MRes Dissertation

Do you keep your promise because you want to keep it, or you just want to do what you have promised?

Jingwei Li

Supervised by Maria Montero and Robin Cubitt

Abstract

We keep our promises. However, the reason behind it remains unresolved. We try to explore the two dominant underlying motivations of it (EBE and CBE) using a novel experimental design. In addition, we compare the moral costs of violating different types of promise and separate self-selection effects in people's promise-keeping behavior.

1. Introduction

Considerable evidence has demonstrated that people are willing to keep their promises even when they are contrary to their self-interest. The natural question that arises is why promisers are reluctant to break their words. Dixit (2009) suggests three main reasons. The first one is the existence of a third-party enforcement mechanism, as studied in the formal contracting literature beginning with Mirrlees (1976) and Holmström (1979). The second one is the reputational incentive that arises when a party is concerned that reneging on a promise may hurt her benefit in the future, as studied extensively in the literature on relational contracting (Bull, 1987; Levin, 2003; Macaulay, 2018). The last reason for honoring an obligation, which is the focus of the present research (Ederer & Stremitzer, 2017), is the moral force of promise-keeping behavior. In the one-shot trust game without opportunities for binding contracting or reputation formation, conventional economic theory predicts no trusting because there is no incentive for trustworthiness (Ben-Ner & Putterman, 2009). According to rational man hypothesis, if people only care about their own income, then promise would be useless since violating it usually leads to a higher payoff (Chen & Zhang, 2021). Under this condition, the anticipated outcome of trust game is quite simple: because of selfishness, trustee would always prefer to keep all proceed rather than returning anything to trustor even they commit to do so; predicting this, trustor would keep all initial endowment and

trustee would not feel the need to give any commitment even there is an opportunity for pre-game communication as they know trustor would not believe it. However, beginning with the investment game of Berg et al. (1995), a string of experimental research, on the other hand, has found a common result that many individuals engage in trusting and trustworthy behaviors. Moreover, trustees are inclined to commit themselves to a promise to their counterparts if they are given the chance and follow their promises with a non-negligible frequency.

While the existence of moral force of promise-keeping behavior is undisputed, there is a multidisciplinary debate about why people tend to stick to their promises in the absence of contractual and reputational concerns. Two leading explanations have emerged in the literature.

- Expectation-based explanation (EBE): Charness and Dufwenberg (2006) first propose the expectation rationale for promise-keeping behavior: people are guilt averse. One would feel guilty from not fulfilling other people's expectation. Therefore, a promiser would avoid violating what she believes to be the expectation of her promisee.
- Commitment-based explanation (CBE): in contrast, CBE argues that people have a preference for promise-keeping behavior *per se*. People are inherently averse to default on their commitments. This is sometimes referred to as promise-breaking aversion or lying aversion in the literature.

This debate remains unresolved, with no consensus reached by far. Experimental evidence has been presented in favor of, respectively, EBE (e.g., Charness & Dufwenberg, 2006; Ederer & Stremitzer, 2017) and CBE (e.g., Vanberg, 2008). Furthermore, most of the discussion in this field is plagued by a simple dichotomy. In principle, however, the two explanations are not mutually exclusive. A more balanced view suggests that guilt aversion describes one important aspect of human motivation, while lying aversion describes another, with neither explanation showing the whole picture (Charness & Dufwenberg, 2010). A promiser may honor her promise to avoid disappointing others and to avoid suffering an additional moral cost of breaking a promise which is independent of other people's expectation at the same time, yet very limited research has attempted to disentangle the two underlying motivations within one experimental design. Our main aim is to fill this gap.

The type of promise is significant. Most of the research on promise focuses on free-form and voluntary promise (Chen & Zhang, 2021). In this common type of promise, senders of the message are free to determine the content of it (they can refrain from sending a message as well). Those who choose to give a commitment are identified as promisers, and their messages are considered as promises. However, not all promises are of this type. Promisers are sometimes nudged or elicited to do so, by the other player (e.g., Ismayilov & Potters, 2017), the experimenter (e.g., Charness & Dufwenberg, 2010), or a third party (e.g., Belot et al., 2010). While a considerable body of research has been carried out on volunteered promise, much less is known about elicited promise. Therefore, the second aim of our research is to enrich existing findings in this field. Specifically, we experimentally explore if and why people obey their promises even the promises are elicited by the experimenter and compare the moral costs of breaking a volunteered and an elicited promise.

Last, an important issue is that promise is endogenous, and promisers are self-selected to be promisers. It is very plausible that people who choose to give a commitment are different from other people, which implies that they keep their words not only because of the possible losses from EBE and CBE, but may also because they are more willing to do what they have promised than non-promisers. This is consistent with the opinion of Ederer and Stremitzer (2017). They believe the difference in the behavior of promisers and non-promisers is a combination of expectation effect, commitment effect, and, selection effect. This self-selection problem is a common problem in this field; however, most of the research simply ignores it. Our last aim is to try to separate the effects of EBE and CBE while considering self-selection problem at the same time using the model of social preferences.

To conclude, the research questions are as follows.

- Why are promisers inclined to keep their voluntary and elicited promises?
- Is the moral cost of breaking a voluntary promise higher than that of breaking an elicited one?
- Do promisers keep their promises because they want to keep them, or they just want to do what they have promised?

2. Related literature and theory

2.1 Motivations of promise-keeping behavior

2.1.1 Expectation-based explanation (EBE)

Expectation-based explanation (EBE), one of the two dominant rationales, is based on the guilt aversion theory formulated by Charness and Dufwenberg (2006). Building on psychological game theory (Geanakoplos et al., 1989), C&D introduce and test a new behavioral motivation, which is referred to as guilt aversion. Its basic idea presumes that a decision maker would experience guilt if she believes she lets other people down. This implies that a promiser would be more likely to keep her promise if she believes her promisee expects her to keep her words in order to avoid bad feelings.

Consider the trust game in Figure 1. Names and choices anticipate the experimental design. Figures represent monetary payoff. There are two players, A (he) and B (she). First, A decides whether or not to opt out of the game. If A chooses *In*, the game will proceed to the next stage, otherwise, the game will end immediately and both players will get 5. In the second stage, B decides whether or not to roll a six-sided dice. If B chooses *Roll*, she will get 10 for sure, and the dice will determine the income of A (A will get 12 with a probability of 5/6 and 0 with a probability of 1/6); if B chooses *Not Roll*, she will get 14 while A will get nothing.

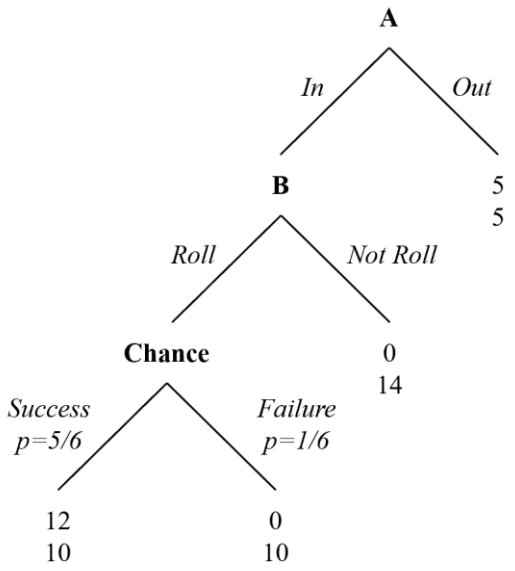


Figure 1. Trust game of Charness and Dufwenberg (2006)

C&D believes that B suffers from guilt to the extent she believes she hurts A relative to what A believes he will get, which means B is motivated by her belief about A's belief. Let $\pi_A \in [0,1]$ and $\pi_B \in [0,1]$ denote first-order belief of A and second-order belief

of B respectively: π_A is the probability of B choosing *Roll* in A's expectation, π_B is B's expectation regarding A's expectation. For A, he believes he will get $\pi_A \cdot[(5/6)\cdot 12+(1/6)\cdot 0]+(1-\pi_A)\cdot 0=10\pi_A$; for B, she believes A believes he will get $\pi_B \cdot[(5/6)\cdot 12+(1/6)\cdot 0]+(1-\pi_B)\cdot 0=10\pi_B$. Therefore, in B's opinion, if she chooses *Not Roll* rather than *Roll*, A will get 0 rather than $10\pi_B$, and as a result of this, B would feel guilty in proportion to $10\pi_B$, which leads to a non-standard concept of utility in the viewpoint of traditional game theory. Specifically, B would suffer from a disutility of $\gamma_B \cdot 10\pi_B$, where γ_B denotes the sensitivity to guilt of B, which is a stable personality trait that reflects the degree to which people are prone to shame and guilt (Tangney, 1995). Figure 2 models this. C&D simply assume that the guilt sensitivity varies among B and is independent of B's second-order belief π_B. In this case, guilt aversion theory provides a route by which promise makers adhere to their promises: by giving a promise to *Roll*, B strengthens A's expectation; if this is believed by B, then this strengthens the incentive for B to *Roll* as the disutility associated with *N*ot *R*oll increases; finally, promise makes (*In*, *Roll*) more frequently observed. By implementing a trust game experiment with measurement of individual's belief, C&D find the empirical relevance of this and experimental evidence consistent with EBE.

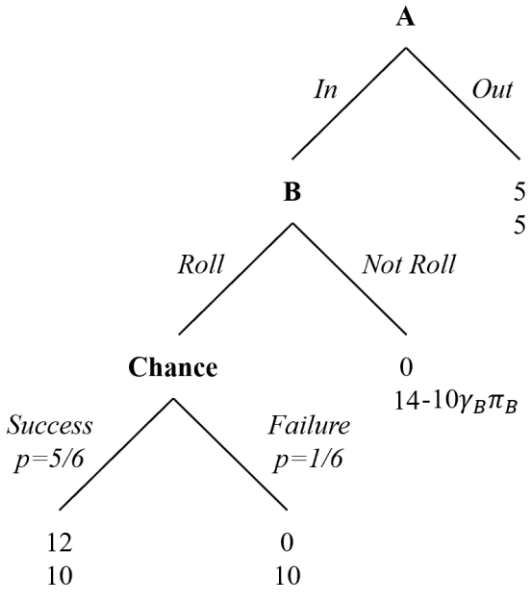


Figure 2. Adjusted trust game of Charness and Dufwenberg (2006)

However, C&D's findings can only lend limited support for EBE. Herein lie two potential problems. First, although C&D's experiment find that a promiser holds higher second-order belief as well as larger likelihood to play *Roll*, Di Bartolomeo et al. (2019)

criticize that their results could only reflect correlation rather than causation since they are randomly assigned neither to their messages (promises) nor to beliefs. Then, as a result of the first problem, C&D cannot independently evaluate EBE. For example, if promisers believe that promisees could predict their behavior accurately, then those who choose to stick to their promises because of CBE would report higher second-order beliefs as well (Bhattacharya & Sengupta, 2016). As C&D themselves say, their experiment is designed to test EBE but not to pit it against any alternative theory that may explain their results from a different perspective.

Some researchers try to test EBE as a cause of promise-keeping behavior while addressing the endogeneity problem of second-order belief. Bhattacharya and Sengupta (2016) allow promisees to purchase an insurance which can partly alleviate the worst payoff thereby introducing a signal from which give promisers the chance to infer the belief their promisees hold. Another methodology comes from Ederer and Stremitzer (2017). They propose an ingenious design, including a reliable and an unreliable random device that could exogenously induce expectations of both promisers and promisees. Their experimental results report support of EBE: exogenous increment in promiser's expectation leads to a significant improvement in their promise-keeping behavior. Furthermore, E&S find a conditional structure of guilt aversion: the sensitivity to other people's expectation is only switched on by making a promise. In another word, this new conditional guilt aversion theory argues that a promiser is influenced by her promisee's expectation but only if the expectation is supported by the promise her made, which could now nest some previous inconsistent results as special cases of EBE. However, although E&S offer evidence in favor of EBE, they do not isolate CBE in their experiment, which means CBE may play a role at the same time.

2.1.2 Commitment-based explanation (CBE)

Commitment-based explanation (CBE), the other leading rationale, posits that people have a preference for promise-keeping behavior *per se*, which is supported by some experimental results as well. For example, Vanberg (2008) proposes a novel methodology by implementing a partner-switching design and rematching half promisers and promisees. Vanberg finds that when being paired with a new partner who receives a promise from someone else instead of with their own recipient, promisers do not adhere to their promises, even though they know promisees are not aware of the rematch. This means that such behavior cannot be accounted for expectation as promisers know promisees have the same expectation under the two different

treatments. This implies people are inherently averse to default on their promises. However, Vanberg does not address the empirical relevance of EBE. Evidence of CBE cannot fundamentally disprove the significance of expectation since the inclination to keep a promise may still positively related to expectation. Furthermore, as mentioned before, this could be nested as a special case of conditional guilt aversion theory of Ederer and Stremitzer (2017): if the promiser is paired with a new player rather than her initial partner, then the promissory link breaks, and as a result of this, promisers in switching treatment do not care about their new partner's expectation anymore as conditional guilt aversion theory requires that the promiser is directly responsible in inducing an increase in her promisee's expectation. All of these indicate the necessity to manipulate the two motivational mechanisms respectively. In our experiment, we try to exogenously move promiser's second-order belief without breaking the promissory link to separate the effect of EBE. On the other hand, to isolate CBE, we have to design an experiment that varies the promissory commitment while keeping expectation unchanged.

Clearly, the empirical implications of the two mechanisms are substantially different. EBE combines guilt aversion theory with the idea that promise affects expectation, which means it requires not only the promiser makes the promise, but also the promisee learns about it. On the other hand, what CBE emphasizes is the intrinsic motivation. Here we use the same methodology as Ellingsen and Johannesson (2004), who first model this by introducing a personal cost of being inconsistent, to define the intrinsic moral cost of breaking a promise as fixed and independent of consequences. According to Di Bartolomeo et al. (2019), CBE implies that people are likely to maintain their commitments once they have given their words. This means a promiser's adherence to her promise does not depend on whether someone else may be affected by it or even knows it. Therefore, a natural way to manipulate the two motivations is to control the delivery time of promise.

To the best of our knowledge, Ismayilov and Potters (2016) are the first to exogenously vary the delivery condition to investigate the commitment rationale. Unlike the clear explanatory route provided by EBE (from belief to behavior), the interpretation under CBE remains ambiguous. According to I&P, this preference may derive from a more general preference for consistency (Ellingsen & Johannesson 2004), which means people wish to avoid any inconsistency between words and actions, or may be because promises establish a moral obligation, making promisers feel they should fulfill them

(Vanberg 2008), or they would suffer from a discomfort. I&P tweak the trust game of Charness and Dufwenberg (2006) by blocking the transmission of half messages. Social obligation predicts a promiser is more likely to keep her promise than a player who does not make a promise if and only if the promise is delivered successfully, while consistency argues that the adherence is irrespective of whether the promise is known to the promisee. Their initial results support the latter mechanism. However, I&P completely ignore the role EBE may play.

Inspired by them, and paying attention to EBE at the same time, we further control whether the promise is delivered before or after the promisee is able to take an action. Our key idea is the realization of social obligation does require the promise to be transmitted (Ismayilov & Potters, 2016), but when the transmission occurs does not matter. Therefore, the fulfillment of an on-time-arriving promise could be attributed to EBE and CBE, while the fulfillment of a late-arriving promise could only be ascribed to CBE. Here we simply define CBE as the integration of the two interpretations I&P propose, without attempting to compare the relative importance of them.

2.2 Voluntary and elicited promise

Elicited promise has not received enough attention in the literature by far. However, in everyday life, this type of promise is pervasive, encompassing situations such as oaths made during weddings, informal commitments to repay privately borrowed money, etc.; therefore, it is necessary to elucidate the influence of elicited promise from the standpoint of economic efficiency (Chen & Zhang, 2021).

A prevailing perspective in this field posits that only volunteered promises are effective in enhancing trust and trustworthiness. This means promisees are less likely to be convinced by an elicited promise, and promisers are less likely to uphold it either. On the one hand, compared with a voluntary promise, a promise that arises in response to an explicit question or request is not believed and is not expected to be believed. The follow-up experiment of Charness and Dufwenberg (2010) proves this: this type of promise would not affect the expectations by both sides. Therefore, EBE predicts weaker incentive to fulfill an elicited promise as the guilt the promiser would experience when breaking it is lower.

On the other hand, with respect to CBE, there exits two main contrasting viewpoints. Belot et al. (2010) suggest that the cost of lying is lower when the promise is elicited as people may feel compelled to make that promise. This provides an alternative explanation of why people are less likely to honor such an obligation. In contrary, Charness and Dufwenberg (2010) argues that a player who makes any type of promise to make a particular choice has to bear the cost of lying, if she subsequently takes a different action, which is the same whether the corresponding promise is full-blooded and carefully worded or pre-fabricated. In our perspective, we are more inclined toward the former one and attempt to experimentally test it. Specifically, we compare the levels of adherence of promisers who make different types of commitment, while controlling the level of second-order belief to ensure that EBE would not affect our comparison.

2.3 Self-selection effect

An important caveat is that promise is endogenous. This is to say, the message category is not randomly assigned, and a promiser is self-selected to be a promiser. Even in research on elicited promise, people are given the right to choose whether to send a promise or not. A problem is that the promiser's trustworthy behavior and *Roll* behavior are closely intertwined; therefore, promisers may be inherently more likely to play *Roll* (this will be identified as promise-keeping behavior) than those who refuse to give their commitments or even those who do not have the opportunity to give one. In another word, promises are just more likely to be sent by "*Rollers*" than by "*non-Rollers*" (Ismayilov & Potters, 2016).

We first explain why people may choose an action that does not maximize their own monetary income when the action has an impact on other people's payoff using the model of social preferences. Consider the simplified trust game of Charness and Dufwenberg (2006) (see Figure 3). Here we have three non-standard concepts of utility. $10\gamma_B\pi_B$ and μ_B that depend on the type of promise come from EBE and CBE respectively. Note that the two factors will only be taken into account by those who make the commitment first. In addition, α_B denotes the non-monetary benefit B would gain from choosing *Roll*. α_B may derive from concerns for altruism, reciprocity, efficiency, inequity aversion, etc. For example, compared with *Not Roll*, *Roll* seems to be a more altruistic choice with narrower distribution gap, both of which may motivate B to sacrifice some economic benefit. In addition, B may be more likely to choose *Roll* for the reason of reciprocity as well. By choosing *In*, A improves B's situation, so B may choose *Roll* to express her gratitude. Note that these considerations differ across

individuals but are independent of whether a commitment is made or not.

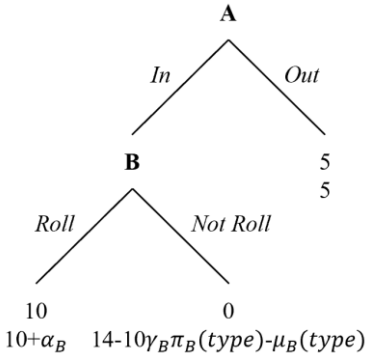


Figure 3. Simplified trust game of Charness and Dufwenberg (2006)

Now consider the trade-off between *Roll* and *Not Roll*. For promisers, the condition for them to choose *Roll* (namely, the condition for keeping their promises) is:

$$10+\alpha_{promiser}>14-10\gamma_{promiser}\pi_{promiser}(type)-\mu_{promiser}(type)$$

For those who do not give any commitment and those who do not have the opportunity to give one (we define this as the baseline treatment without pre-game communication), the conditions are:

$$10+\alpha_{non-promiser}>14$$

$$10+\alpha_{baseline}>14$$

By comparing these conditions, it is obvious that a promiser honor her promise due to three possible reasons: losses from EBE, losses from CBE, and a higher social preference propensity. This is consistent with the opinion of Ederer and Stremitzer (2017). They believe the difference in the behavior of promisers and players who send an empty talk or nothing is a combination of selection effect (subjects who promise are inherently different from those who do not), expectation effect (i.e., EBE, subjects care about their partner's expectation that could affect their own feelings and utilities), commitment effect (i.e., CBE, subjects feel compelled to contribute just because of the promise *per se*). To separate the effect of EBE as accurately as possible, we introduce a random device in our experiment that exogenously determines the timing of promise delivery. This device could exogenously induce promiser's second-order expectation while keeping the promissory link and social preference propensity unchanged to controlling the other two effects; therefore, we could get a relatively clear causal effect of EBE. However, the effect of CBE is harder to manipulate. Self-selection problem always exists in the experiment as we can never force our subjects to make a

commitment. Actually, this problem that promise is endogenous is not unique to our experiment. To try to make this point clearer, we compare α in different groups. By comparing the actions of promisers and non-promisers in baseline treatment without pre-game communication, we can compare $\alpha_{promiser}$ and $\alpha_{non-promiser}$ since social preferences are stable over a long period of time (Carlsson et al., 2014), which means they would not change in different treatments. Higher *Roll* rate implies larger α. If $\alpha_{promiser}$ is significantly larger, then we cannot exclude self-selection effect, i.e., we cannot know the better performance of promiser is because of CBE or self-selection or both. If this is the case, we have an alternative methodology to compare the moral costs of breaking different types of promise. First consider player's actions in the two period as an integrated strategy as Ismayilov and Potters (2017). There are four strategies in total: (*P*,*R*), (*P*,*NR*), (*NP*,*R*), (*NP*,*NR*). *P* represents giving a promise, *R* presents playing *Roll*, and *N* represents the opposite behavior. The expected utility of each strategy writes as followed.

- (*P*,*R*): $\pi_P(type)\cdot(10+\alpha_B)+(1-\pi_P(type))\cdot 5=(5+\alpha_B)\cdot\pi_P(type)+5$
- (*P*,*NR*): $\pi_P(type)\cdot(14-10\gamma_B\pi_B(type)-\mu_B(type))+(1-\pi_P(type))\cdot 5=9\pi_P(type)-(10\gamma_B\pi_B(type)+\mu_B(type))\cdot\pi_P(type)+5$
- (*NP*,*R*): $\pi_N\cdot(10+\alpha_B)+(1-\pi_N)\cdot 5=(5+\alpha_B)\cdot\pi_N+5$
- (*NP*,*NR*): $\pi_N\cdot 14+(1-\pi_N)\cdot 5=9\pi_N+5$

$\pi_P(type)$ is the probability of A choosing *In* when receiving a promise, and π_N is the probability of A choosing *In* when not receiving a promise. $\pi_B(type)$ is B's second-order belief of choosing *Roll*, $10\gamma_B\pi_B(type)$ and $\mu_B(type)$ are losses in utility from EBE and CBE respectively. Then consider the trade-off between (*P*,*NR*) and (*NP*,*NR*). It is obvious that $\pi_P(type)$ is higher than π_N (Ismayilov & Potters, 2017), which means an insincere promise would increase one's expected utility as it could enhance the likelihood of her counterpart choosing *In*. However, on the other hand, making a promise that will not be fulfilled make the promiser feel uncomfortable at the same time: she would suffer from two types of disutility. The key idea is that if the sum of the two types of disutility is large enough, then those B who do not want to *Roll* would not choose to make a fake promise. Our key design is to prevent promisees from knowing the type of promise to eliminate the effects of type on π_A, π_B and π_P. If promisees are only aware of the content of the promise, then they are expected to have same π_A and π_P, and as a result of this, π_B should be the same as well, which means the loss from EBE should be the same for a volunteered and an elicited promise. Now we can write the condition that prevents those B from making a volunteered fake promise:

$$(P,NR)<(NP,NR)$$

$$9\pi_P\text{-}(10\gamma_B\pi_B+\mu_B^V)\cdot\pi_P+5<9\pi_N+5$$

$$\mu_B^V>9\text{-}10\gamma_B\pi_B\text{-}9\cdot\frac{\pi_N}{\pi_P}$$

Similarly, the condition that prevents those B from making an elicited fake promise is:

$$9\pi_P\text{-}(10\gamma_B\pi_B+\mu_B^E)\cdot\pi_P+5<9\pi_N+5$$

$$\mu_B^E>9\text{-}10\gamma_B\pi_B\text{-}9\cdot\frac{\pi_N}{\pi_P}$$

Under our key design, it is obvious that the only way for the type of promise to affect B's behavior is through the moral cost μ_B. While controlling the expectation effect, we successfully control self-selection effect at the same time. Since subjects are randomly assigned to voluntary promise treatment and elicited promise treatment, the proportion of those who want to *Roll* and those who do not want to *Roll* in each promise treatment group should be approximately equal. For those who want to *Roll*, their dominant strategy is always to promise since $(5+\alpha_B)\cdot\pi_P+5>(5+\alpha_B)\cdot\pi_N+5$; for those who do not want to *Roll*, the greater the moral cost of breaking a promise, the more likely for the condition to be satisfied, the less likely those B would give a fake promise. Therefore, any significant difference in promise-giving rate can be attributed to different moral costs. Lower promise-giving rate implies higher moral cost. To the best of our knowledge, we are the first to compare moral costs of breaking a volunteered and an elicited promise through promise-giving rate.

3. Experimental design and hypotheses

3.1 Experimental design

Experimental economists use simple laboratory experiments to investigate the underlying motivational mechanisms of promise-keeping behavior, predominantly in the context of the investment game (Berg et al., 1995). The form of promise is generally written. This is because there may be many confounding and uncontrolled effects in face-to-face interaction (Roth, 1995) that we try to avoid. Our experiment adopts a simple version of the trust game of Charness and Dufwenberg (2006) (see Figure 4) with several differences. First, in our experiment, if B chooses *Roll*, then the outcome of A would be 10 GBP for sure rather than depending on the dice's result. This is mainly to exclude the effect of responsibility aversion: a preference to minimize one's causal role in outcome generation (Leonhardt et al., 2011). In some situations, decision makers prefer to delegate the decision to another person or to let a random device determine the outcome. Now we use *X* and *Y* instead of *Roll* and *Not Roll* to indicate the two

actions respectively, to avoid any farming effect as well. Second, we adjust the payoff of *X* from (10,10) to (9,10) to reduce the effect of inequity aversion: some B may choose the previous *X* as they have a preference to behave in a fair manner. Here we give B the higher monetary payoff as a distribution such as (10,11) may lower the reliability of the promise, which would eliminate the loss from EBE. Third, we vary the type of promise and the timing of promise delivery. B have an opportunity to send a message to their counterparts as in C&D's experiment; however, sometimes they cannot write their own messages (i.e., voluntary promise treatment) and can only choose from two pre-written ones (i.e., elicited promise treatment), and not all messages will be transmitted to their counterparts immediately. Here we introduce a random device that exogenously determine the timing of promise delivery. Half messages will be delivered before A is able to take an action (i.e., on-time delivery treatment), while the rest half will be delivered after that (i.e., delayed delivery treatment). Last, we add a baseline treatment where pre-game communication is forbidden. In this treatment, A and B simply decide their actions sequentially. Table 1 summarizes the 5 treatments in our experiment.

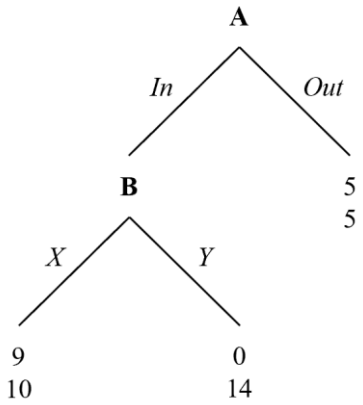


Figure 4. Modified trust game

Table 1. Treatments

		On-time delivery	Delayed delivery
Pre-game communication	Voluntary promise[1]	*VO*	*VD*
	Elicited promise[2]	*EO*	*ED*
No pre-game communication	Baseline	*B*	

1 In voluntary promise treatment, B can send a free-form message to their counterparts. By not restricting the content of it, we can learn which endogenous messages subjects prefer to use; however, they are not allowed to reveal any identifying information (name, gender, race, outfit, hair color, seat number, etc.). Subjects who violate this rule will be excluded from the experiment. In the case where B waive the opportunity themselves,

a blank message will be sent to their counterparts instead of no message at all. All messages collected from B will be identified by real persons according to if they contain a promise to play *X*. This yields three categories: (voluntary) promise, empty talk, blank message.

2 In elicited promise treatment, B can only choose from two pre-written messages: one contains a commitment and the other one is a negative message refusing to make a commitment. We use the latter one instead of a blank message as the experiment of Chen and Zhang (2021). If a commitment and a blank message are provided at the same time as the experiment of Charness and Dufwenberg (2010), and the receiver will not be aware of the type of message, then one might assume a promise made under such a condition is a volunteered one. We have a set of different pre-written messages. The references come from real messages collected in our own experiment.

Our key experimental design is to provide incomplete information to A. On the one hand, if no message is received by A, he will not be able to discern whether his counterpart attempts to send one that is intercepted by the device or his counterpart is not given the chance to send one. This is mainly to separate CBE by controlling subject's first-order and second-order beliefs in delayed delivery treatment and in baseline treatment. If A knows his counterpart has a message she wants to send to him, he may have a slightly higher expectation, which may have a minor influence on B's behavior as EBE predicts. On the other hand, to compare the moral costs of breaking a volunteered and an elicited promise, we have to prevent promisees from knowing the type of promise to eliminate the effect of type on expectation. In this case, the promisee would have the same level of belief when receiving two promises in different types; knowing this, according to EBE, the losses from EBE would be the same when reneging on different types of promise. This can be realized by only informing receivers of the content of the message without disclosing its type. When receiving the message (regardless of the content of it), A will only be told that this message is from his counterpart to make sure A cannot tell whether it is a message written or chosen by B. Note that A may default to the message as an endogenous and voluntary one by implementing this design, while this will not affect our result. To conclude, A cannot tell the difference between treatment *VO* and treatment *EO* and differences among treatment *VD*, treatment *ED* and baseline treatment.

We use a between-subject design in role-assigning. All subjects will be randomly assigned to the role of A or B with equal probability (50%), which will be told to them

immediately. There is no role-switching during the whole experiment and all subjects will play the same role in all rounds. This is because A with experience in playing the role of B may be able to distinguish between different types of commitments, which is what we do not want to see. We use a between-subject design in voluntary promise treatment and elicited promise treatment as well to avoid any possible effect of pre-written messages on content in the former treatment. With respect to delivery condition, we still exploit a between-subject design in on-time delivery treatment and delayed delivery treatment, while the participation in baseline treatment is compulsory. This means subjects will participate in one of treatment *VO*, treatment *VD*, treatment *EO*, treatment *ED* once and baseline treatment once respectively. Although this design may compel subjects to make different choices under different conditions, it gives us a measure of individual level consistency, which is crucial in our experiment since it allows us to compare moral costs while controlling the effect of self-selection. Last, since subjects are exposed to two treatments sequentially, there may be potential ordering effect in which scenarios are presented to them. To prevent our responses from being biased by this, the order will be randomized. Half subjects will participate in the baseline treatment first, while the other half will participate afterwards.

The experiment consists of two main sessions. In session 1, each pair of A and B will participate in voluntary promise treatment (i.e., subjects playing the role of B can write their own messages) once and baseline treatment once; in session 2, each pair will participate in elicited promise treatment (i.e., B can only choose from two pre-written messages) once and baseline treatment once. The purpose of this design is to provide reference for our pre-written promises in session 2. Some of the messages collected from B in voluntary promise treatment will be rewritten to pre-defined messages in elicited promise treatment for B to choose from. Here we provide different options rather than two fixed options as we want subjects to focus on the intrinsic variance (whether contains a promise or not) as much as possible and to minimize the impact of any other aspects of the message such as communication style on people's behavior. Moreover, randomization check of all pre-written messages should be conducted according to B's choices and A's reactions. However, due to limited fundings, the number of round attended by one subject may be adjusted.

Last, we collect first-order and second-order expectations during our experiment by letting subjects to make a guess about the behavior of their counterparts. Specifically, A will be asked to guess the probability of B choosing *X*, and B will be asked to guess

A's guess. Their answers represent our measurements for first-order belief of A and second-order belief of B respectively. Here we use a seven-point Likert scale and give B with an accurate guess a monetary reward. We do not ask them to state their beliefs in probability like most research does as we believe the mental scale is easier for them to understand and to choose with less bias, and we want to avoid any possible experimenter effect as experimenters can define the criteria for an accurate guess (i.e., the margin of error) themselves. In addition, A will not be rewarded in our experiment. Some research gives A an extra payoff as well if A's guess matches B's actual behavior; however, this may potentially change A's report by providing an opportunity to spread the risk. Consider the simplest binary scenario: A can choose between 1 (he believes his counterpart will choose *X*) and 0 (he believes his counterpart will choose *Y*). Let ω denotes the extra reward for an accurate statement, and π_A is the first-order belief of A. If A reports 1, he will get 10+ω with probability π_A and 0 with probability 1-π_A; if he reports 0, he will get 10 with probability π_A and ω with probability 1-π_A. In this case, a risk-averse A may report 0 to spread his risk by alleviating the worst payoff. Last, we will first collect their guesses and then collect their actual choices. The order is done deliberately to reduce any salient effect of action on reported belief.

3.2 Hypotheses

Our design allows us to disentangle the two motivational mechanisms of promise-keeping behavior within one experiment. First, since we exogenously manipulate first-order and second-order expectations by introducing a random device while preserving the promissory link and controlling the self-selection effect, we expect promisers are more likely to keep their promises in on-time delivery treatment than in delayed delivery treatment. This is the key hypothesis of expectation-based explanation (EBE).

Hypothesis 1 The promise-keeping rate is higher in on-time delivery treatment than in delayed delivery treatment.

This requires a precondition: promisers believe that promisees are more convinced that promisers will choose *X* in on-time delivery treatment than in delayed delivery treatment. It is worth mentioning that here we do not need a significant difference in the first-order belief under the two treatments to prove EBE as it is the second-order belief of B rather than the first-order belief of A that affect B's promise-keeping behavior. This leads to the second hypothesis.

Hypothesis 2 Promisers have higher second-order belief in on-time delivery treatment than in delayed delivery treatment.

Then consider commitment-based explanation (CBE). We first focus on the moral cost of breaking a volunteered and an elicited promise respectively. This leads to hypotheses 3.

Hypothesis 3A The *X* rate is higher in Treatment *VD* than in baseline treatment.

Hypothesis 3B The *X* rate is higher in Treatment *ED* than in baseline treatment.

To compare the moral costs of violating different types of promise, first we have to control the level of second-order expectation to ensure that EBE would not affect our comparison. In principle, this is satisfied in our experiment as the type of promise is not known to A, which is known to B. This leads to hypothesis 4.

Hypothesis 4 Promisers have the same level of second-order belief in Treatment *VO* and in Treatment *EO*.

According to Belot et al. (2010), people do not want to volunteer to lie but may have no compunction in lying if they feel compelled to do so, which means the moral cost of violating a promise is lower when the promiser is "forced" to make the promise; therefore, we assume the moral cost is lower when the promise is elicited by the experimenter. This leads to hypothesis 5.

Hypothesis 5 The promise-keeping rate is higher in *Treatment VO* than in *Treatment EO*.

Last, consider self-selection bias. The difference between a promiser's behavior and the behavior of a player in baseline treatment (in hypothesis 3), and the difference between a voluntary promise maker's performance and an elicited promise maker's performance (in hypothesis 5), may partially come from the subjects themselves, as promisers are self-selected to be promisers. A way to isolate self-selection effect is to compare the actions of different people in baseline treatment. This leads to hypotheses 6.

Hypothesis 6A In baseline treatment, there is no significant difference in *X* rate between voluntary promiser makers and the whole population.

Hypothesis 6B In baseline treatment, there is no significant difference in *X* rate between elicited promiser makers and the whole population.

Hypothesis 6C In baseline treatment, there is no significant difference in *X* rate between voluntary promiser makers and elicited promise makers.

Hypothesis 6A and 6B are preconditions of hypothesis 3A and 3B respectively, and hypothesis 6C is another precondition of hypothesis 5 (the first precondition of hypothesis 5 is hypothesis 4). If these preconditions do not hold, we cannot know the better performance is because of CBE or self-selection or both. In addition, for hypothesis 5, we have an alternative hypothesis to compare the moral costs of breaking different types of promise that does not need any precondition.

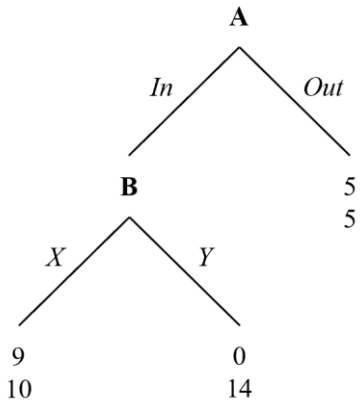
Hypothesis 7 The promise-giving rate is lower in voluntary promise treatment than in elicited promise treatment.

Appendix A. Instructions

Thank you for participating in this experiment. The purpose of this experiment is to study how individuals make decisions in a particular situation. Please read the instructions carefully and remain quiet during the whole experiment. In case you have any question at any time, please raise your hand. This experiment consists of 2 independent rounds. In each round, you will be paired with another participant who is randomly selected by the computer. You will never interact with a same person more than once. In addition, you will never know the identity of the persons with whom you interacted during the experiment. Depending on the decisions you made and your partners made, you have the chance to earn some money. At the end of the experiment, one of the two rounds will be randomly chosen for your payment. Each round is equally likely to be selected.

Overview

You will play the following game in every round. **Bold letters** represent roles. *Italic letters* represent choices you can choose from. Figures represent monetary payoffs in GBP: number in the first line is the payoff of A, and number in the second line is the payoff of B.



There are two players in this game, A and B. First, A can decide whether or not to opt out of the game. If A chooses *In*, the game will proceed to the next stage, otherwise, the game will end immediately and both players will get 5 GBP. In the second stage, B can choose *X* or *Y*. If B chooses *X*, B will get 10 GBP and A will get 9 GBP; if B chooses *Y*, B will get 14 GBP and A will get nothing.

You will be randomly assigned to the role of A or B with equal probability (50%). You will be informed of your role immediately. You will play this role during the whole experiment (2 rounds). Based on your role, you will be asked to perform different tasks.

You and your partner **may** have an opportunity for communication. However, the communication is one-way: only B can send a message to A over the computer. A cannot reply to it. In addition, there is a random device that determine the timing of message delivery. This device will generate one of the two following results with equal probability (50%).

- Result 1: the message will be delivered immediately. This is to say, A can receive and read the message before choosing *In* or *Out*.
- Result 2: the message will be delivered after A choosing *In* or *Out* and B choosing *X* or *Y*. This is to say, A will receive the delayed message at the same time as being informed the final payoff.

Each round consists of 3 steps, which are described below.

- **Step 1: communicating.** As mentioned before, B **may** have an opportunity to send a message to A, although the message has a probability of 50% to be postponed by our random device. Otherwise, you and your partner will start the game directly without any form of communication. It is possible that you face different scenarios in two rounds. Note that only B will be informed of the result of the device. This is to say, if you play the role of A and do not receive a message in the beginning, maybe you will receive one later, which means your partner's message is intercepted by the device, or maybe your partner is not given the chance to send one in this round.
- **Step 2: guessing.** You will be invited to make some guesses during the game. You may be rewarded for an accurate guess. You will learn more about this later.
- **Step 3: playing.** Then the game starts. After you and your partner make your choices, you will be told the final payoff of both of you immediately. In addition, if the message from B is intercepted by our random device, A will receive it now.

Do you have any questions?

Appendix B. Examples of pre-written messages in elicited promise treatment

Elicited promises:

Choose In and I will choose X.

Cooperate for maximum outcome?

Messages refusing to make a promise:

Good luck.

Choose whatever you want.

References

Belot, M., Bhaskar, V., & van de Ven, J. (2010). Promises and cooperation: Evidence from a TV game show. *Journal of Economic Behavior & Organization*, *73*(3), 396-405.

Ben-Ner, A., & Putterman, L. (2009). Trust, communication and contracts: An experiment. *Journal of Economic Behavior & Organization*, *70*(1-2), 106-121.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, *10*(1), 122-142.

Bhattacharya, P., & Sengupta, A. (2016). Promises and guilt. *Available at SSRN 2904957*.

Bull, C. (1987). The existence of self-enforcing implicit contracts. *The Quarterly journal of economics*, *102*(1), 147-159.

Carlsson, F., Johansson-Stenman, O., & Nam, P. K. (2014). Social preferences are stable over long periods of time. *Journal of public economics*, *117*, 104-114.

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579-1601.

Charness, G., & Dufwenberg, M. (2010). Bare promises: An experiment. *Economics letters*, *107*(2), 281-283.

Chen, Y., & Zhang, Y. (2021). Do elicited promises affect people's trust?—Observations in the trust game experiment. *Journal of Behavioral and Experimental Economics*, *93*, 101726.

Dixit, A. (2009). Governance institutions and economic activity. *American economic*

review, *99*(1), 5-24.

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

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*(495), 397-420.

Holmström, B. (1979). Moral hazard and observability. *The Bell journal of economics*, 74-91.

Ismayilov, H., & Potters, J. (2016). Why do promises affect trustworthiness, or do they?. *Experimental Economics*, *19*, 382-393.

Ismayilov, H., & Potters, J. (2017). Elicited vs. voluntary promises. *Journal of Economic Psychology*, *62*, 295-312.

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

Leonhardt, J. M., Keller, L. R., & Pechmann, C. (2011). Avoiding the risk of responsibility by seeking uncertainty: Responsibility aversion and preference for indirect agency when choosing for others. *Journal of Consumer Psychology*, *21*(4), 405-413.

Levin, J. (2003). Relational incentive contracts. *American Economic Review*, *93*(3), 835-857.

Macaulay, S. (2018). Non-contractual relations in business: A preliminary study. In *The Sociology of Economic Life* (pp. 198-212). Routledge.

Mirrlees, J. A. (1976). The optimal structure of incentives and authority within an organization. *The Bell Journal of Economics*, 105-131.

Roth, A. E. (1995). Bargaining experiments. *Handbook of experimental economics*, *1*, 253-348.

Tangney, J. P. (1995). Recent advances in the empirical study of shame and guilt. *American Behavioral Scientist*, *38*(8), 1132-1145.

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations 1. *Econometrica*, *76*(6), 1467-1480.