instead of investing in the bar, had invested in the MSCI World ETF, and how does *this* compare to what she will receive if I take it easy?" Our main hypothesis is that the answer to this question is relevant to Bob. In particular, a testable and substantively interesting implication of our theory is that Bob's willingness to work hard and thereby fulfill his promise will depend on the realized value of the MSCI World ETF that Ann would have bought if she had not invested in Bob's bar. Specifically, if the return that she would have earned on the ETF is high, Bob will feel more obligated to fulfill his promise than if it is low.

Note that in this example, the realized payoff from the foregone opportunity (MSCI world) was *unknown* to Ann when she chose to invest in Bob's bar. Thus, our conjecture is distinct from the idea that Bob may feel more obligated to keep his promise when (and because) Ann has *chosen* to incur a greater cost, as might be predicted for example by theories of reciprocity or belief-based guilt aversion. Instead, our conjecture is that Bob will feel more motivated to keep his promise if the MSCI return is high even when this was unknown to Ann when she made her choice and therefore has *no relevance for*

[3]As we will emphasize later, our conjecture is distinct from the related idea that the promisor may be affected by the mere fact that the promisee has *chosen* to give up something of lesser or greater value, e.g., because this reveals something about their expectations, or because it makes their act of trust more or less "generous".

how he judges or interprets her choice. We conjecture two (potentially compatible) reasons why Bob might be concerned about the (initially uncertain) "reliance damage" implied by a breach of his promise.

Our first mechanism is based on the notion that people incur an intrinsic "moral cost" from breaking a promise. Ellingsen and Johannesson (2004) and others have modeled such a cost as fixed and independent of consequences. We conjecture that it could be moderated by what we are referring to as "reliance damage". Thus, the first mechanism we consider is that people intrinsically dislike causing reliance damage by breaking a promise. We will call this the "moral cost" mechanism.

Our second mechanism is based on the observation that a promisee is likely to experience *regret* if the outcome that obtains as a result of her having relied on a promise is worse than what they could have obtained from foregone opportunities (Bell, 1982; Loomes and Sugden, 1982; Zeelenberg et al., 1996). If Bob breaks his promise to work hard, Ann is likely to regret her decision to rely on him.[4] Moreover, she is likely to feel more regret if she learns (*ex post*) that the MSCI World ETF that she was planning to buy performed well. We conjecture that a promisor may care about the regret that a promisee will experience if a promise is broken. Such a concern might be based on a combination of *sympathy* and the wish to avoid *anger* directed at oneself due to betrayal of trust. That is, Bob may dislike causing "regret" in Ann simply because he does not wish her to experience a negative emotion, or because he believes that such emotion will cause her to think worse of him and/or to be angry with him for having broken his promise.[5] If Bob cares about Ann's regret, he will keep his word more often if her alternative investment performed well and *in addition* (he believes that) she is aware of this. We will call this the "second-order regret aversion" mechanism.

[4]The associate editor points out that there is a close link between this idea and the notion of "betrayal aversion": Ann's regret for having invested in Bob's bar is likely to be augmented by the feeling of having been betrayed by Bob. This suggests that the emotional "disutility" caused by a broken promise may be stronger than the regret generated in an equivalent situation where the bad outcome is due to bad luck. Indeed, Bohnet et al. 2008 show that individuals are less willing to take a strategic risk by trusting another person than they are to take a probabilistically equivalent "natural" risk.

[5]In the latter case, one might say that Bob is concerned about his "social image", as for example in Schütte and Thoma (2014); Grubiak et al. (2019). Our theory differs from these in that the degree to which Bob's image is harmed depends on the amount of "reliance damage" that he causes.

An important difference between the two conjectured mechanisms is that the former involves a "direct" concern for reliance damage *per se*, whereas the latter involves an "indirect" concern, since Bob cares about reliance damage only because - and to the extent that - it affects Ann's regret. A testable difference between these theories is therefore that the "second order regret aversion" mechanism can arise only in a setting where Ann (eventually) learns what she would have obtained if she had not relied on Bob to keep his promise. In contrast, the "direct" concern driving the "moral cost" mechanism can arise even if Ann remains uncertain about this foregone payoff.

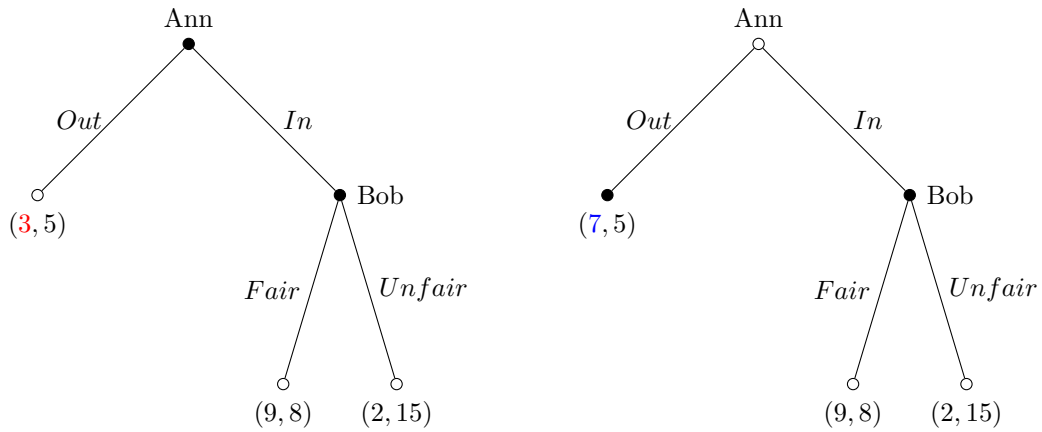
We test our theory in the context of an experimental investment game involving an uncertain payoff from the first mover's outside option. The key design feature is that the second mover learns the realized (but counterfactual) payoff from the outside option before making his own choice. In case a promise was made, the second mover is therefore aware of the "reliance damage" that would result if that promise is broken. We investigate (i) whether promise-keeping is affected by reliance damage, and if so (ii) which of the two motivational mechanisms introduced above can explain this behavior. Our results lend support to (i) and suggest that (ii) the underlying mechanism involves a desire not to cause regret in others.

The rest of the paper is organized as follows. In Section 2, we introduce the design, outline our predictions, and explain how our design rules out alternative explanations based on social preferences, reciprocity, or expectation-based guilt aversion. Section 2.3 presents the experimental procedure. Section 3 presents the empirical analysis. Section 4 concludes.

2. Design and Procedure

2.1. Design

Figure 1 presents two versions of a standard investment game that can be used to *illustrate* our theory but not to test it. Ann moves first and decides whether to *In*(vest) in Bob's project or choose *Out*. Choosing *In* increases the total available amount to be shared between Bob and Ann. However, if Ann chooses *In*, Bob decides the amount he wants to return. Bob can return a *Fair* amount, in which case both players receive similar payoffs, or, Bob can choose *Unfair*, in which case he keeps most of the money. In this game, Bob has both an incentive to promise to choose *Fair* if he can communicate with Ann, and an incentive to choose the *Unfair* amount if Ann chooses *In*. This, in turn, creates a disincentive for Ann to choose *In*.



(a) Low Outside Payoff

(b) High Outside Payoff

Figure 1: Investment Games

Note that the games presented in Figure 1 differ only in the payoff that Ann would receive if she chooses *Out*. An implication of our theory is that Bob's motivation to keep a promise to choose *Fair* will be affected by this foregone payoff. Specifically, both versions of the theory predict that Bob will be more likely to keep a promise to choose *Fair* in the game on the right, where Ann has foregone a payoff of €7, than in the game on the left, where she has foregone only €3. In the "moral cost" version, the reason is that Bob feels that he harms Ann more in the game on the right. In the "second order regret aversion" version, he anticipates that Ann will experience greater regret for having relied on his promise.

Although (both versions of) our theory make clear predictions for a comparison of promise-keeping behavior across the investment games in figure 1, such a comparison is not sufficient to distinguish our theory from others or between the two conjectured mechanisms.

The problem with such a comparison is that Ann is making a *different choice* in the two games, so that Bob may judge or interpret that choice differently. For example, it seems plausible that Ann's choice of *In* may induce higher *second-order beliefs* when the outside payoff that she knowingly foregoes is *High*. If so, the theory of *expectations-based guilt aversion* would

predict the same effect as our theory.[6] Similarly, Bob may feel more inclined to reciprocate Ann's trust if her outside payoff is *High*. Therefore, intuitive notions of trust and reciprocity may also predict that Bob will more often keep his promise in that case.[7]

For these reasons, we base our experiment on a modified investment game involving an *uncertain payoff* from the first mover's outside option. This modified investment game resembles our motivating story involving Ann's investment in an MSCI world ETF in that the payoff from the outside option is unknown to the first mover when she makes her choice, but becomes known to the second mover before he makes his. The central advantage of this design is that it introduces variation in reliance damage while keeping expectations and reciprocity motives (or *anything* related to how Bob judges or interprets Ann's choice of *In*) constant. In addition, our design will allow us to distinguish between the two mechanisms discussed above.

Figure 2 schematically represents our modified investment game which involves uncertainty in the outside payoff. At $t = 1$, Ann decides whether to choose *Out* or *In*(vest). After Ann makes her choice, Nature moves and flips a fair coin to determine Ann's payoff in case she chose Out. This outside payoff is either *High* (€7) or a *Low* (€3), with equal probability. Bob receives €5 if Ann chooses *Out*. If Ann chose *In*, Bob decides whether to split €17

[6]Stone and Stremitzer (2020) find evidence for an effect of opportunity costs on the expectations-based motivation for promise keeping.

[7]Perhaps surprisingly, standard formalizations of reciprocity preferences such as Rabin (1993); Dufwenberg and Kirchsteiger (2004) do not yield this intuitive prediction. Rabin's model does not predict trust and trustworthiness at all in games like ours, a fact that he acknowledged (ibid. p. 1296). Isoni and Sugden (2019) refer to this property of Rabin's and related models as the "paradox of trust". They trace its origins to the fact that such models are based on the idea of "reciprocal kindness". In Rabin's formalization, Ann's choosing *In* cannot be perceived as "kind" if she expects Bob to reciprocate, since in that case it is in her own material interest to trust him. The model of Dufwenberg and Kirchsteiger (2004) avoids the "paradox" through a modified definition of kindness. But even in that model, Ann's kindness does not depend on her own outside payoff. The intuition we mention corresponds closely to conventional but informal concepts of reciprocity used in the literature. For example, McCabe et al. (2003) argue informally that "the TR [Trust and Reciprocity] hypothesis predicts that the cooperative move in the positive opportunity cost games will generate greater reciprocity than the same move in the zero opportunity cost game." A conceptually consistent and plausible generalization of this idea would hold that a cooperative move in a *High* opportunity cost game will generate greater reciprocity than the same move in a *Low* opportunity cost game.

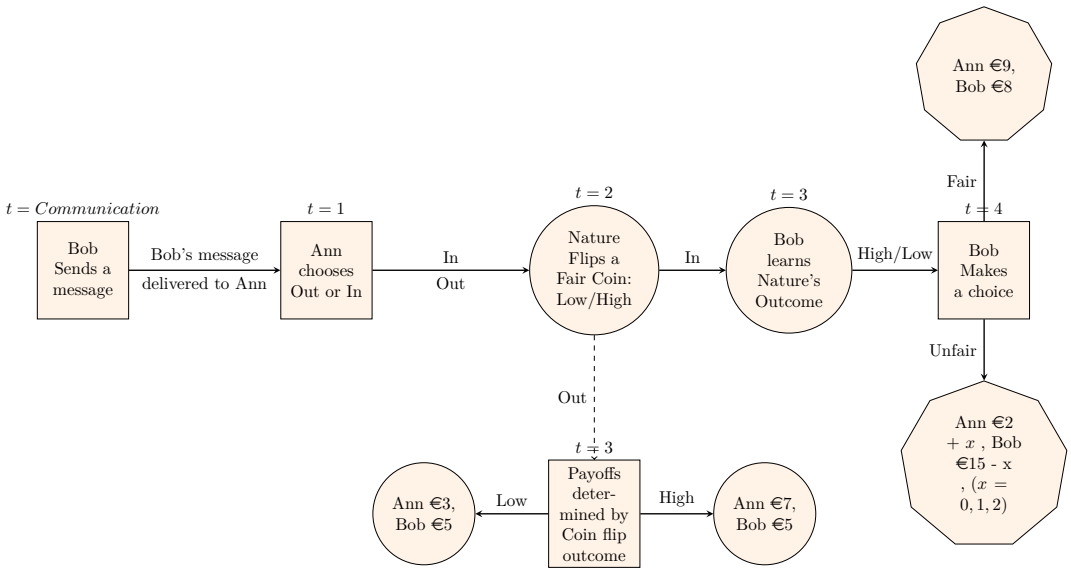


Figure 2: Modified Investment Game Timeline

in a *Fair* or an *Unfair* way. If Bob Chooses *Fair*, Ann receives €9 and Bob receives €8. If Bob Chooses *Unfair*, Ann receives €3 and Bob receives €14. A key feature of the design is that Bob observes the result of the coin flip before he makes his decision. That is, he knows the *realized* value of the opportunity that Ann has foregone by choosing *In*, and thus the "reliance damage" that he would cause if he breaks a promise to choose *Fair*.

Since the uncertainty regarding Nature's move is resolved only after Ann makes her choice, Ann's *expectation* concerning Bob's behavior should be unaffected by the realized value of the outside option. Since Bob knows this, we expect his *second-order expectation* to be independent of the potential reliance damage. Therefore, the theory of expectations-based guilt aversion does not predict an effect of reliance damage on Bob's choice.[8] Likewise, Bob has no reason to perceive Ann's choice as any more or less "kind" depending

[8]If Bob can communicate and issue a promise, Ann's expectations are likely to change. However, by the same logic, both first and second-order expectations should remain unaffected by the realized foregone payoff. Moreover, note that Bob also communicates before the uncertainty is resolved. This implies that the motivation to *issue* a promise, as well as the nature of that promise (wording, etc.) will be unaffected by the foregone payoff.

on her outside payoff, given that she was unaware of it when she made her choice. For this reason, theories of reciprocity also do not predict such an effect.

To test whether the potential reliance damage has the predicted effect on promise-keeping, and to disentangle the two mechanisms which could drive such an effect, we vary whether (i) Bob can send a message to Ann before Ann makes her decision (Communication vs. No Communication) and (ii) whether Ann learns the realized foregone payoff at the end of the interaction (Feedback vs. No Feedback). Table 1 shows all 4 treatments. Our main outcome variable of interest is the proportion of *Fair* choices made by subjects in the role of Bob. In Table 1, this proportion is denoted F, with a superscript indicating whether the foregone payoff is *H*igh or *L*ow and a subscript abbreviating the treatment (e.g. 'ncf' = No Communication and Feedback). We now lay out our hypotheses in terms of our main variable of interest, the proportion of *Fair (F)* choices.

Table 1: Treatments

	No Feedback	Feedback
No Communication	F^H_{ncnf}, F^L_{ncnf}	F^H_{ncf}, F^L_{ncf}
Communication	F^H_{cnf}, F^L_{cnf}	F^H_{cf}, F^L_{cf}

ncnf = no communication, no feedback; ncf = no communication feedback; cnf = communication no feedback; cf = communication feedback. For example, F^H_{cf} represents the proportion of Fair choices when the foregone payoff was high in the Communication Feedback treatment. There are 9 matching group level observations for each treatment. There are two observations for each matching group, one for low and one for high outside payoff.

2.2. Hypotheses

Our first hypothesis concerns the effects of the outside payoff on promise-keeping in a context where the first mover (Ann) learns the realized outside payoff following the interaction. Recall that in this case, we conjecture that the outside payoff may affect promise-keeping for two reasons. First, Bob may be concerned about the intensity of regret that Ann will experience if he breaks a promise to choose *Fair*. This intensity is likely to be increasing in

the realized value of her outside option, which in this treatment will become known to Ann ex-post. Second, Bob may be directly concerned about the reliance damage that he imposes on Ann, which corresponds to her foregone outside payoff. Note that both of these motivations work in the same direction when Ann learns the foregone payoff at the end of the interaction. Therefore, we expect that $F^H_{cf} > F^L_{cf}$.

Naturally, it is possible that Bob entertains similar thoughts even in the absence of a promise. That is, Bob may be more likely to chose *Fair* when the foregone payoff is large, even in our "No Communication" treatments ($F^H_{ncf} > F^L_{ncf}$). To test the idea that the foregone payoff increases the *effect of promises*, we test the hypothesis that the difference in the proportion of *Fair* choices between the low and the high outside payoff condition will be larger when individuals can communicate (and presumably promise) than when they cannot.

Hypothesis 1: *With Feedback, the difference in the proportion of Fair choices between the High and the Low outside payoff conditions will be larger in the Communication treatment than in the No Communication treatment.* $H_a : [(F^H_{cf} - F^L_{cf}) - (F^H_{ncf} - F^L_{ncf})] = \Delta_F > 0$

Without feedback, Ann never learns the cost she has incurred by choosing to Invest. Thus, although she may experience regret when she observes a broken promise, the intensity of her regret cannot depend on the realization of the outside payoff. As a consequence, Bob's second-order regret can also not dependend on this realization. Therefore, any effect of the outside payoff on promise keeping in the No Feedback treatments could be explained only by Bob being *directly* concerned about reliance damage, as suggested by the "moral cost" mechanism. Thus, the following hypothesis is consistent only with the "moral cost" version of our theory.

Hypothesis 2: *With No Feedback, the difference in the proportion of Fair choices between the High and the Low outside payoff conditions will be larger in the Communication treatment than in the No Communication treatment.* $H_a : [(F^H_{cnf} - F^L_{cnf}) - (F^H_{ncnf} - F^L_{ncnf})] = \Delta_{NF} > 0$

In case Hypothesis 1 is supported and Hypothesis 2 is not, we would conclude that the evidence supports our theory and that the underlying mechanisms appears to be the second order regret aversion. If both are supported, this would be consistent with both versions of our theory. In that case, we could attempt to decompose the effect according to the two motivational mechanisms proposed. If both are at work, Δ_{NF} would capture the effect attributable to the "moral cost" mechanism, while the *additional* difference,

$Z = \Delta_F - \Delta_{NF}$, would capture the part due to "second-order regret aversion". Thus, if both mechanisms have an important effect on promise keeping behavior, we should find support for the following hypothesis.

Hypothesis 3: *The effect of the outside payoff on promise-keeping behavior will be larger in the Feedback treatments (Δ_F) than in the No Feedback treatments (Δ_{NF}). $H_a : \Delta_F - \Delta_{NF} > 0$*

2.3. Experimental Procedure

The experiment was conducted at the experimental laboratory of the Alfred-Weber Institute of Economics at Heidelberg University in Germany. Subjects were students of the university and were recruited using Hroot and Sona. Each subject participated only in one of the four different treatments (cf,cnf,ncf,ncnf). Thus, the Feedback and Communication treatment conditions were varied *between subjects*. The High/Low outside payoff condition was randomly varied *within each of the sessions* (depending on the coin flips). We ran 3 sessions per treatment, with 18 subjects in each session. All Feedback treatments were run between 4-11th June 2019 and all No Feedback treatments were run between 9-18th November 2019. An average session lasted about 40-45 minutes. Subjects received on average a payment of 12 Euro, including a 3 Euro show up fee. 57.41% (62/108) of participants in the role of Bob were female, and the average age of the participants was 22.6 years. The experiment was programmed using Ztree (Fischbacher, 2007). In what follows, we will explain the procedure for the Communication and Feedback treatment in detail. The procedure for the other three treatments was identical except for the treatment variation.

Subjects were randomly assigned to a computer terminal in the lab. Instructions (in German) were displayed to the subjects on the computer screen. (Appendix A contains the English version of the instructions.) The instructions appeared over several pages, and the subjects could move forward and backward between the pages while reading the instructions. They were also given written instructions that remained available for later reference. When all subjects indicated that they had finished reading the instructions by pressing the "done reading" button which appeared on the last page of the instructions, the experiment proceeded to the main stage.

At the beginning of the experiment, each subject was randomly assigned to a matching group of size six. All interactions took place within the matching group. Half of the subjects in a matching group were assigned the role of Bob or Ann, respectively (referred to as "Participant A" and "Participant

B".). Their roles remained the same for the entire experiment. There were nine rounds. In each round, Ann and Bob were randomly matched within their matching group and they played the game represented in Figure 2.[9]

Within a round, Bob first had an opportunity to send a free form message to Ann. If he did not want to send a message, he could leave the message box blank. After Ann observed Bob's message, she decided whether to *Invest* or stay *Out* without knowing the payoff associated with "Out". Not knowing Ann's choice, participants in the role of Bob were asked to sequentially do two things. First, they flipped a (virtual, fair) coin to resolve Ann's payoff (High or Low) in case she chose "Out". Second, Bob chose between two options neutrally labeled Option 1 and 2 to determine both players' payoffs for the case that Ann chose "In". By making participant Bob perform these tasks in sequence, we made sure that Bob would know the value of the outside option without explicitly informing him of this value in a way that might induce demand effects.

After both participants made their choices, they stated their beliefs. Ann stated on a scale of 1-5 how confident she is that Bob is going to choose the *Fair* option, with 1 indicating that she is sure that he will not and 5 indicating that she is sure that he will. Bob was asked to guess the number Ann had stated. The belief elicitation was incentivized using a modified scoring rule, as in Vanberg (2008). (See pages 4 and 5 of the experimental instructions for details.) After the beliefs were stated, the round ended and a new round began. No information about others' choices or outcomes was given to players between rounds. Choice information and outcomes were displayed for all nine rounds at the end of the experiment. In the Feedback treatments, Ann also learned the realized value of the outside option.

If Ann chose *Out*, she received a payoff of €3 (Low outside payoof) or €7 (High outside payoff) and Bob received €5. If Ann chose *In* and Bob chose *Fair*, Ann received €9 and Bob received €8. We used three different payoff constellations for Ann and Bob if Ann chose *In* and Bob chose *Unfair*. Bob

[9]Since each matching group had six subjects and the game was repeated nine times, subjects interacted in the same pair more than once. The pairing was constructed such that the same pair never met twice in a row. No feedback about past choices or outcomes was given until the very end of the experiment. Thus, although subjects may have been able to identify in the communication treatments that they were again matched with the same person, they could not establish a reputation or react in any way to what others have done.

received either €13 or €14 or €15, and Ann received the remaining of the €17 that Bob was splitting. In a given round, only one of the three *Unfair* payoffs was used, but the *Unfair* payoff varied across rounds. The exact payoff structures were common knowledge before Ann and Bob made any decisions. They also knew that the payoffs resulting from the *Unfair* choice would change across rounds. The ordering of the *Unfair* choice payoffs over rounds was randomized across sessions. In addition to the payoffs from the game, subjects received a show-up fee of €3.

3. Results

In this section, we present the results of the experiment. Our main variables of interest are (i) the proportion of *Fair* choices by participants in the role of B and (ii) the difference in this proportion between the *High* and the *Low* outside payoff condition. In section 3.1.1, we present results from the Feedback treatments and show that the outside payoff affects promise-keeping behavior (supporting Hypothesis 1 and consistent with our main conjecture). In Section 3.1.2, we show that this effect vanishes in the absence of feedback (contradicting Hypothesis 2 and inconsistent with the "moral cost" mechanism). Together, these results constitute support for the "second order regret aversion" mechanism. (As indicated in Section 2.2, these results render an investigation of Hypothesis 3 redundant.).

We use non-parametric tests to test Hypotheses 1 and 2. All analyses are done at the matching group level, i.e., each independent observation is the proportion of Fair choices at the matching group level, conditional on whether the realized outside payoff was Low or High. We ran three sessions per treatment, each involving three matching groups of size 6. Thus, we have nine independent observations per treatment.[10] The nonparametric tests for Hypotheses 1 and 2 each use only data from the Feedback and No Feedback treatments, respectively. In section 3.1.3, we supplement these analyses using a linear regression based on the pooled data from all treatments.[11]

[10]The statistical analysis is done by pooling choices across all rounds for each matching group. Recall that the *Unfair* payoff for Ann and Bob varied across rounds. This was done in order to increase the expected variation in the binary dependent variable *Fair/Unfair*. Given the number of observations, we are under-powered to condition our analysis on the unfair payoff. For illustration, we present summary averages in Tables F.19 and F.20 in the Appendix.

[11]Since our main interest is in Bob's promise-keeping behavior, the statistical analysis

Table 2: Messages

	Strong Promise	Weak Promise	Promise	Empty
No Feedback	42.39%(103/243)	19.75%(48/243)	62.14%(151/243)	37.86%(92/243)
Feedback	63.79% (155/243)	11.11% (27/243)	74.90%(182/243)	25.10% (61/243)

3.1. Bob's Choice

3.1.1. Feedback

In the Communication treatments, Bob could send a free-form message to Ann before she made a decision. The messages were coded into three categories. They were coded as a *Strong Promise* if Bob clearly expressed his intent to choose the fair option (e.g. "let me choose. I will share 9 and 8 euro."). They were coded as a *Weak Promise* if the message merely suggests a choice of the fair option (e.g. "let me choose, we will get more money."). All other messages were classified as *Empty Talk* (e.g. blank messages or "Hello, how are you?").[12] Table 2 shows the number of messages in each category. For our statistical analyses, we consider both a weak and a strong promise as a promise.[13]

Figure 3 represents the percentages of *Fair* choices by subjects in the role of Bob in the treatments involving Feedback, i.e. those where Ann learns the value of the outside option at the very end of the experiment. In the Feedback Communication treatment, the proportion of *Fair* choices is 36.09% when the outside payoff is *Low* vs 54.64% when it is *High*, corresponding to an average difference of 18.55% points across all matching groups. Figure 4 displays the differences in the proportion of *Fair* choices between the Low and High outside payoff condition at the matching group level. The observations are

of Bob's second-order beliefs and Ann's choices and beliefs are presented in Appendix B and Appendix C, respectively.

[12]Messages were coded independently by one of the authors and a research assistant. The coders disagreed on 22 out of 486 observations, mostly on Weak Promises. The discrepancies were resolved together. Additional examples for messages and how they were coded are provided in Appendix E. Only one subject sent a conditional promise stating "In any case, I will decide that your payout is higher than if you tossed a coin. If the coin toss is €3 you get €4 from me. Otherwise €9". This message was sent 4 times and was coded as *Empty Talk*. Given that messages were restricted to 200 characters, and only one message can be sent, messages that can trigger closeness through aligned or mutual interests do not occur.

[13]Our results do not change if we restrict our data to only Strong Promises.

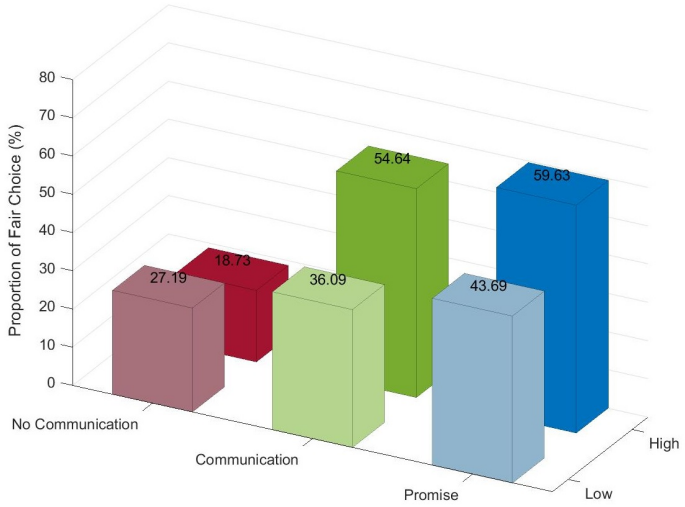


Figure 3: **Feedback -** Proportion of Fair choices

Note: Communication contains all observations from the Communication treatment. Promise contains only observations where a promise was made in the communication treatment.

arranged in ascending order, and the rank of observation is signed according to the sign of the observation. If the differences were randomly distributed, the observations in Figure 4 would have been equally distributed around zero. However, in the Feedback Communication treatment, the differences are positive for most of the matching groups. The median value is 14.29% points (vertical blue line) and is significantly different from zero (Sign-Rank, $p = 0.02$).

In the Feedback Communication treatment, 74.90% (182/243) of all messages were coded as a promise. If we restrict attention to the corresponding observations, the proportion of Fair choices is 43.69% and 59.63% in the Low and High outside payoff condition, respectively. The difference in the proportion of Fair choices between the High and the Low condition is 15.94% points. At the matching group level, the median value of this difference is 16.66% points, and it is significantly different from 0 (Sign-Rank, $p = 0.06$).

Result 1a: *In the Feedback Communication treatment, the proportion of Fair choices by subjects in the role of Bob is higher in the High than in the Low outside payoff condition.*

These observations are *consistent* with the conjecture that the motivation to keep a promise is affected by the degree of reliance damage, as measured by the realized value of a foregone opportunity. It is possible, however, that the outside payofgf would affect behavior even in the absence of communication and promises. In oder to conclude that reliance damage affects the force of promises, we need to rule out a similar effect in the No Communication treatment.

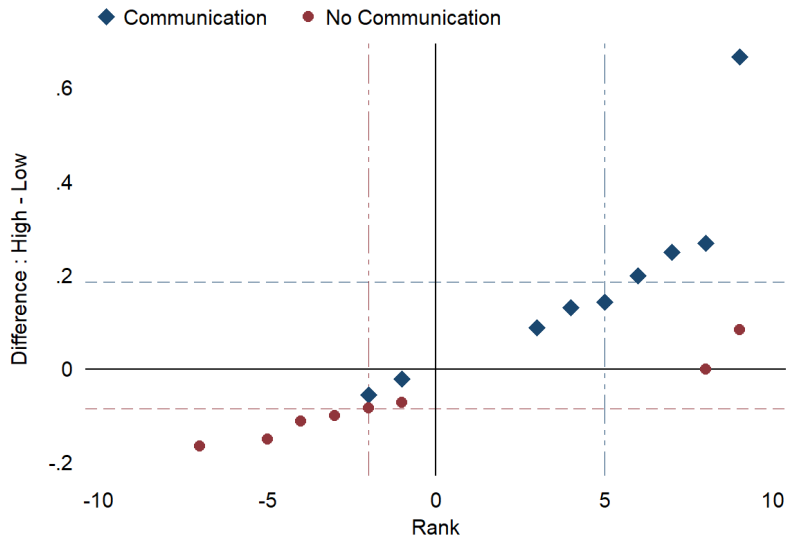


Figure 4: Distribution of differences (Feedback)

Note: Each observation represents a single matching group. The vertical axis measures the difference in the proportion of fair choices between high and low foregone payoff. The observations are ranked in ascending order within a treatment. The horizontal position of each observation corresponds to the signed-rank of that observation, where the sign of the rank is positive if the difference is positive and negative if the difference is negative. The dashed horizontal lines indicate the average of the observations. The dashed vertical lines mark the median observations.

In the Feedback No Communication treatment, the proportion of *Fair* choices was 27.19% when the outside payoff was Low and 18.73% when it was High. As can be seen in Figure 4, the difference is negative for most of the matching groups. The median is negative and significantly different from 0 (Sign-Rank, $p = 0.02$). This result is contrary to our expectation that subjects in the role of Bob may choose *Fair* more often when the outside

value is High even in the absence of communication.[14]

Result 1b: *In the Feedback No Communication treatment, the proportion of Fair choices by subjects in role B is lower when the outside payoff is High than when it is Low.*

Results 1a and 1b establish that the proportion of Fair choices increases with Ann's foregone outside payoff in the Communication Feedback treatment, but not in the No Communication Feedback treatment. To assess the significance of this difference in effects, we compare the differences between the High and Low conditions across the Communication and No Communication treatments. The average *difference in differences* (High vs. Low outside payoff) between the Communication and No Communication treatments is 27.03% points. Figure 4 shows that while the matching group level differences in the Communication treatment are mostly positive, they are mostly negative in the No Communication treatment. The distributions of the differences are significantly different between the Communication and the No communication treatments (Rank Sum, $p < 0.001$). If we restrict our data only to second movers that have issued a promise, then the difference in differences between the Communication and No Communication treatments is 24.41% and the distributions of the differences are also significantly different between the two (Rank Sum, $p = 0.004$). This suggests that the positive differences observed in the Feedback Communication treatment are indeed related to an *interaction effect* of motivations to keep promises and the cost of reliance that the counterpart has incurred.

[14]Since Bob does not know if Ann has chosen In when he makes his decision, a possible explanation for this result might be based on the idea that Bob might feel that he is choosing between two lotteries for Ann, $(\pi_{out}, p; \pi_{fair})$ and $(\pi_{out}, p; \pi_{unfair})$, where p is the probability that Ann chooses Out, π_{out} is her payoff in that event, and π_{fair}/π_{unfair} is the payoff that he assigns to her for the event that she chose In. Then a possible conjecture is that Bob is more willing to assign a lower payoff to Ann when π_{out} is large, since she "already" has a chance at a decent payoff. Holding every other motivation constant, this would mean that Bob would be more likely to choose the Fair option in the Low than in the High condition. As we use the strategy method in all our treatments, this motivation is present in all our treatments. Our difference in difference analysis therefore should take care that this motivation does not affect our results unless there is an interaction effect between communication and this motivation. Given that we see a significant positive difference in the fraction of Fair choices High vs. Low in the Communication treatment, we do not think that this can plausibly be accounted for by such an interaction effect, as it would require Bob to reason in an exactly opposite way when there is communication.

Further evidence of Hypothesis 1 is provided by the linear probability model in column 1 of Table 3.[15] The dependent variable is Bob's choice with 1 indicating a *Fair* choice and 0 otherwise. $High$ takes the value 1 in the High outside payoff condition. $Communication$ takes the value 1 in the Communication treatment. $HighXCommunication$ is the interaction term between the two. The coefficient on $High$ is negative, indicating that when individuals cannot communicate, they choose the Fair option more often in the Low than in the High condition. The coefficient on communication is also positive, indicating that the proportion of *Fair* choices is greater in the Communication treatment given a Low outside payoff, but this difference is not significant. Finally, our main variable of interest, the interaction term between $High$ and $Communication$, which measures the difference in differences in the proportion of Fair choices between the High and Low conditions moving from the No communication to the Communication treatment, is positive and significant, consistent with Hypothesis 1. We obtain qualitatively similar results if we restrict observations in the Communication treatment to only those who have sent a promise (Table D.12).

Result 1: *With Feedback, the difference in the proportion of Fair choices between the High and the Low outside payoff conditions is larger in the Communication treatment than in the No Communication treatment. This is consistent with Hypothesis 1 and the idea that Bob's motivation to keep a promise is enhanced by the reliance damage that would result from a breach.*

3.1.2. No Feedback

In the last section, we found that a High outside payoff leads to more promise-keeping. This pattern is consistent with both the *second-order regret-aversion* and the *moral cost* mechanism. In this section, we analyze the data from the *No Feedback* treatments, in which the second mover never learns the realized value of the outside option. In these treatments, the degree of reliance damage cannot affect the intensity of *regret* that subjects in the role of Ann experience if Bob chooses the *Unfair* option. Assuming that subjects in the role of Bob understand this, any effect of the outside payoff on their behavior could not be attributed to second order regret aversion, but would be consistent with the "moral cost" mechanism.

[15] To check for robustness, we also use a logistic and probit model with clustering at the matching group level. The results are reported in Table D.10 and D.11 and are qualitatively similar to the results reported in the main text.

Figure 5 represents the proportion of *Fair* choices in the No Feedback treatments. When Bob could communicate, the proportion of Fair choices is 58.50% when the outside payoff is Low and 50.05% when it is High. As can be seen from Figure 6, the matching group level differences are equally likely to be positive, negative or very close to zero, and the median difference does not differ significantly from 0 (Sign-Rank, $p = 0.26$).

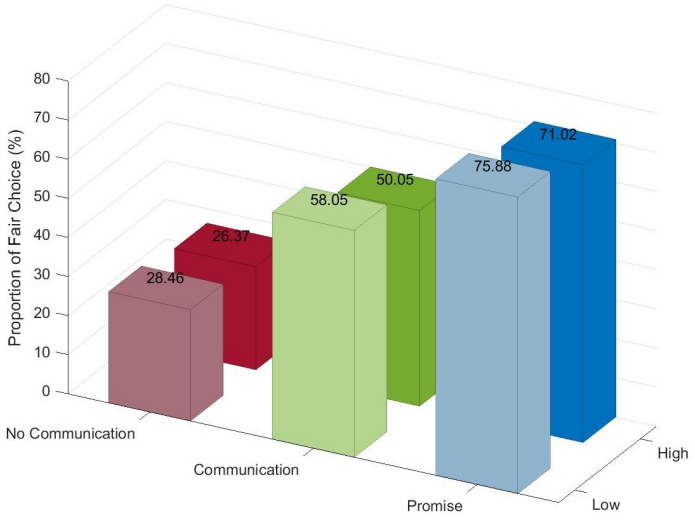


Figure 5: **No Feedback -** Proportion of Fair Choices

Note: Communication contains all observations from the Communication treatment. Promise contains only observations where a promise was made in the communication treatment.

In the No Feedback Communication treatment, 62.14% (151/243) of the messages sent were coded as a promise. If we restrict attention to only the corresponding observations, the proportion of Fair choices is 75.88% and 71.02% in the Low and High outside payoff conditions, respectively. Again, this difference is not statistically significant (Sign-Rank, $p = 0.44$).

Result 2a: *In the No Feedback Communication treatment, the proportion of Fair Choices does not differ between the High and Low outside payoff conditions, both in the full sample and among subjects who made a promise.*

Result 2a *suggests* that the degree of reliance damage did not affect promise keeping rates in the absence of feedback. To verify this interpretation, we also need to look at what happened in the treatment where Bob

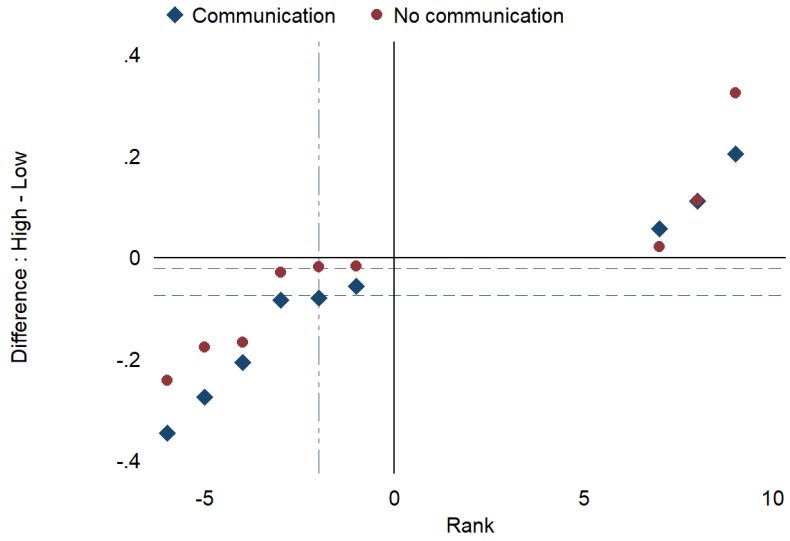


Figure 6: Distribution of differences (No Feedback)

did not have the opportunity to promise. In the No Feedback No Communication treatment, the proportion of *Fair* choices is 28.46% and 26.37% in the Low and High outside payoff conditions, respectively. This difference does not differ significantly from 0 (Sign-Rank, $p = 0.51$).

Result 2b: *In the No Feedback No Communication treatment, the proportion of Fair Choices does not differ between the High and Low outside payoff conditions.*

Finally, if we compare the differences in the proportion of *Fair* choices between the High and Low outside payoff conditions across the Communication and No Communication treatments, we cannot reject the null that the distributions of differences are identical across the two treatments (Rank Sum, $p = 0.45$). This result remains unchanged if we restrict the observations from the Communication treatment to those where Bob has issued a promise (Rank Sum, $p = 0.92$).

Further evidence of this observation is provided by regression specification II in Table 3. The only coefficient that is significant and positive is the coefficient on *Communication*. The coefficient on *High* and on the interaction term between *High* and *Communication* is not significantly different from zero. The results are similar if we restrict observations in the Com-

munication treatment to only those where Bob has sent a promise (Table D.12).

Result 2: *With No Feedback, the difference in the proportion of Fair choices between the High and the Low outside payoff conditions does not differ significantly between the Communication and the No Communication treatments. This is inconsistent with Hypothesis 2.*

Result 2 is inconsistent with the "moral cost" mechanism, i.e. the idea that the intrinsic cost of breaking a promise depends directly on reliance damage. Together, Results 1 and 2 suggests that the observed effect of reliance damage can be attributed mainly to a concern about the counterpart's regret for having relied on the promise.

3.1.3. Pooled Data Analysis

In this section, we analyze the pooled data using a linear probability model. This will serve as a robustness check of our non-parametric analyses and allows us to verify directly that the effect of the outside payoff on promise keeping is larger when there is feedback than when there is not.

Specification (3) in Table 3 reports the results of the pooled analysis. The variable $Feedback$ takes the value 1 in the Feedback treatments and 0 otherwise. The model includes interaction terms between $Feedback$ and $Communication$, $Feedback$ and $High$, and a triple interaction term between $High$, $Feedback$ and $Communication$. The interaction term of $High$ and $Communication$ captures the effect of the outside payoff on the promise-keeping rate in the No Feedback treatments. Thus, a positive and significant coefficient would be consistent with the "moral cost" mechanism. The triple interaction term captures the effect of the outside payoff on the promise-keeping rate in the Feedback treatment after controlling for the same effect in the No Feedback treatment. Thus, a positive and significant coefficient is consistent with the second-order regret aversion mechanism as a driver for increased promise keeping.

We find that the coefficient of the triple interaction term is positive and significant. Furthermore, the interaction term between $Communication$ and $High$ in specification (3) is not significant. This is consistent with the interpretation that the moral cost of breaking a promise does not vary with reliance damage. These two observations together confirm that promise keeping is affected by the outside payoff only when there is feedback. Thus the regression analysis using the pooled data, like the non-parametric analyses above, supports the "second order regret aversion" mechanism version of our

Table 3: Linear Probability Model for Bob's Choice of *Fair*

	(1) Fair	(2) Fair	(3) Fair
High	-0.087**	-0.032	-0.032
	(0.035)	(0.058)	(0.057)
Communication	0.084	0.306***	0.306***
	(0.096)	(0.075)	(0.074)
High*Communication	0.248***	-0.064	-0.064
	(0.069)	(0.083)	(0.082)
Feedback			-0.016
			(0.076)
Communication*Feedback			-0.222*
			(0.120)
High*Feedback			-0.055
			(0.067)
High*Communication*Feedback			0.312***
			(0.107)
Constant	0.271***	0.287***	0.287***
	(0.060)	(0.048)	(0.047)
Observations	486	486	972
Feedback	Yes	No	Combined
# of Clusters	18	18	36

The dependent variable $Fair$ takes the value 1 if Bob chooses the Fair option and zero otherwise. All specifications are clustered at the matching group level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

theory, and not the "moral cost" mechanism. Results are qualitatively similar if we restrict attention to only those who promised (Table D.12).

Result 3: *In our pooled data analysis, the coefficient of $HighXCommunicationXFeedback$ is significant and positive, while the coefficient of $CommunicationXHigh$ is not significant. Taken together, this is consistent with the idea that reliance damage affects promise keeping due to the "second order regret aversion" mechanism, but inconsistent with the "moral cost" mechanism.*

4. Discussion

We set out to investigate the idea that a person's motivation to keep a promise is enhanced by the "reliance damage" that a counterpart would suffer as a result of a breach. Our results are consistent with this idea, provided that the counterpart learns, *ex post*, how large her reliance damage actually was. The interpretation we suggest is that subjects dislike causing their counterpart to *regret* having relied on their promise. To the best of our knowledge, our paper is the first to present evidence for this phenomenon. We do not find evidence that reliance damage directly affects the intrinsic "moral cost" of breaking a promise in a setting where the counterpart does not know the value of the option foregone.

More broadly, the phenomenon described in this paper is an additional example of how *sunk costs* can significantly influence behavior. In the trust game context, similar effects are predicted by other theories, including reciprocity and belief based guilt aversion. These theories predict that the outside payoff that a counterpart foregoes may influence behavior because it signals something about their intentions and / or expectations.[16] Our paper takes a step forward by showing that another person's cost of reliance can influence behavior even when such signaling effects are excluded. Moreover, our evidence suggests that reliance costs are especially relevant if an individual has made a promise, perhaps because this causes him to feel *responsible*

[16]Olivola (2018) provide evidence from vignette studies for what they call an 'interpersonal sunk cost effect'. In their scenarios, the hypothetical counterpart in one condition incurred sunk cost and not in the other. They find that more people are willing to engage in activities that they do not prefer when the counterpart incurred sunk costs. Their results could be explained by reciprocating good intentions (signaling effect of sunk cost) of others.

for the costs incurred. Incorporating these patterns into a formal model of promise keeping based on second order regret aversion could present a fruitful challenge for future research.

An additional contribution that we want to highlight is that we introduce a trust game involving an uncertain payoff from the first mover's outside option. We are unaware of literature that specifically distinguishes this type of structure from the typical game in which the value of the outside option is fixed. However, uncertainty of this type appears to be a relevant feature in many important economic transactions. As an example, consider a firm hiring an employee. The firm is foregoing the opportunity to hire another applicant, while the employee is foregoing the opportunity to take another job. Both parties will typically face significant uncertainty as to how the other available alternatives would work out in the future. And both are likely to experience *regret* if they later learn that another opportunity would have produced a more favorable outcome. Finally, in this context and others like it, both formal and informal promises often are exchanged when the relationship is first being formed.

Our results suggest that in such situations, an individual's motivation to abide by informal agreements will be affected by (possibly updated) information about the opportunities that his counterpart has foregone by relying on him. For example, our main result suggests that an employee will be more motivated to abide by an informal agreement if he believes that his employer would otherwise regret having hired him. Conversely, the employer will be more motivated to live up to his part of such an agreement if she believes that the employee would otherwise regret having accepted the job. If true, this would have potentially interesting implications for the strategic use of information to influence the behavior of partners in various types of economic relationships. For example, the employer may consider giving the employee information about the quality of the pool of applicants from which he was chosen. If new information about the success of other applicants arrives (perhaps from another department), the employer may or may not want the employee to know this. Conversely, an employee may benefit from informing the employer about job offers that he has turned down. More generally, our main result suggests that individuals have an incentive to make their partners aware of the costs that they have incurred (and perhaps continue to incur) in order to enter (and remain in) the relationship.

Finally, our paper also relates to the literature on contract law. In contract law, a non-binding promise can be viewed as a binding promise en-

forceable by law if such a promise induced the promisee to take an action that he would otherwise not have taken (Restatement (second) of contracts § 90(1) (1979)). Under promissory estoppel, generally, two forms of penalties can be imposed on the party which is guilty of breaking the promise.[17] The promisor could be asked to pay "expectation damage", which requires the promisee to be put in the position she would have been in had the promise been kept. Or the promisor could be asked to pay "reliance damage", which requires the promisee to be restored to the position she would have been in had the promise never been made. Our results suggest that, even in the absence of legal punishment, people experience (or anticipate) an intrinsic psychological cost of a broken promise that closely corresponds to the legal concept of "reliance damage".

References

Damages for Breach of Contract. NYU Law. URL: `https://www.law.nyu.edu/sites/default/files/ECM_PRO_063763.pdf`.

Bell, D.E., 1982. Regret in decision making under uncertainty. Operations research 30, 961–981.

Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games and economic behavior 10, 122–142.

Bhattacharya, P., Sengupta, A., 2016. Promises and guilt. Available at SSRN 2904957 .

Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74, 1579–1601.

Di Bartolomeo, G., Dufwenberg, M., Papa, S., Passarelli, F., 2019. Promises, expectations & causation. Games and Economic Behavior 113, 137–146.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and economic behavior 47, 268–298.

[17] See https://www.law.nyu.edu/sites/default/files/ECM_PRO_063763.pdf for an exposition on the damages promisors pay when there is a breach in contract.

Ederer, F., Stremitzer, A., 2017. Promises and expectations. Games and Economic Behavior 106, 161–178.

Ellingsen, T., Johannesson, M., 2004. Promises, threats and fairness. The Economic Journal 114, 397–420.

Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. Experimental economics 10, 171–178.

Grubiak, K., et al., 2019. Exploring Image Motivation in Promise Keeping: An Experimental Investigation. Technical Report. School of Economics, University of East Anglia, Norwich, UK.

Hart, O., Moore, J., 1988. Incomplete contracts and renegotiation. Econometrica: Journal of the Econometric Society , 755–785.

Isoni, A., Sugden, R., 2019. Reciprocity and the paradox of trust in psychological game theory. Journal of Economic Behavior & Organization 167, 219–227.

Knack, S., Keefer, P., 1995. Institutions and economic performance: cross-country tests using alternative institutional measures. Economics & politics 7, 207–227.

Knack, S., Keefer, P., 1997. Does social capital have an economic payoff? a cross-country investigation. The Quarterly journal of economics 112, 1251–1288.

Loomes, G., Sugden, R., 1982. Regret theory: An alternative theory of rational choice under uncertainty. The economic journal 92, 805–824.

Lorenz, E., 1999. Trust, contract and economic cooperation. Cambridge journal of economics 23, 301–315.

McCabe, K.A., Rigdon, M.L., Smith, V.L., 2003. Positive reciprocity and intentions in trust games. Journal of Economic Behavior & Organization 52, 267–275.

Olivola, C.Y., 2018. The interpersonal sunk-cost effect. Psychological science 29, 1072–1083.

Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without a sword: Self-governance is possible. American Political Science Review 86, 404–417.

Rabin, M., 1993. Incorporating fairness into game theory and economics. The American economic review , 1281–1302.

Ross, L., Greene, D., House, P., 1977. The "false consensus effect": An egocentric bias in social perception and attribution processes. Journal of experimental social psychology 13, 279–301.

Schütte, M., Thoma, C., 2014. Promises and image concerns. Technical Report. SFB/TR 15 Discussion Paper.

Sengupta, A., Vanberg, C., 2020. Promises and opportunity cost. Technical Report. AWI Discussion Paper Series No. 692.

Stone, R., Stremitzer, A., 2020. Promises, reliance, and psychological lock-in. The Journal of Legal Studies 49, 33–72.

Vanberg, C., 2008. Why do people keep their promises? an experimental test of two explanations. Econometrica 76, 1467–1480.

Vanberg, C., 2019. A short note on the rationality of the false consensus effect. Technical Report. Discussion Paper Series.

Whiteley, P.F., 2000. Economic growth and social capital. Political studies 48, 443–466.

Zak, P.J., Knack, S., 2001. Trust and growth. The economic journal 111, 295–321.

Zeelenberg, M., Beattie, J., Van der Pligt, J., De Vries, N.K., 1996. Consequences of regret aversion: Effects of expected feedback on risky decision making. Organizational behavior and human decision processes 65, 148–158.

Appendix A. Instructions: Communication and Feedback (Lab Language: German)

On screen instructions

Page 0: (Displayed when subjects enter the lab)

Thank you for participating in this experiment. Please read the following instructions carefully. If you have a question, please raise your hand quietly.

General rules.

- This experiment will take approximately 60 minutes. During this time, you should not leave your station.
- Please turn off and stow away your phone. Starting now, there should be nothing on your table. (A drink is okay.)
- Please remain quiet for the duration of the experiment, and do not speak to other participants.
- At the end of the experiment, please remain at your station until your number is called. You will then receive your payment and sign a receipt.
- You will receive further instructions once all participants have been seated.

Page 1: Rounds, Roles, and Groups. Today's experiment consists of **9 Rounds** which are conducted **independently of one another**. This means that your decisions in any given round have no influence on what will happen in other rounds. At the end of the experiment, **one round will be randomly chosen for payment**. Your payment will depend only on the decisions made in this randomly chosen round.

At the beginning of the experiment, a **"role"** will be assigned to each participant. Half of the participants will receive **"Role A"**, the other half **"Role B"**. Your role will remain the same **for all 9 rounds** of the experiment.

At the beginning of each round, groups will be **randomly** formed, each consisting of one participant A and one participant B. (You will never be matched with the same participant twice in a row. It is possible that you will be matched to the same participant in a later round. However this is not predictable and the participants remain *anonymous* in all rounds.)

Page 2: Process within a round. **In general:** The process within a round is basically the same in all rounds. First, participant A chooses one of the following options:

1. **Option "Coin flip":** If this option is chosen, participant B will subsequently toss a (virtual) coin to determine the payments to the participants. Participant A will receive **7 EUR** in case of "Heads" and **3 EUR** in case of "Tails". (Heads and Tails are equally likely.) Participant B will receive **5 EUR** in both cases.
2. **Option "Participant B decides":** If this option is chosen, participant B will subsequently decide how **17 EUR** will be split between the participants. He will chose between**two available divisions** (for example **9 for A and 8 for B** or **4 for A and 13 for B**). Which divisions will be available to chose from will vary from round to round. Both participants will be informed about the available choices at the *beginning* of each round.

Details: In order to better understand the participants' decisions, we will ask participant B to determine an outcome for **both of the options** that participant A may chose. While participant A makes his choice, participant B will first toss a (virtual) coin and then chose between the two available divisions of 17 EUR. *Whether the coin toss or the chosen division counts depends on participant A's choice.* (At this point, participant B will not know what choice participant A made.)

Page 3: Communication at the beginning of a round. Before the participants make their decisions, participant B will have the opportunity to send a **message (maximum 200 characters)** to participant A. Participant B may not reveal his identity or identifying characteristics (e.g. "I am the person with the blue T-Shirt" or similar). Other than this, it is up to you to decide whether and how you use this opportunity.

Page 4: Bonus questions at the end of a round (Participant A). After the participants have made their decisions, they will have the opportunity to receive an additional payment by answering a question. Participant A will attempt to guess what decision participant B has made. And participant B will attempt to guess, how participant A answers that question. The concrete procedure works as follows:

- **Participant A** will report, on a scale of 1-5, how likely he feels it is that participant B chose the first or the second division. The bonus

payment that participant A will receive for this task depends on his report and on participant B's actual choice. This is summarized in the table below (all numbers represent EUR.)

	(1) certainly division 1	(2) probably division 1	(3) unsure	(4) probably division 2	(5) certainly division 2
B's choice is division 1	0.65 EUR	0.60 EUR	0.50 EUR	0.35 EUR	0.15 EUR
B's choice is division 2	0.15 EUR	0.35 EUR	0.50 EUR	0.60 EUR	0.65 EUR

(For example, if participant A feels certain that participant B chose division 2, he should report "Certainly division 2", because if participant B actually chose division 2, he will then receive the largest payment. However, he will then receive the smallest payment if participant B chose division 1 after all. Therefore, if participant A feels uncertain, he should consider the other reports as well.)

Page 5: Bonus questions at the end of a round (Participant B).

- **Participant B** will attempt, at the same time, to guess what answer participant A will give to this question. That is, participant B will also chose a column in the table. If he guesses correctly, he will receive 1 EUR.

	(1) certainly division 1	(2) probably division 1	(3) unsure	(4) probably division 2	(5) certainly division 2
B's choice is division 1	0.65 EUR	0.60 EUR	0.50 EUR	0.35 EUR	0.15 EUR
B's choice is division 2	0.15 EUR	0.35 EUR	0.50 EUR	0.60 EUR	0.65 EUR

(For example, if participant B chooses "(2) probably division 1", and if participant A actually chose "(2) probably division 1", then Participant B will receive 1 EUR. If participant A chose a different column, participant B will receive 0 EUR.)

At the end of the experiment, **one round** will be randomly chosen for payment of the bonus question. This round will be **different** from the round that is chosen for payment of the decisions.

Page 5: Summary. To summarize, the following steps are performed sequentially in each round:

1. Both participants learn what divisions will be available to participant B if participant A chooses the option "Participant B decides".
2. Participant B can send a message to participant A. If he does so, the message is immediately displayed to participant A.
3. Participant A chooses one of the options "Coin flip" or "Participant B decides". (B is not immediately informed of A's decision.) At the same time, participant B throws a (virtual) coin and chooses a division. (Whether the coin or his choice counts depends on A's decision.)
4. Both participants answer the bonus question.

After all groups have completed these steps, a new round will immediately begin. That is, you will not immediately be informed about the decisions made in your group, the result of the coin flip, or what payments were realized. You will receive this information only after all 9 rounds of the experiment have been completed. (See next page.)

Page 7.

Feedback at the end of the experiment. At the end of the experiment, you will receive a detailed summary of your decisions and the results in all 9 rounds.

- If participant A chose "coin toss", he will learn only the result of the coin toss, and not which division participant B would have chosen.
- If participant A chose "B choses", he will learn both the division that B chose and the result of the coin toss.

In addition, you will be informed about which round was chosen for payment of the decision and which round was chosen for payment of the bonus question.

Payment. Your payment (including 3 EUR show up fee) will be displayed on the final screen. Please enter this amount on your receipt and sign it. Then please wait quietly at your station until your number is called.

Appendix B. Bob's Second Order Beliefs

The experiment was designed to rule out the possibility that Ann's outside payoff might influence Bob's behavior indirectly through an effect on his second order beliefs. Since Bob knows that Ann is unaware of the outside payoff, we expected his second order belief to be independent of it. Here we verify that indeed the effects we observe are not driven by changes in second order beliefs.

Table B.4: Bob's Second Order Beliefs

		Low	High	Difference
Feedback	No Communication	2.38	2.11	-0.27 ($p = 0.02$)
	Communication	3.58	3.84	0.26 ($p = 0.14$)
	Promise	4.32	4.24	-0.08 ($p = 0.94$)
No Feedback	No Communication	2.19	1.99	-0.20 ($p = 0.19$)
	Communication	3.63	3.52	-0.11 ($p = 0.37$)
	Promise	3.97	4.22	0.25 ($p = 0.12$)

p- values reported are from Sign-Rank test.

Table B.4 summarizes the average second order beliefs reported by subjects in the role of Bob. We find a statistically significant difference between the High and Low outside payoff condition only in the Feedback No Communication treatment. All other differences are not significant. The same is true when we condition on a promise being sent in the Communication treatments.

Table B.5 shows the results of a regression that includes second order beliefs and an interaction term with Communication as explanatory variables. As can be seen, the observed variation in beliefs does not explain the effect of the outside payoff on promise keeping. The coefficient of $HighXCommunication$ in specification I and the coefficient of $HighXCommunicationXFeedback$ in specification III are still positive and significant.[18]

[18]Beliefs are positively correlated with the probability of a *Fair* choice. However, this need not be causal. See (Ross et al., 1977; Vanberg, 2019).

Table B.5: Linear Probability Model: Controlling for second order beliefs

	(1) Fair	(2) Fair	(3) Fair
High	-0.058*	-0.001	-0.003
	(0.031)	(0.046)	(0.045)
Communication	-0.031	0.055	0.099
	(0.171)	(0.192)	(0.152)
High*Communication	0.183**	-0.066	-0.067
	(0.065)	(0.076)	(0.073)
Belief	0.136***	0.161***	0.149***
	(0.029)	(0.021)	(0.018)
Belief*Communication	-0.013	0.003	-0.004
	(0.052)	(0.042)	(0.034)
Feedback			-0.042
			(0.065)
Communication*Feedback			-0.177
			(0.119)
High*Feedback			-0.052
			(0.054)
High*Communication*Feedback			0.242**
			(0.098)
Constant	-0.050	-0.063	-0.037
	(0.051)	(0.066)	(0.062)
Observations	486	486	972
Feedback	Yes	No	Combined
# of clusters	18	18	36

The dependent variable *Fair* takes the value 1 if Bob chooses the Fair option and zero otherwise. All specifications are clustered at the matching group level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Appendix C. Ann's Choices and Beliefs

Recall that subjects in the role of Ann made their decisions without knowing the outside payoff. Therefore we expected no differences in Ann's choices or beliefs between the Low and High outside payoff conditions. This is observed in our data. Table C.6 and C.7 show investment rates and average beliefs reported by Ann conditional on the realized outside payoff. In all four treatments, neither investment rates nor average beliefs differ significantly between the Low and High outside payoff conditions.

Table C.6: Ann's investment rates by treatment and outside payoff condition

		Low	High	Difference
Feedback	No Communication	39.48%	46.41%	6.93% ($p = 0.21$)
	Communication	61.78%	68.96%	7.18% ($p = 0.21$)
	Promise	69.51%	74.11%	4.60% ($p = 0.34$)
No Feedback	No Communication	39.03%	44.56%	5.53% ($p = 0.21$)
	Communication	55.53%	60.83%	5.30% ($p = 0.59$)
	Promise	72.60%	71.31%	-1.29% ($p = 0.68$)

p- values reported are from Sign-Rank test.

Table C.7: Ann's average belief by treatment and outside payoff condition

		Low	High	Difference
Feedback	No Communication	2.34	2.40	0.06 ($p = 0.81$)
	Communication	3.41	3.29	-0.11 ($p = 0.77$)
	Promise	3.76	3.62	-0.14 ($p = 0.39$)
No Feedback	No Communication	2.43	2.41	-0.02 ($p = 1.00$)
	Communication	3.21	3.21	0.00 ($p = 0.86$)
	Promise	3.57	3.64	0.06 ($p = 0.54$)

p- values reported are from Rank Sum test.

Having established that the outside payoff has no impact on investment rates and beliefs, Table C.8 and C.9 display these numbers pooling over the two outside payoff conditions. We observe that communication has a significant positive effect on investment rates and beliefs, irrespective of the Feedback condition. Feedback has no impact on Ann's stated beliefs or investment rates.

Table C.8: Ann's investment rate by treatment condition

	No Communication	Communication	Promise	Column Difference
No Feedback	41.15%	59.26%	70.00%	18.11 ($p = 0.03$)
Feedback	43.21%	62.96%	71.20%	19.75 ($p = 0.06$)
Row Difference	2.06 ($p = 0.89$)	3.70 ($p = 0.63$)		

p- values reported are from Rank Sum test.

Table C.9: Ann's average belief by treatment condition

	No Communication	Communication	Promise	Column Difference
No Feedback	2.41	3.23	3.48	0.82 ($p = 0.008$)
Feedback	2.37	3.38	3.71	1.00 ($p = 0.001$)
Row Difference	.04 ($p = 0.96$)	0.15 ($p = 0.82$)		

p- values reported are from Rank Sum test.

Appendix D. Additional Regressions

Appendix D.1. Logit and Probit Model

In section 3.1.3, we presented regression results from a linear probability model. To investigate the robustness of our results, Tables D.10 and D.11 summarize estimates using Logit and Probit models with errors clustered at the matching group level. The results are qualitatively similar to the linear probability model.

Table D.10: Logit Model for Bob's Choice of *Fair*

	(1) Fair	(2) Fair	(3) Fair
High	-0.500**	-0.164	-0.164
	(0.212)	(0.295)	(0.291)
Communication	0.392	1.286***	1.286***
	(0.444)	(0.334)	(0.329)
High*Communication	1.162***	-0.225	-0.225
	(0.324)	(0.379)	(0.374)
Feedback			-0.077
			(0.376)
Communication*Feedback			-0.894
			(0.547)
High*Feedback			-0.336
			(0.358)
High*Communication*Feedback			1.387***
			(0.491)
Constant	-0.988***	-0.911***	-0.911***
	(0.303)	(0.232)	(0.229)
Observations	486	486	972
Feedback	Yes	No	Combined
# of Clusters	18	18	36

The dependent variable *Fair* takes the value 1 if Bob chooses the Fair option and zero otherwise.. All specifications are clustered at the matching group level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table D.11: Probit Model for Bob's Choice of *Fair*

	(1) Fair	(2) Fair	(3) Fair
High	-0.291**	-0.098	-0.098
	(0.120)	(0.176)	(0.173)
Communication	0.238	0.797***	0.797***
	(0.269)	(0.204)	(0.201)
High*Communication	0.703***	-0.145	-0.145
	(0.194)	(0.230)	(0.227)
Feedback			-0.046
			(0.225)
Communication*Feedback			-0.559*
			(0.333)
High*Feedback			-0.193
			(0.210)
High*Communication*Feedback			0.849***
			(0.297)
Constant	-0.609***	-0.563***	-0.563***
	(0.181)	(0.139)	(0.137)
Observations	486	486	972
Feedback	Yes	No	Combined
# of clusters	18	18	36

The dependent variable *Fair* takes the value 1 if Bob chooses the Fair option and zero otherwise.. All specifications are clustered at the matching group level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Appendix D.2. Promise vs. No Promise

Our main analysis takes an intention-to-treat approach to deal with the problem that the making of a promise is endogenous. Therefore we compare Bob's behavior in the communication treatments, regardless of whether a promise was made, to the no communication treatments. An alternative approach (ignoring the endogeneity issue) would be to condition our analysis on a promise actually being made in the communication treatments. As a robustness check, we investigate two ways of approaching such an analysis: 1) a between-treatments approach and 2) a within-treatment approach.

Appendix D.2.1. Between Treatments Analysis (Promise vs. No Communication)

For a between-treatments analysis, we restrict the data from the Communication treatment to only those observations where Bob actually made a promise, and otherwise estimate models analogous to those presented thus far. In these regressions, the variable "Promise" takes the value zero only for observations coming from the No Communication treatments. Thus, the comparison is between subjects who made a promise in the Communication treatments vs. all subjects in the No Communication treatments.

Table D.12: Promise vs. No Communication (Linear Probability Model)

	(1) Fair	(2) Fair	(3) Fair
High	-0.087**	-0.032	-0.032
	(0.035)	(0.058)	(0.057)
Promise	0.114	0.438***	0.438***
	(0.106)	(0.083)	(0.081)
High*Promise	0.246***	-0.034	-0.034
	(0.065)	(0.097)	(0.095)
Feedback			-0.016
			(0.076)
Promise*Feedback			-0.324**
			(0.133)
High*Feedback			-0.055
			(0.067)
High*Promise*Feedback			0.280**
			(0.115)
Constant	0.271***	0.287***	0.287***
	(0.060)	(0.048)	(0.047)
Observations	420	394	814
Feedback	Yes	No	Combined
# of clusters	18	18	36

The dependent variable *Fair* takes the value 1 if Bob chooses the Fair option and zero otherwise. All specifications are clustered at the matching group level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table D.12 shows the Linear Probability Model. Table D.13, includes Bob's second-order beliefs as an independent variable. Tables D.14 and D.15 show the logit and probit estimations, respectively. In all cases, results are qualitatively similar to those obtained using the intent-to-treat approach which we regard as more appropriate.

Table D.13: Promise vs. No Communication (Linear Probability Model)

	(1)	(2)	(3)
	Fair	Fair	Fair
High	-0.058*	-0.001	-0.003
	(0.031)	(0.046)	(0.045)
Promise	-0.156	-0.068	0.021
	(0.193)	(0.236)	(0.179)
High*Promise	0.172**	-0.019	-0.023
	(0.061)	(0.085)	(0.086)
Belief	0.136***	0.161***	0.149***
	(0.029)	(0.021)	(0.018)
Belief*Promise	0.015	0.038	0.023
	(0.059)	(0.051)	(0.039)
Feedback			-0.042
			(0.065)
Promise*Feedback			-0.228*
			(0.119)
High*Feedback			-0.052
			(0.054)
High*Promise*Feedback			0.187*
			(0.103)
Constant	-0.050	-0.063	-0.037
	(0.051)	(0.066)	(0.062)
Observations	420	394	814
Feedback	Yes	No	Combined
# of clusters	18	18	36

The dependent variable *Fair* takes the value 1 if Bob chooses the Fair option and zero otherwise. All specifications are clustered at the matching group level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table D.14: Promise vs. No Communication (Logit Model)

	(1) Fair	(2) Fair	(3) Fair
High	-0.500**	-0.164	-0.164
	(0.212)	(0.295)	(0.291)
Promise	0.521	1.878***	1.878***
	(0.477)	(0.409)	(0.403)
High*Promise	1.144***	-0.146	-0.146
	(0.305)	(0.440)	(0.433)
Feedback			-0.077
			(0.376)
Promise*Feedback			-1.358**
			(0.619)
High*Feedback			-0.336
			(0.358)
High*Promise*Feedback			1.290**
			(0.527)
Constant	-0.988***	-0.911***	-0.911***
	(0.303)	(0.232)	(0.229)
Observations	420	394	814
Feedback	Yes	No	Combined
# of Clusters	18	18	36

The dependent variable *Fair* takes the value 1 if Bob chooses the Fair option and zero otherwise. All specifications are clustered at the matching group level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table D.15: Promise vs. No Communication (Probit Model)

	(1) Fair	(2) Fair	(3) Fair
High	-0.291**	-0.098	-0.098
	(0.120)	(0.176)	(0.173)
Promise	0.317	1.159***	1.159***
	(0.291)	(0.245)	(0.241)
High*Promise	0.693***	-0.090	-0.090
	(0.182)	(0.268)	(0.264)
Feedback			-0.046
			(0.225)
Promise*Feedback			-0.842**
			(0.375)
High*Feedback			-0.193
			(0.210)
High*Promise*Feedback			0.783**
			(0.319)
Constant	-0.609***	-0.563***	-0.563***
	(0.181)	(0.139)	(0.137)
Observations	420	394	814
Feedback	Yes	No	Combined
# of clusters	18	18	36

The dependent variable *Fair* takes the value 1 if Bob chooses the Fair option and zero otherwise. All specifications are clustered at the matching group level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Appendix D.2.2. Promise vs. No Promise within the Communication Treatments

As a final robustness check, we restrict attention to the communication treatments only and directly compare behavior in games where subjects in the role of Bob choose to promise vs. those where they did not. As might be expected, there is relatively little within-subject variation in the variable "promise". The majority of subjects either *always* or *never* issued a promise. Only slightly more than 1/3 did both. (See Table D.16.)

Table D.16: Promising behavior by Subject

	Never Promised	Sometimes Promised	Always Promised
Feedback (n=27)	3/27 (11.11%)	9/27 (33.33%)	15/27 (55.55%)
No Feeback (n=27)	6/27 (22.22%)	10/27 (37.03%)	11/27 (40.74%)

Table D.17: Promise vs. No Promise (Communication only, Linear Probability Model)

	(1) Fair	(2) Fair	(3) Fair	(4) Fair
High	0.066	0.005	0.098	0.115
	(0.120)	(0.067)	(0.103)	(0.124)
Promise	0.414**	0.570***	0.397**	0.582***
	(0.163)	(0.114)	(0.145)	(0.137)
High*Promise	0.054	0.008	0.097	-0.009
	(0.106)	(0.054)	(0.094)	(0.149)
Constant	0.199	-0.024	-0.440**	-0.601**
	(0.168)	(0.125)	(0.151)	(0.150)
Observations	243	243	81	90
Feedback	Yes	No	Yes	No
All Subjects	Yes	Yes	No	No
# of clusters	9	9	7	6

The dependent variable *Fair* takes the value 1 if Bob chooses the Fair option and zero otherwise. All Subjects 'No' means subjects who sent a promise at least once but not always. All specifications include subject fixed effect and are clustered at the matching group level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table D.17 shows the results from the regression analysis. We run the same regression specification as in the main text and additionally control for subject fixed effect. In columns (1) and (2), we run the analysis with all the subjects in the communication treatment. In column (3) and column (4), we include only those subjects who exhibit individual variation in promise making across experimental periods. The coefficients of our variables of interest are directionally similar to the between-subject analysis in the main text, but not statistically significant. This likely reflects the fact that these regressions are under-powered due to a lack of within-subject variation in promising.

Appendix E. Example Messages (translated from german)

Table E.18: Types of Messages

Promise	**Weak Promise**	**Empty**
B decides ->I choose Split 2	Let me choose, we will get more money than through the coin toss	
If I get to choose, I'll split the money in 9 for you and 8 for me	In Any Case, I Choose So That You are better off if you choose B	Hi, How are you?
I choose the most sensible breakdown where you get 9	You are Better Off If You Choose B	let us make 50/50
Hey :) If you choose B decides "I'll Choose Split 2 of which we both have the most.:) Do you trust me?	If You Let Me Decide, I Will Act according to the egalitarian maxim.	Hey I Would Take Split 1 - You can decide :)
I Will Choose Split 2 If You Choose Option B	Playing Fair!	Do What You Want And We'll See what comes out of it. :)

Appendix F. Variation of unfair payoff

We varied the payoffs associated with Bob's 'Unfair' choice between periods. In Table F.19, we look at the proportion of *Fair* choices, further conditioning our analysis on the unfair payoff. We find the same pattern as in the pooled data. In the Feedback Communication treatment, for each of the possible unfair payoffs, the proportion of fair choices was higher under the High compared to the Low outside payoff condition. If there was feedback but no opportunity to communicate, the pattern was reversed. Without feedback, the outside payoff did not impact the proportion fair choices, regardless of whether Bob could communicate or not.

Table F.19: Feedback: Variation of unfair payoff

		low	high	Difference = high - low
Unfair	Communication	42.53%	60.55%	18.01 ($p = 0.02$)
(4,13)	No Communication	25%	19.25%	-5.74 ($p = 0.02$)
Unfair	Communication	34.37%	41.25%	6.87 ($p = 0.02$)
(3,14)	No Communication	16.50%	14.81%	-1.69 ($p = 0.02$)
Unfair	Communication	26.81%	48.95%	22.54 ($p = 0.02$)
(2,15)	No Communication	34.20%	24.81%	-9.49 ($p = 0.02$)

Table F.20: No Feedback: Variation of unfair payoff

		low	high	Difference = high - low
Unfair	Communication	54.33%	53.33%	-1.0 ($p = 0.905$)
(4,13)	No Communication	25.55%	25.18%	-0.3 ($p = 0.81$)
Unfair	Communication	66.34%%	53.80%	-12.53 ($p = 0.44$)
(3,14)	No Communication	32.26%	17.51%	-14.74 ($p = 0.12$)
Unfair	Communication	50.18%	43.09%	-7.08 ($p = 0.44$)
(2,15)	No Communication	27.40%	32.77%	0.05($p = 0.67$)

Appendix G. Experiment Screen Shots

Figure G.7: B's Message Stage

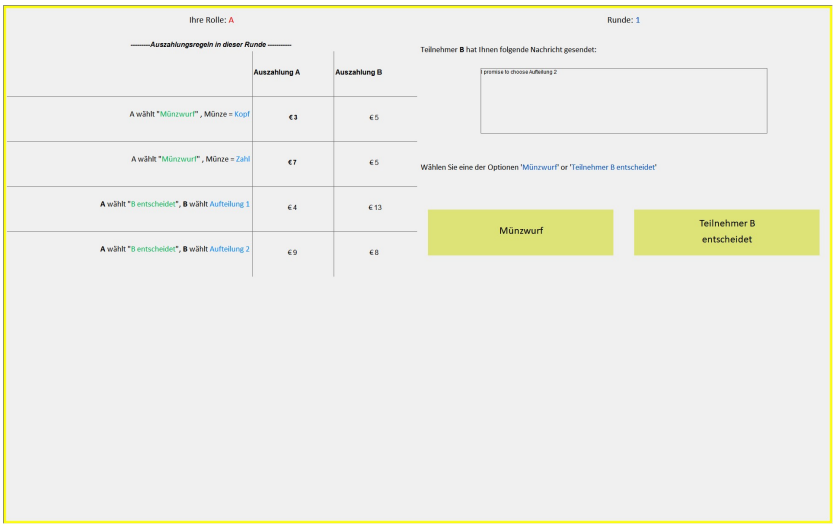


Figure G.8: A's Message and Choice Stage

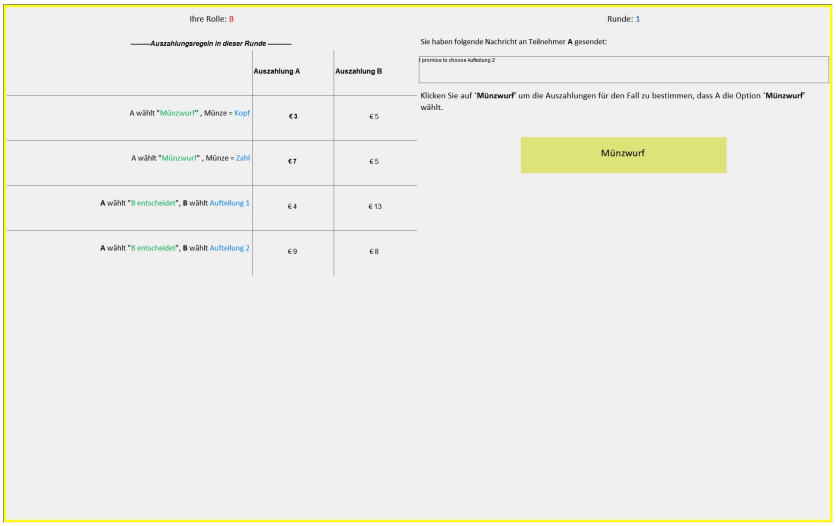


Figure G.9: B's Coin Flip Stage

Figure G.10: B's Choice Stage

Figure G.11: A's Belief Stage

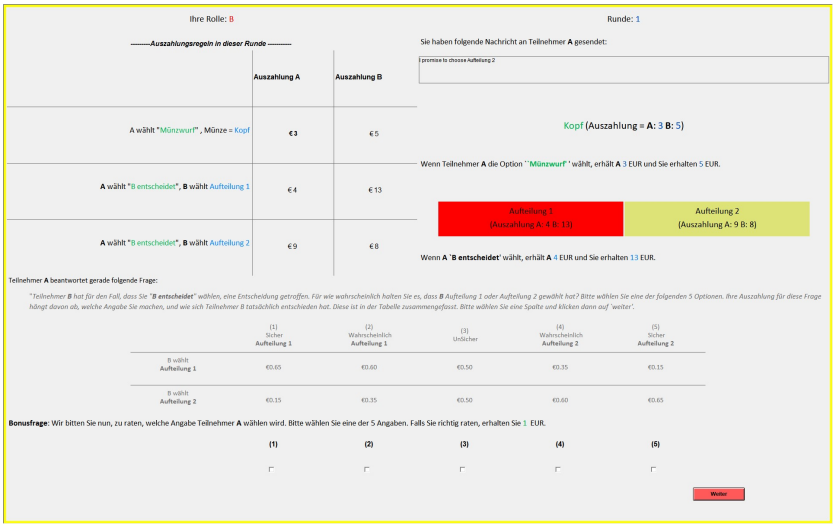


Figure G.12: B's Belief Stage