frontiers | Frontiers in Behavioral Economics

TYPE Original Research
PUBLISHED 03 September 2025
DOI 10.3389/frbhe.2025.1631806

OPEN ACCESS

EDITED BY
Joel Van Der Weele,
University of Amsterdam, Netherlands

REVIEWED BY
Vera Te Velde,
The University of Queensland, Australia
Egor Bronnikov,
University College Maastricht (UCM), Netherlands

*CORRESPONDENCE
Kevin P. Grubiak
✉ kevin.grubiak@uni-passau.de

RECEIVED 20 May 2025
ACCEPTED 21 July 2025
PUBLISHED 03 September 2025

CITATION
Grubiak KP (2025) Promises, image concerns, and excuses–An experimental investigation. *Front. Behav. Econ.* 4:1631806. doi: 10.3389/frbhe.2025.1631806

Promises, image concerns, and excuses–An experimental investigation

Kevin P. Grubiak*

School of Business, Economics and Information Systems, University of Passau, Passau, Germany

This paper tests the robustness of promise keeping in economic interactions using a laboratory experiment. Our design allows us to examine the roles of both social- and self-image concerns, and to investigate whether these concerns are diminished when participants are provided with responsibility-diffusing excuses. When the responsibility for a broken promise is undeniable, promise keeping is high. However, when plausible excuses are available that allow participants to preserve their social image, a significant number choose to break their promises. Yet, cooperation remains higher compared to treatments without a communication stage, and we find no evidence of participants engaging in self-deception to evade their promise-induced commitments. These findings suggest that while some individuals keep their promises reluctantly, others exhibit stable preferences for promise keeping that are not easily eroded by moral wiggle room.

KEYWORDS

communication, promises, image concerns, excuses, moral wiggle room

1 Introduction

Many economic and everyday interactions offer opportunities for mutual benefit, provided there is a foundation of trust and cooperation. Trust is often viewed as an essential prerequisite for initiating agreements, particularly in situations where formal contracts are unenforceable or too costly to implement. It serves as a social lubricant, reducing bureaucratic frictions and the need for excessive oversight, thereby enhancing overall efficiency. A key inhibitor of trust is its inherent risk of being betrayed by entrusted parties for private benefit. As a result, a substantial body of literature has developed to understand factors and circumstances that most effectively support trust-based interactions.

A prominent strand of experimental research, starting with the seminal work of Ellingsen and Johannesson (2004), examines the role played by non-binding verbal communication. In contrast to the standard economic model, which treats such communication as "cheap talk," it has frequently been observed that communication—and more specifically, the exchange of *promises*—exerts a remarkably strong effect on trust and cooperation. Various mechanisms may contribute to the strength of promises, including a commitment-based internal preference for keeping one's word (Vanberg, 2008), the desire to fulfill expectations generated by promises (Charness and Dufwenberg, 2006), the avoidance of reliance damage resulting from broken promises

(Sengupta and Vanberg, 2023), and an aversion to being seen as a promise breaker (Kingsuwankul et al., 2023; Lang and Schudy, 2023).[1]

This paper explores the robustness of promise keeping by providing laboratory participants with devices that can serve as excuses to obscure responsibility for broken promises. Excuses can alter the costs of breaking a promise in multiple ways. Most notably, excuses can prevent the damage inflicted on one's social image. In addition, not feeling personally responsible for breaking a promise allows to maintain a positive self-image and may even result in reduced sensitivity to guilt or reliance damage. Prior research has shown that people frequently use excuses across various social contexts (Gino et al., 2016), allowing them to be less altruistic (Dana et al., 2007; Exley, 2015), reciprocal (Malmendier et al., 2014; Regner, 2018), norm-enforcing (Kriss et al., 2016), or groupish (Robbett et al., 2024). An open question concerns the generalizability of these findings to the morally-rich context of promise keeping.[2] Will promise keeping prove equally vulnerable to moral wiggle room?

Closely related to our work are studies in the promise keeping literature that directly varied whether promise keeping is visible to other participants or the experimenter, rather than providing participants with responsibility-diffusing excuses. Whereas early studies fail to find a significant effect (Deck et al., 2013; Schütte and Thoma, 2014), or find mixed evidence (Cadsby et al., 2015), more recent contributions show that the observability of promise keeping does have an impact (Kingsuwankul et al., 2023; Lang and Schudy, 2023). These divergent findings warrant further research into the role of social-image concerns in promise keeping, which our study provides. A further and more novel contribution of our study lies in its test of self-image concerns in promise keeping. Such concerns have featured much less prominently in the literature, presumably due to their reliance on subtle cognitive processes—such as self-deception—that are more difficult to target and manipulate than social-image concerns.

Our design is based on a modified version of the "plausible deniability" mechanism introduced by Dana et al. (2007). Participants first exchange promises stating their intent to select a generous (as opposed to selfish) allocation in a prospective double-sided dictator game (as in Vanberg, 2008). Whether participants are able to act on their promises is tied to their performance in a simple effort task, which, upon successful completion, grants them the required decision right in the dictator game. The twist is that participants can be cut off from the task before completion, which results in their decision right in the dictator game being delegated to the computer. In this case, the computer implements the generous or selfish allocation with equal probability on their behalf. To test for social-image concerns, we manipulate the plausibility of using the cut-off mechanism as an excuse for selfish allocations by varying whether cut-offs can occur early or late. To test for self-image concerns, we analyze participants' performance in the effort task to identify motivated delays aimed at delegating the likely implementation of the selfish allocation to the computer.

When the cause of a broken promise is undeniable, we observe high rates of promise keeping. However, when the cut-off mechanism can be exploited as a plausible excuse to preserve one's social image, a significant number of participants choose to break their promise. Yet, cooperation in the experiment remains higher compared to treatments without a communication stage, and we find no evidence of participants engaging in motivated delays to evade their promise-induced commitments. Thus, while some subjects are sensitive to whether they are seen as a promise breaker, others exhibit stable preferences for promise keeping that are not easily eroded by moral wiggle room.

The remainder of this paper is structured as follows. Section 2 presents the experimental design and elaborates on the hypotheses and procedures of the experiment. Section 3 reports the results. Section 4 provides a discussion and situates our paper within the broader literature. Section 5 concludes.

2 Materials and methods

2.1 Experimental design

Our experimental design builds on the "plausible deniability" framework introduced by Dana et al. (2007), with a few key modifications. First, we embed the cut-off mechanism in a real-effort task to test whether participants intentionally delay task completion in order to trigger cut-offs that delegate the implementation of the selfish outcome to the computer—we refer to these as our Plausible Deniability treatments. Second, we introduce conditions similar to those in Andreoni and Bernheim (2009), in which the cut-off mechanism (or, as they would say, *nature*) cannot be blamed for selfish outcomes—we refer to these as our No Deniability treatments. We adopted the methods contained in these two studies as they are frequently used in the literature to vary self- and social- image concerns, respectively. Third, to test the relevance of these concerns in the specific domain of promise keeping, we add treatments that vary whether participants can engage in pre-play communication, adopting Vanberg (2008)'s sequential promise-making structure. Comparing treatments with and without communication allows us to separate image concerns tied to the act of promise-making from those related to distributional considerations—such as appearing selfish, greedy, or unfair. Table 1 summarizes our 2 × 2 factorial design. The sequence of stages in the experiment is illustrated in Figure 1. We now turn to a more detailed discussion of our design.

1 Several studies focus on the relative importance of the first (commitment-based) and second (expectations-based) explanation of promise keeping (Ismayilov and Potters, 2016; Ederer and Stremitzer, 2017; Mischkowski et al., 2019; Di Bartolomeo et al., 2019; Schwartz et al., 2019; Di Bartolomeo et al., 2023a,b). Image concerns reflect a preference for viewing oneself (or being viewed by others) in a positive light (Bodner and Prelec, 2003; Bénabou and Tirole, 2004, 2006, 2011; Andreoni and Bernheim, 2009; Linardi and McConnell, 2011; Grossman and Van der Weele, 2017) and have featured less prominently in the literature on promise keeping.

2 Image concerns have featured more prominently in a related literature on lying aversion that mainly concerns individual decision making (e.g., Mazar et al., 2008; Gneezy et al., 2018; Bašić and Quercia, 2022; Bicchieri et al., 2023). In contrast, we consider strategic interactions and explicit promises which arguably impose stricter moral constraints than norms of simple truth-telling.

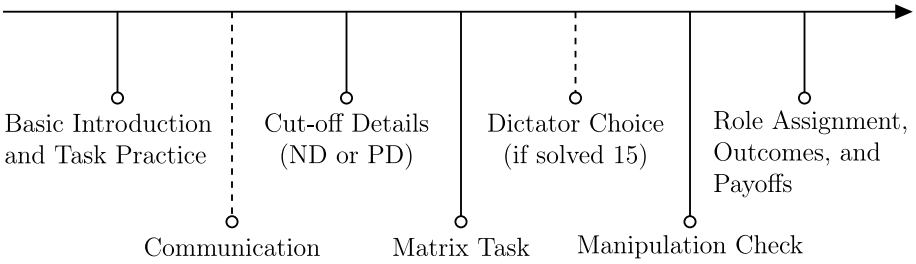

FIGURE 1
Sequence of stages in the experiment. Dotted lines indicate the structural changes of our treatments to the timeline of the experiment. The ability to make promises was varied by adding/removing a pre-play communication stage. All treatments had in common that a decision right in the dictator game was only obtained if a preceding matrix task was solved successfully. Plausible deniability of actions in the dictator game was introduced and varied via a cut-off mechanism that could interrupt subjects at an early (plausible deniability) or late (no deniability) stage during the task. A cut-off implied that the dictator allocation (either generous of selfish) was selected by the computer with equal chance.

TABLE 1 Factorial treatment design.

Factor A: communication	Factor B: deniability	
	No deniability	Plausible deniability
No communication	NC_ND	NC_PD
Communication	C_ND	C_PD

Treatments vary along two dimensions: whether communication (promise exchange) is possible, and whether actions are deniable. This design allows to separate distributional image concerns (appearing selfish, greedy or unfair) from image concerns related to promise keeping *per se*.

Subjects are randomly paired into groups of two. Role assignment takes place at the end of the experiment; that is, all subjects simultaneously play as *A* players (potential dictators) knowing that outcomes in this role will count for only half of them, while the other half will at the end assume the role of player *B* (recipient).[3] As a result, all subjects receive identical instructions from the outset.

All treatments have in common that a decision right in the dictator game stage is only obtained if a preceding matrix task is solved successfully. In case of *success*, the subject obtains the decision right and selects how to allocate money between him- or herself and their counterpart by choosing one of two possible allocations: *A*=(£10,£0) or *B*=(£6,£6). Conversely, in case of *no success*, the subject does not obtain the decision right and is forced to let the computer randomly implement either of the two allocations with equal probability on their behalf.

The matrix task, borrowed from Abeler et al. (2011), consists of subjects counting ones (1s) in a series of 5x5 matrices comprised of randomly ordered zeros and ones. Importantly, we modified the task to feature a cut-off mechanism which (in some of our treatments) can serve as a plausible excuse for the implementation of the selfish allocation *A* (£10, £0).[4] Successful completion requires a subject to solve a target number of 15 matrices within the allotted time— i.e. before being cut off by the computer.

We employ different variants of the cut-off mechanism in our experiment. In our No Deniability (ND) treatments (Table 1, second column), subjects are given 300 s (5 min) to work on the task until a cut-off occurs. The time allotted in these treatments is intentionally generous, based on the results of an informal and unincentivized pretest where subjects needed on average 104 s to solve 15 matrices and no subject took longer than 138 s. Our aim was to erase the plausibility of using the cut-off mechanism as an excuse for selfish allocations while keeping the experimental protocol as close as possible to the treatments we describe next.

In our Plausible Deniability (PD) treatments (Table 1, third column), instead of telling subjects that the cut-off would occur after exactly 300 s, we tell them that the cut-off can occur at any randomly determined second within the 300 s interval.[5] The PD treatments offer room for two distinct dimensions of deniability:

Deniability toward the counterpart. Subjects can exploit the fact that their counterpart cannot verify whether an allocation was the result of a subject's own choice or was implemented by the computer. Our plausible deniability treatments therefore reduce the social-image cost typically associated with selfish behavior under full transparency.

Deniability toward the self. Subjects who feel compelled to choose the other-regarding allocation because they do not want

3 This design choice is motivated by Vanberg (2008)'s double-dictator game. In the instructions, we refer to "you" and "your counterpart" instead of "dictator" and "recipient." Instructions can be found in Appendix B.

4 In Dana et al. (2007), 24% of the subjects allowed themselves to be cut off by the computer, thereby preferring a mixture of two outcomes over each one separately. This observation is "inconsistent with a theory of rational choice with utilities defined only over outcomes" (p. 74). For subjects who feel compelled to choose the other-regarding option in order not to threaten their self-image, however, being cut off can be desirable. In half of the cases, the outcome selected would align with what the dictator felt compelled to choose anyway. In another half, the opportunistic outcome would be implemented, allowing the subject to uphold the illusion of not being responsible for its selection.

5 If a cut-off occurred, a subject worked on a follow-up task for the remainder of the 300 s. The task was not incentivized and consisted of adding up numbers on the screen. The purpose of this task was to maintain a constant background sound of mouse clicks, thereby avoiding that subjects could infer from the lack of this sound information about the timing of cut-offs of their peers.

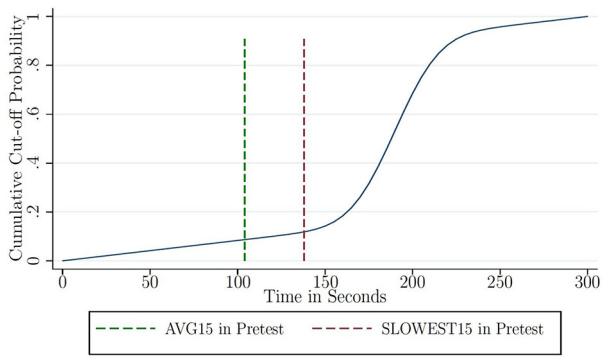

FIGURE 2
Cumulative distribution function of the cut-off timer in the plausible deniability treatments and the control. The distribution was calibrated to favor late cut-offs, ensuring limited data loss and minimizing confounds from selection effects. AVG15 and SLOWEST15 refer to the success times (solving 15 matrices) of the average as well as the slowest subject in an informal pretest.

to think badly of themselves may prefer to be cut off by the computer. A cut-off results in a 50% chance of obtaining the selfish allocation, while allowing them to maintain the illusion of not being responsible for its implementation.

We expected that self-deceivers would work on the task half-heartedly, waste time or commit more errors—all of which would delay task completion.[6] To identify whether subjects in our PD treatments indeed procrastinated, we conducted an additional control treatment. This treatment was designed to mirror the NC_PD treatment as closely as possible. The only difference was the absence of a counterpart. In this treatment, successful completion of the matrix task allowed dictators to choose their own payoff only (£10 or £6). Since incentives to procrastinate were removed in this control, we expected to obtain an unbiased distribution of task performance to serve as a comparison benchmark for performance in our main treatments. Instructions for the control treatment can be found in Appendix B.2.

No information was disclosed to subjects regarding the underlying distribution that generated the cut-offs in our PD treatments (and the control). While it is technically true that a cut-off could occur anywhere within the specified time interval, we used a distribution that favored later cut-offs. To achieve this, we combined a discretized normal distribution with a uniform distribution, such that cut-offs were drawn from the function: $f(x) = \mathcal{N}(190, 20) + \mathcal{U}\{1, 300\}$.[7] Figure 2 displays the corresponding cumulative distribution function, illustrating the probability of being cut off in the matrix task as a function of time. Dotted lines mark the times that the average and slowest subjects took to successfully complete the matrix task in the informal pretest. These times were used as benchmarks for our calibration. Our cut-off distribution was calibrated to achieve the following two objectives:

Data efficiency. Early cut-offs are associated with data loss because neither is the time data of subjects rich enough to identify procrastination, nor do we obtain choice data in the subsequent dictator game. To minimize data loss, our cut-off distribution is shifted to the right. Recall that in the pretest, subjects needed an average of 104 s to succeed in the matrix task. But even by the 150-s mark, the cumulative probability of being cut off from the task was only 12%, after which it increased more rapidly.

Internal validity. Some of our hypotheses tested in Section 3.2 compare aggregate behavior in the dictator game stage between our ND and PD treatments. For these tests to be reliable, we have to rule out the possibility that our cut-off mechanism altered the composition of our PD samples compared to the ND samples. This would be the case, for example, if cut-off subjects were disproportionally selfish or other-regarding. The rightward shift of our cut-off distribution was specifically motivated to handle this concern. Since, in most cases, a cut-off would not occur until very late, we made it very difficult for subjects to sustain a self-deceptive strategy. A cut-off could only be enforced through excessive procrastination which we expected to be incompatible with maintaining the perception of not being responsible. Consequently, we expected most subjects to complete the task, with only few being cut off. In Section 3.2 we confirm that this was indeed the case in our experiment.

The second dimension of our factorial treatment design varied whether subjects could exchange promises with their counterpart before starting the matrix task. We adapted the sequential promise-making structure used by Vanberg (2008) to successfully induce promise exchange, but opted for pre-formulated instead of free-form messages, in order to avoid the subjective process of message coding. Within each group, one subject was randomly selected to choose the first message:

Message 1: "I promise to do my best to implement Option B, if you promise to do the same."
Message 2: "I don't want to commit myself to anything."

The second subject could then reply by choosing between:

Message 1: "I promise to do my best to implement Option B."
Message 2: "I don't want to commit myself to anything."

Payoffs were calibrated to provide both an equality-based and a total-earnings-maximizing argument in favor of option B(£6, £6) over the selfish option A(£10, £0). We expected that subjects would use the communication stage to exchange promises as a means of coordinating on the former allocation.

6 Previous studies which used a cut-off mechanism required self-deceivers to be passive and to wait for the computer to intervene. We chose to embed our cut-off mechanism into a real effort task instead of the dictator game itself to reduce potential demand effects and to mimic a richer (and, in our view, more realistic) environment that would allow subjects to conceal their intentions in an inconspicuous way—by masking their true ability during an active task.

7 We refrained from shifting the cut-off distribution entirely to the right and included a uniformly distributed element to avoid deception and prevent subjects from deducing the underlying distribution of cut-offs ex-post—e.g. through communication with other participants after the experiment.

The experiment was designed such that the deniability manipulations occurred only after the communication stage had concluded. Thus, at the time of exchanging messages, subjects did not know whether they would be assigned to the No Deniability or Plausible Deniability condition. It was only after the communication stage that they learned which condition applied to them.[8] This design allowed us to vary, by treatment, whether deniability was possible, without influencing the content of exchanged messages.

We opted for a one-shot version of the game, anticipating that repeated play would likely diminish or eliminate the scope for self-deception. To aid subjects' understanding of the rules and processes of the experiment, we initiated a practice phase in which they were guided through the stages of the experiment, supplemented by detailed explanations. During this practice phase, subjects also worked on scaled-down versions of the matrix task with computer-simulated counterparts. A late cut-off round (60 s) demonstrated how the matrix task worked, followed by an early cut-off round (12 s), which familiarized subjects with the cut-off mechanism and its consequences.[9] The practice phase concluded with a quiz to ensure that subjects understood the instructions and procedures of the experiment. To check whether our cut-off mechanism successfully diffused the perceived responsibility for outcomes, we elicited on a 5-point Likert scale participants' first- and second-order beliefs about the likelihood that their counterpart solved the task on time (1 = very unlikely, 5 = very likely). This manipulation check took place after the conclusion of the dictator game, but before roles, outcomes and payoffs were determined.

2.2 Hypotheses

We start this section with general hypotheses about the content and effects of exchanged messages, before turning our attention to image concerns in particular.

Since the focus of our paper is on promise keeping, it was our aim to facilitate high rates of promise exchange in the experiment. To this end, we adapted Vanberg (2008)'s sequential promise-making structure, which in his study led 79% of messages to contain a promise. We also expected promise-induced cooperation on the generous allocation to be appealing to many subjects, due to its equal and total-earnings-maximizing payoff properties. Moreover, our restrictive communication protocol with pre-formulated messages made promise exchange suggestive and eliminated the ambiguities surrounding the classification of messages that are often observed under free-form communication protocols.

Hypothesis 1: *Subjects will use the communication stage to exchange promises.*

It is a well-documented finding in the literature that promises are often kept, even in one-shot interactions and in the absence of observers or punishment threats (e.g. Ellingsen and Johannesson, 2004; Vanberg, 2008; Di Bartolomeo et al., 2019). If people keep their promises because they intrinsically value doing so, we would expect promise keeping to persist even when excuses are available that allow them to diffuse responsibility. Consequently, we would expect promises to increase generosity under both our No Deniability and Plausible Deniability conditions.

Hypothesis 2: *Generosity is higher in treatments featuring promise exchange.*

2.2.1 Social-image concerns

Social-image concerns suggest that individuals gain (or lose) utility from being perceived in a positive (or negative) light by others (e.g. Andreoni and Bernheim, 2009; Ariely et al., 2009; Bursztyn and Jensen, 2017; Friedrichsen and Engelmann, 2018). In our No Deniability conditions, broken promises are easily attributed to subjects' opportunistic intentions, thereby threatening their social image. In our Plausible Deniability treatments, on the other hand, subjects can exploit possible cut-offs as an excuse to break promises, which mitigates a cost to their reputation. Under the assumption that some promise keeping is driven by social-image concerns, we would expect communication and the exchange of promises to be less effective under PD than ND.

Hypothesis 3: *Promises induce generosity less effectively under PD than ND.*

2.2.2 Self-image concerns

Self-image concerns refer to how individuals internally evaluate the decisions they make (Bem, 1972; Bodner and Prelec, 2003), and the relevance of these concerns has been documented across numerous studies (e.g. Mazar et al., 2008; Johansson-Stenman and Svedsäter, 2012; Falk, 2021). Self-image concerns may also apply to the act of promise keeping, and we therefore ask whether the strength of promises is compromised when subjects are also able to *self*-deceive about the cause of a broken promise. In our experiment, a subject who feels compelled to honor their promise in order to protect their self-image may choose to procrastinate on the matrix task in the hope of being cut off by the computer. A cut-off results in a fair chance of obtaining the selfish outcome while allowing the subject to maintain the perception of not having acted against their promise. To identify procrastination, we compare matrix task performance in treatment C_PD to performance in NC_PD and CONTROL. Recall that no counterparts were involved in the control treatment, and that successful completion of the matrix task allowed a subject to choose their own payoff only. The idea behind this control treatment was that image-related incentives for procrastination would be removed, thereby providing an unbiased benchmark of participants' ability in the task against which we compare performance in our PD treatments

8 In the instructions, we only provide minimal information about the cut-off mechanism. Subjects are told that additional details would follow later in the experiment. After the communication stage concluded, treatment-specific details regarding the cut-off mechanism were read aloud by the experimenter. Scripts can be found in Appendix B.3.

9 To hint at the possibility that a cut-off could be desirable, we programmed the computer to select the selfish outcome in the early cut-off round. Thus, every subject experienced at least once that a cut-off could result in their favor. Appendix C provides screenshots from the practice phase.

TABLE 2 Overview of message profiles by treatment.

$Message_{F\text{-}Mover}/Message_{S\text{-}Mover}$	By treatment			Pooled
	C_ND	C_PD	*p*-value	C_ND + C_PD
Promise intent/promise	17/24 (70.8%)	25/32 (78.1%)	0.551	42/56 (75%)
Promise intent/no commitment	3/24 (12.5%)	1/32 (3.1%)	0.303	4/56 (7.1%)
No commitment/promise	1/24 (4.2%)	1/32 (3.1%)	1.000	2/56 (3.6%)
No commitment/no commitment	3/24 (12.5%)	5/32 (15.6%)	1.000	8/56 (14.3%)

Message exchange patterns during the communication stage, by treatment and pooled. The majority of pairs used the communication stage to exchange promises, confirming its intended function in line with hypothesis 1. Reported *p*-values are based on two-sided Fisher's exact tests.

(where we expected incentives for procrastination to exist). Since promises induce commitments and moral pressure, we expected motivated delays to be more pronounced in treatment C_PD compared to NC_PD and CONTROL.

Hypothesis 4: *Task performance declines in C_PD relative to NC_PD and CONTROL.*

2.3 Procedures

The study was approved by the ECO ethics committee of the University of East Anglia. The experiment was programmed in z-Tree (Fischbacher, 2007) and conducted in the Laboratory for Economic and Decision Research (LEDR). A total of 254 participants, recruited from the local student population, took part in the experiment. We ran 16 sessions, each lasting between 35 and 45 min, depending on the treatment. We conducted more PD sessions to compensate for the small data loss expected to occur due to early cut-offs. The number of sessions per treatment was as follows: 3 × NC_ND, 3 × C_ND, 4 × NC_PD, 4 × C_PD, 2 × CONTROL. 16 subjects participated in each session, except for one NC_PD session where only 14 subjects showed up. Average earnings were £10, with a minimum of £4 and a maximum of £16 (including a £3 participation fee). Further details on the procedures are provided in Appendix B.5.

3 Results

Section 3.1 examines the content of communication in our experiment. Section 3.2 analyzes the effects of communication, focusing on the role of social-image concerns in Section 3.2.1 and self-image concerns in Section 3.2.2. Despite robust empirical evidence for directional effects in prior work, we take a conservative stance by using two-sided statistical tests throughout.

3.1 Communication contents

Table 2 summarizes the observed message profiles (pairs of messages) broken down by treatment condition. Recall that by design, our deniability manipulations occured only after the communication stage concluded. Up to that point, the experimental protocol, including the instructions, was identical across treatments. We would therefore expect no significant differences in the content of exchanged messages. This is confirmed by our data, which is why we henceforth refer to the pooled data provided in the last column of Table 2.

By looking at the first two rows of Table 2, we can see that 46 out of 56 first-movers (82.1%) sent the cooperative message 1, stating a promise intent. Among the 46 second-movers who received a promise intent, 42 (91.3%) reciprocated with a promise thereby establishing mutual promise exchange. Unsurprisingly, among the few cases (10 out of 56) where first-movers refrained from proposing a mutual exchange of promises by stating that they do not want to commit themselves, the majority of second-movers (eight out of 10) also chose not to commit. Two subjects decided to commit despite not having received a willingness to commit from their counterpart. In line with hypothesis 1, we can state the following result:

Result 1. *The majority of pairs (75%) used communication to exchange promises.*

3.2 Communication effects

Having established that subjects used the communication stage to exchange promises, we can investigate whether and how promise exchange increased generosity in our communication treatments. Table 3 reports the frequency of cut-offs and choices in the dictator game stage, broken down by treatment and, where applicable, communication history. We pool data from first- and second-movers in the sequential message exchange, as we found no statistically significant differences in the behavior of these two subgroups (see Appendix A.2).

Our analysis is based on subjects who successfully completed the matrix task and for whom choice data in the dictator game is available. Losing data on subjects who were cut off before completing the task may raise internal validity concerns. As discussed earlier, we designed our experiment to minimize these concerns. As expected, the proportions of subjects who were

TABLE 3 Allocation and cut-offs by treatment.

Treatment	*n*	Cut off *n*(%)	Generous *n*(%)	Selfish *n*(%)
Communication	112	8 (7.1%)	49 (47.1%)	55 (52.9%)
C_ND	48	2 (4.2%)	27 (58.7%)	19 (41.3%)
C_ND_PromiseEx.	34	1 (2.9%)	25 (75.8%)	8 (24.2%)
C_ND_NoPromiseEx.	14	1 (7.1%)	2 (15.4%)	11 (84.6%)
C_PD	64	6 (9.4%)	22 (37.9%)	36 (62.1%)
C_PD_PromiseEx.	50	5 (10.0%)	22 (48.9%)	23 (51.1%)
C_PD_NoPromiseEx.	14	1 (7.1%)	0 (0.0%)	13 (100%)
No communication	110	9 (8.2%)	20 (19.8%)	81 (80.2%)
NC_ND	48	0 (0.0%)	10 (20.8%)	38 (79.2%)
NC_PD	62	9 (14.5%)	10 (18.9%)	43 (81.1%)
CONTROL	32	4 (12.5%)	1 (3.6%)	27 (96.4%)

Observed choices by subjects in the dictator game stage and frequency of cut-offs across treatments. The exchange of promises increases generosity (in line with hypothesis 2), but plausible deniability reduces the effectiveness of promises (in line with hypothesis 3). Cut-off rates were small across treatments, supporting the internal validity of our comparisons.

cut off in our Plausible Deniability conditions were small: 6/64 (9.4%) in treatment C_PD, 9/62 (14.5%) in treatment NC_PD, and 4/32 (12.5%) in treatment CONTROL. Moreover, if selection effects were present—e.g., if procrastinators successfully managed to enforce a cut-off—we would expect the proportion of cut-offs to be higher in treatments C_PD and NC_PD (where incentives for procrastination were present) compared to treatment CONTROL (where such incentives were removed). However, this was not the case, according to pairwise Fisher's exact tests (p = 0.727 and p = 1.000, respectively). Appendix A.1 provides details on the cut-off times and matrix task progress of subjects who were cut off before completing the task. It is noteworthy that a considerable proportion of these subjects (11/21 or 52.4%) did not manage to solve a single matrix in the practice stage, suggesting that our cut-off mechanism effectively filtered out subjects who lacked a sufficient understanding of the task.

Figure 3 summarizes our main findings by displaying the proportions of subjects choosing the generous allocation for each treatment separately. Our communication protocol proved effective in increasing generous allocations under both No Deniability (20.8 vs. 58.7%; $p < 0.01$, Fisher's exact test) and Plausible Deniability (18.9 vs. 37.9%; p = 0.036, Fisher's exact test). Pooling across treatments, communication increased the proportion of generous choices from 19.8 to 47.1% ($p < 0.01$, Fisher's exact test). This finding is consistent with hypothesis 2 and replicates previous research on the effectiveness of communication and promise exchange.

Result 2. *Generosity is higher in treatments featuring promise exchange.*

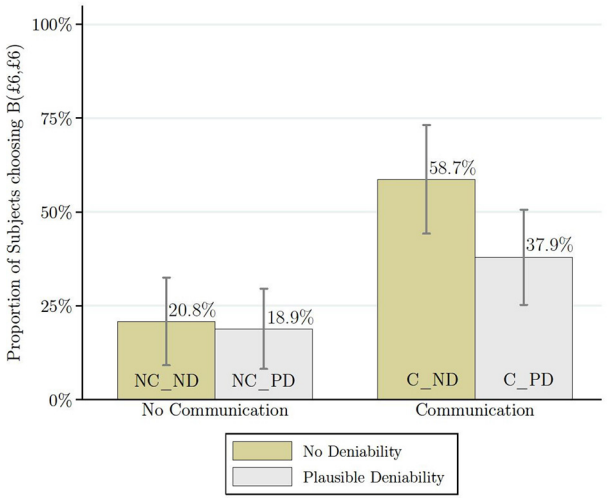

FIGURE 3
Proportion of subjects choosing the generous allocation (£6,£6) across treatments. The exchange of promises increases generosity, but plausible deniability reduces the effectiveness of promises, providing support for hypotheses 2 and 3 and social-image concerns in promise keeping.

3.2.1 Social-image concerns

A closer inspection of the communication bars in Figure 3 reveals that some of the gains from communication are lost when moving from ND to PD (58.7 vs. 37.9%; p = 0.048, Fisher's exact test). This reduction can be traced back to an increased willingness among subjects to break promises, which increases from 24.2% in treatment C_ND to 51.1% in treatment C_PD (p = 0.020, Fisher's exact test), supporting the idea that promise keeping is partly driven by social-image concerns.[10]

10 The successful diffusion of responsibility is also reflected in subjects' second-order beliefs about matrix task success. The mean response decreases from 4.9 in the ND treatments to 4.1 in the PD treatments (where 5 = "very likely" and 4 = "somewhat likely"; $p < 0.01$, Mann-Whitney U-test).

TABLE 4 Success times and accuracy in the matrix task across treatments.

Treatment	*n*	Cut off *n* (%)	Time15 mean/median	Incorrect15 mean/median
C_PD	64	6(9.4%)	103 s/100 s	1.22/1
NC_PD	62	9(14.5%)	102 s/102 s	1.49/1
CONTROL	32	4(12.5%)	111 s/104 s	1.29/1

Summary statistics of task performance (time to complete 15 matrices and number of mistakes) across treatments. We observe no evidence of procrastination or inferior performance in treatment C_PD, suggesting an absence of motivated delays.

Result 3. *Promises induce generosity less effectively under PD than ND.*

Image concerns in our experiment may arise not only from the *process* of promise making. Previous research has shown that people also dislike being responsible for *outcomes* that make them appear selfish or unfair. Our communication-free treatments help demonstrate that purely outcome-based image concerns did not appear to play a major role in our experiment. This is evident from the left bars in Figure 3, which show low levels of generosity even when social image is at stake, due to the absence of deniability. We believe there is a plausible explanation for this discrepancy with previous findings. Unlike in previous research, subjects in our setup have to earn their right to decide as dictators by exerting effort in a prior task. This may have created entitlement effects (as in Cherry et al., 2002), which could justify the choice of the selfish outcome (possibly coupled with the belief that both players had a fair chance to act as dictators). By contrast, communication generates explicit commitments through promises, introducing a distinct source of moral obligation. In Appendix A.3, we confirm the robustness of these results using regression analyses.

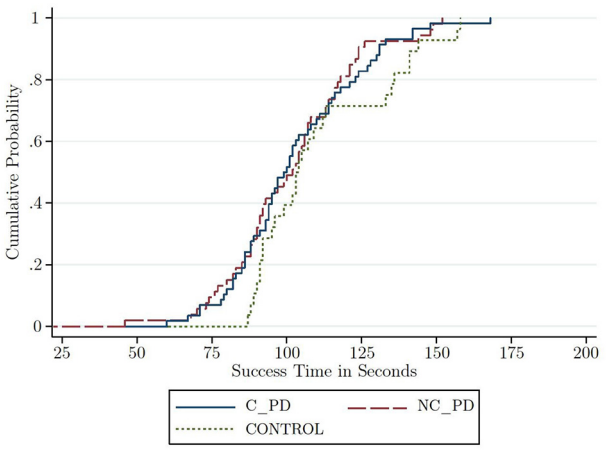

FIGURE 4
Cumulative distribution functions of durations to complete 15 matrices across treatments. No significant differences are observed across treatments, confirming the absence of systematic delays in treatment C_PD, thereby contradicting hypothesis 4.

3.2.2 Self-image concerns

It is also interesting to observe that generosity under PD remains considerably higher in treatment C_PD compared to NC_PD, where no communication was possible. This is due to the substantial proportion of subjects (22/45 or 48.9%) who honored their promise despite having the option to hide behind the cut-off mechanism. Perhaps these subjects had an intrinsic desire to fulfill their promises. Alternatively, they might have done so reluctantly to maintain a positive self-image. To test for self-image concerns, we examined subjects' performance in the matrix task to identify signs of procrastination and motivated delays—for example, in the form of increased errors.

To obtain a benchmark for subjects' abilities in the matrix task—against which to compare performance in our plausible deniability treatments—we conducted our control treatment that removed incentives for procrastination. The following analysis is based on a comparison of matrix task performance observed across treatments C_PD, NC_PD, and CONTROL.

Table 4 reports summary statistics on the speed and accuracy with which subjects solved the target of 15 matrices.[11] Figure 4 presents the associated cumulative distribution functions (CDFs) of success times across treatments. If subjects procrastinated in our main treatments, we would expect the respective CDFs to lie further to the right compared to our control treatment, where incentives for procrastination were removed. Contrary to this expectation, we observe the opposite. Subjects in our main treatments appear to have performed even better than those in the control treatment—a pattern especially pronounced among high-performing subjects. However, according to pairwise Kolmogorov-Smirnov tests, the distributions for treatments C_PD and NC_PD do not differ significantly from CONTROL ($p = 0.221$ and 0.305, respectively).

We also examined within-subject variation in performance on the matrix task. It is possible that procrastination could take the form of subjects slowing down on the task as they approach the target of 15 matrices. Figure 5 displays the average time spent on each of the 15 tasks, broken down by treatment. Once again, visual inspection suggests that subjects in our main treatments performed better than subjects in the control treatment, particularly toward the end of the task.

To quantify the patterns observed in Figure 5, we ran a random-effects panel model estimation. Results are presented in

11 We continue to condition our analysis on the sample of subjects who were not cut off. Recall our previous discussion and Appendix A.1 for a justification of this approach. One advantage of doing so is that our cut-off mechanism simultaneously filtered out subjects who lacked a sufficient understanding of the task. Including these subjects in the analysis would have made it difficult to distinguish motivated procrastination from delays due to confusion.

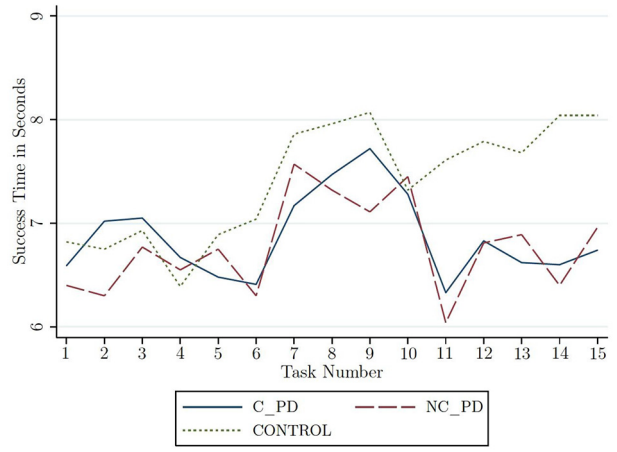

FIGURE 5
Average time spent on each of the 15 matrices across treatments. Performance in treatment C_PD remained stable across time and comparable to treatments C_ND and CONTROL. These patterns provide further evidence against self-image driven procrastination.

Table 5. Our dependent variable is the natural logarithm of the time (in seconds) a subject spent solving a given task. *TREAT* is a dummy variable distinguishing our treatment conditions, with CONTROL serving as the reference category. *TASK_N* denotes the task number, allowing us to measure changes in performance over time. We also include an interaction term between *TREAT* and *TASK_N* to allow performance changes to be treatment-specific. The coefficient for *TASK_N* is positive and significant, suggesting that subjects in the control treatment exhibit performance reductions over time—an effect that could be due to boredom or fatigue. In contrast, no such time trend is observed in treatments NC_PD and C_PD. This is reflected in the negative and significant interaction terms, which fully compensate for the negative time trend observed in our control treatment. Overall, performance in the matrix task appears to be inferior in the control treatment, with no significant difference between treatments C_PD and NC_PD. This result contradicts hypothesis 4 and allows us to conclude:

Result 4. *There is no evidence of inferior performance in treatment C_PD relative to treatments NC_PD and CONTROL.*

4 Discussion

Trust is a fundamental element in many economic and social interactions where formal enforcement is weak or absent. Citizens rely on politicians to fulfill campaign promises, investors depend on financial advisors to act in their best interest, and users of peer-to-peer marketplaces trust that goods will be delivered after payment. In such settings, communication—particularly in the form of promises—often plays a crucial role in promoting cooperation.

We contribute to the literature on promise keeping by providing a more nuanced view on the power of promises. When individuals are equipped with plausible excuses that allow them to disguise their choices, promises remain effective, but to a lesser degree. This reduction can be attributed to the role of social-image concerns: the availability of excuses appears to lower the reputational cost of breaking a promise. This interpretation aligns well with recent studies exploring the influence of image concerns in promise keeping. For instance, Kingsuwankul et al. (2023) examine the impact of honesty oaths in a financial market context, motivated by regulatory practices in the Dutch banking system. They find that public oaths dramatically reduce dishonest behavior. Their comparison of public and private oaths mirrors our contrast between the C_ND and C_PD conditions: when individuals are held accountable, promise keeping is more robust. A similar pattern emerges in Lang and Schudy (2023)'s investigation of political campaign promises. In a dynamic environment with promise competition, they find that transparency reduces promise breaking, but also leads to less generous promises. Another common finding across these studies and ours is that many individuals keep their promises even when promise breaking is unobserved or deniable. This behavior may reflect a genuine preference for honoring moral commitments. Alternatively, it could arise from an impure desire to maintain a favorable self-image. A key contribution of our study lies in its test of self-image concerns by allowing participants to self-deceive about the cause of a broken promise.[12] Our null result could be interpreted as evidence supporting an intrinsic preference for promise keeping. As such, our findings highlight a dual insight: while the availability of excuses weakens the social enforcement of promises, a resilient core of individuals honor their commitments even in the absence of reputational consequences. This suggests that internal moral standards can sustain cooperative behavior where external enforcement fails. Promises, then, retain their power not only through social accountability, but also through the self-regulatory force of personal integrity. Future research could aim to identify screening mechanisms that allow to distinguish between the two types of individuals: those who keep their promises reluctantly, and those who genuinely desire to honor their commitments.

Our study also contributes to the literature on image concerns—specifically, the relative importance of social vs. self-image. As pointed out for example by Bursztyn and Jensen (2017, p. 144), it is important to be able to differentiate the two concerns, as they differ in both their underlying mechanisms and their policy implications. However, most of the existing literature has focused on one concern in isolation, making comparisons across varying experimental setups difficult. There are a few exceptions where both concerns are examined within the same

12 Lang and Schudy (2023) include additional treatments to isolate different mechanisms at play in their dynamic game of promise competition. One treatment targets self-image concerns by introducing information about "economic circumstances", thereby making it harder for participants to downplay the likely consequences of a broken promise. However, such information may trigger image costs unrelated to promises per se, such as being perceived as a greedy person (as in Dana et al., 2007). An advantage of our communication-free treatments is that they isolate promise-specific motivations by ruling out image concerns tied to distributional fairness or greed.

TABLE 5 Random effects panel model estimations.

Dep. variable *LN_TIME*	Coef.	Robust SE	*Z*	*p*-value
TREAT				
NC_PD	−0.015	0.050	−0.30	0.762
C_PD	0.023	0.050	0.47	0.641
TASK_N	0.012	0.004	3.15	0.002
TREAT × *TASK_N*				
NC_PD	−0.010	0.005	−2.23	0.026
C_PD	−0.012	0.004	−2.82	0.005
_CONS	1.847	0.041	45.31	0.000
		Prob > chi2		0.013
		R-Squared		0.015
		Number of groups		139
		Number of observations		2,085

Random effects panel regressions of time spent per matrix, testing for performance differences across treatments. Standard errors are clustered at the subject level. The results confirm that performance did not decline in treatment C_PD, contradicting motivated delays (hypothesis 4).

design. For example, Grossman (2015) uses a probabilistic dictator game where the dictator's choice is only implemented with a certain probability, and varies whether the choice, the outcome, or both the outcome and the implementation probability are revealed to recipients. He finds evidence of social-signaling, but not of self-signaling. Relatedly, Andreoni and Sanchez (2020) compare subjects' stated and true beliefs in a trust game and find a discrepancy that aligns more with social-image than self-image concerns. In our design, both types of image concerns were similarly varied within a unified experimental framework. Our observation that we find evidence for social-image but not for self-image concerns is consistent with previous findings, and may suggest that social-image is the stronger of the two concerns.

Finally, our study contributes to the literature on moral wiggle room and excuse-seeking behavior. This literature originated from very simple dictator game studies and has since documented that individuals not only shift responsibility for outcomes by blaming external circumstances (Dana et al., 2007) or other agents (Hamman et al., 2010), but engage in a wide range of subtle techniques to preserve their image, such as the crafting of excuses based on motivated risk preferences (Haisley and Weber, 2010; Exley, 2015), motivated beliefs (Di Tella et al., 2015; Andreoni and Sanchez, 2020; Bicchieri et al., 2023), motivated memory (Saucet and Villeval, 2019; Amelio and Zimmermann, 2023), or motivated errors (Exley and Kessler, 2024). We add to this list the study of motivated delays in a real-effort task, providing a novel measure to identify self-deception in a subtle and unobtrusive way. Beyond investigating the different forms of excuses individuals rely on, there has also been a growing interest to explore the role of excuses in morally-richer contexts. For instance, van der Weele et al. (2014) use a cut-off mechanism to investigate the robustness of reciprocal behavior in a trust and a moonlighting game. They report a null result and conclude that reciprocal preferences resist moral wiggle room. In contrast, Malmendier et al. (2014) and Regner (2018) find that excuses can undermine reciprocity. Similarly, Kriss et al. (2016) show that third parties misreport the outcome of a private die roll to avoid the cost of punishing norm violators—another example of excuse-driven moral evasion. We add to this growing body of work by testing the role of excuses in yet another important domain of everyday social interaction, namely promise exchange. In light of our null result on motivated delays, an interesting direction for future research would be to challenge or refine our findings, for example, by applying one of the many alternative methods discussed in the literature to vary self-image concerns.

We would also like to acknowledge some limitations of our approach. To generate sufficient data on the exchange of promises, we employed a protocol with pre-formulated messages that made promise exchange suggestive. Previous research has found that promises elicited under such restrictive protocols are less powerful than voluntary, free-form promises (Charness and Dufwenberg, 2010; Chen and Zhang, 2021). While we did not observe a lack of effectiveness of "bare" promises in our setup, this may imply that our results represent a lower bound on the effectiveness of voluntary promises. At the same time, we cannot rule out the possibility that our restrictive communication protocol may have invited experimenter demand effects (Zizzo, 2010). This concern, however, is less applicable to our deniability manipulations, which involved only subtle procedural changes. One of our design choices was to reveal the cut-off details only after the communication stage concluded. This had the advantage of keeping the content of communication constant across treatments. An interesting extension for future research would be to inform participants about the availability of excuses prior to communication and examine whether this systematically affects the content of communication—and the credibility of promises—in a setup with voluntary, free-form communication.

Our identification of self-image concerns relied on participants' incentives to engage in procrastination and self-deception to avoid

the damage inflicted on their self-image when breaking a promise. However, it is possible that a conflicting image concern may have influenced procrastination incentives in the opposite direction—namely, the risk of perceiving oneself as incompetent at solving a simple task. While we cannot entirely rule out this concern, we consider it unlikely to have had a major impact on our results, based on recent evidence from Exley and Kessler (2024) on "motivated errors" in very simple tasks. In their study, subjects choose between a fixed amount for themselves and a sum of amounts for charity. Strikingly, the simple addition of a "0" to the charity amount induces subjects to make calculation errors that justify the selfish choice. When subjects have no personal stake in the allocation decision, these errors disappear. These results support our interpretation that concerns about appearing incompetent likely played a minimal role, as individuals are often willing to appear inattentive or error-prone when doing so serves a self-interested purpose.

5 Conclusion

This study investigated the robustness of promise keeping by providing laboratory participants with responsibility-diffusing excuses designed to reduce the image costs associated with breaking a promise. We find clear evidence of social-image concerns. When the cause of a broken promises is undeniable, individuals are significantly more likely to honor it, suggesting an aversion to being perceived by others as promise breakers. This finding reinforces the idea that reputational considerations are a key driver of cooperative behavior.

To test for self-image concerns, we examined whether participants engaged in self-deceptive strategies—specifically, procrastination—as a way to evade promise-induced commitments. Our analysis revealed no such evidence. This null result may be interpreted as supporting evidence of an intrinsic preference for promise keeping. Thus, while some subjects are sensitive to whether they are seen as promise breakers, others exhibit stable preferences for promise keeping that are not easily eroded by moral wiggle room.

In sum, our study adds nuance to the literature on communication and trust, showing that the context in which a promise is made—and the availability of excuses—can significantly influence its credibility and impact.

Author's note

I would like to thank Anders and Odile Poulsen, Amrish Patel, Christoph Vanberg, David Hugh-Jones, Joël Van der Weele, Kiryl Khalmetski, Robert Sugden as well as the audiences of the 2018 CCC (CBESS-CEDEX-CREED) meeting, the 14th TIBER Symposium on Psychology and Economics, the 2019 European meeting of the Economic Science Association, and the 2023 meeting of the German Association for Experimental Economic Research for helpful comments and feedback. A preliminary version of this paper has been circulated under the title "Exploring Image Motivation in Promise Keeping–An Experimental Investigation."

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the ECO Ethics Committee of the University of East Anglia. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

KG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The author thanks Anders Poulsen, Odile Poulsen, Robert Sugden, and Johann Graf Lambsdorff for financially supporting the experiments and the publication of this research.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frbhe.2025.1631806/full#supplementary-material

References

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *Am. Econ. Rev.* 101, 470–492. doi: 10.1257/aer.101.2.470

Amelio, A., and Zimmermann, F. (2023). Motivated memory in economics - a review. *Games* 14:15. doi: 10.3390/g14010015

Andreoni, J., and Bernheim, B. D. (2009). Social image and the 50–50 norm: a theoretical and experimental analysis of audience effects. *Econometrica* 77, 1607–1636. doi: 10.3982/ECTA7384

Andreoni, J., and Sanchez, A. (2020). Fooling myself or fooling observers? Avoiding social pressures by manipulating perceptions of deservingness of others. *Econ. Inq.* 58, 12–33. doi: 10.1111/ecin.12777

Ariely, D., Bracha, A., and Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *Am. Econ. Rev.* 99, 544–555. doi: 10.1257/aer.99.1.544

Bašić, Z., and Quercia, S. (2022). The influence of self and social image concerns on lying. *Games Econ. Behav.* 133, 162–169. doi: 10.1016/j.geb.2022.02.006

Bem, D. J. (1972). "Self-perception theory," in *Advances in Experimental Social Psychology, Vol. 6* (Amsterdam: Elsevier), 1–62. doi: 10.1016/S0065-2601(08)60024-6

Bénabou, R., and Tirole, J. (2004). Willpower and personal rules. *J. Polit. Econ.* 112, 848–886. doi: 10.1086/421167

Bénabou, R., and Tirole, J. (2006). Incentives and prosocial behavior. *Am. Econ. Rev.* 96, 1652–1678. doi: 10.1257/aer.96.5.1652

Bénabou, R., and Tirole, J. (2011). Identity, morals, and taboos: beliefs as assets. *Q. J. Econ.*, 126, 805–855. doi: 10.1093/qje/qjr002

Bicchieri, C., Dimant, E., and Sonderegger, S. (2023). It's not a lie if you believe the norm does not apply: conditional norm-following and belief distortion. *Games Econ. Behav.* 138, 321–354. doi: 10.1016/j.geb.2023.01.005

Bodner, R., and Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. *Psychol. Econ. Decis.* 1, 105–26. doi: 10.1093/oso/9780199251063.003.0006

Bursztyn, L., and Jensen, R. (2017). Social image and economic behavior in the field: identifying, understanding, and shaping social pressure. *Annu. Rev. Econom.* 9, 131–153. doi: 10.1146/annurev-economics-063016-103625

Cadsby, C. B., Du, N., Song, F., and Yao, L. (2015). Promise keeping, relational closeness, and identifiability: an experimental investigation in china. *J. Behav. Exp. Econ.* 57, 120–133. doi: 10.1016/j.socec.2015.05.004

Charness, G., and Dufwenberg, M. (2006). Promises and partnership. *Econometrica* 74, 1579–1601. doi: 10.1111/j.1468-0262.2006.00719.x

Charness, G., and Dufwenberg, M. (2010). Bare promises: an experiment. *Econ. Lett.* 107, 281–283. doi: 10.1016/j.econlet.2010.02.009

Chen, Y., and Zhang, Y. (2021). Do elicited promises affect people's trust?-observations in the trust game experiment. *J. Behav. Exp. Econ.* 93:101726. doi: 10.1016/j.socec.2021.101726

Cherry, T. L., Frykblom, P., and Shogren, J. F. (2002). Hardnose the dictator. *Am. Econ. Rev.* 92, 1218–1221. doi: 10.1257/00028280260344740

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Econ. Theory* 33, 67–80. doi: 10.1007/s00199-006-0153-z

Deck, C., Servátka, M., and Tucker, S. (2013). An examination of the effect of messages on cooperation under double-blind and single-blind payoff procedures. *Exp. Econ.* 16, 597–607. doi: 10.1007/s10683-013-9353-0

Di Bartolomeo, G., Dufwenberg, M., and Papa, S. (2023a). Promises and partner-switch. *J. Econ. Sci. Assoc.* 9, 77–89. doi: 10.1007/s40881-023-00128-4

Di Bartolomeo, G., Dufwenberg, M., Papa, S., and Passarelli, F. (2019). Promises, expectations & causation. *Games Econ. Behav.* 113, 137–146. doi: 10.1016/j.geb.2018.07.009

Di Bartolomeo, G., Dufwenberg, M., Papa, S., and Passarelli, F. (2023b). Promises or agreements? Moral commitments in bilateral communication. *Econ. Lett.* 222:110931. doi: 10.1016/j.econlet.2022.110931

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: avoiding altruism by distorting beliefs about others' altruism. *Am. Econ. Rev.* 105, 3416–3442. doi: 10.1257/aer.20141409

Ederer, F., and Stremitzer, A. (2017). Promises and expectations. *Games Econ. Behav.* 106, 161–178. doi: 10.1016/j.geb.2017.09.012

Ellingsen, T., and Johannesson, M. (2004). Promises, threats and fairness. *Econ. J.* 114, 397–420. doi: 10.1111/j.1468-0297.2004.00214.x

Exley, C. L. (2015). Excusing selfishness in charitable giving: the role of risk. *Rev. Econ. Stud.* 83, 587–628. doi: 10.1093/restud/rdv051

Exley, C. L., and Kessler, J. B. (2024). Motivated errors. *Am. Econ. Rev.* 114, 961–987. doi: 10.1257/aer.20191849

Falk, A. (2021). Facing yourself-a note on self-image. *J. Econ. Behav. Organ.* 186, 724–734. doi: 10.1016/j.jebo.2020.11.003

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* 10, 171–178. doi: 10.1007/s10683-006-9159-4

Friedrichsen, J., and Engelmann, D. (2018). Who cares about social image? *Eur. Econ. Rev.* 110, 61–77. doi: 10.1016/j.euroecorev.2018.08.001

Gino, F., Norton, M. I., and Weber, R. A. (2016). Motivated bayesians: feeling moral while acting egoistically. *J. Econ. Perspect.* 30, 189–212. doi: 10.1257/jep.30.3.189

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *Am. Econ. Rev.* 108, 419–453. doi: 10.1257/aer.20161553

Grossman, Z. (2015). Self-signaling and social-signaling in giving. *J. Econ. Behav. Organ.* 117, 26–39. doi: 10.1016/j.jebo.2015.05.008

Grossman, Z., and Van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *J. Eur. Econ. Assoc.* 15, 173–217. doi: 10.1093/jeea/jvw001

Haisley, E. C., and Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games Econ. Behav.* 68, 614–625. doi: 10.1016/j.geb.2009.08.002

Hamman, J. R., Loewenstein, G., and Weber, R. A. (2010). Self-interest through delegation: an additional rationale for the principal-agent relationship. *Am. Econ. Rev.* 100, 1826–1846. doi: 10.1257/aer.100.4.1826

Ismayilov, H., and Potters, J. (2016). Why do promises affect trustworthiness, or do they? *Exp. Econ.* 19, 382–393. doi: 10.1007/s10683-015-9444-1

Johansson-Stenman, O., and Svedsäter, H. (2012). Self-image and valuation of moral goods: stated versus actual willingness to pay. *J. Econ. Behav. Organ.* 84, 879–891. doi: 10.1016/j.jebo.2012.10.006

Kingsuwankul, S., Tergiman, C., and Villeval, M. C. (2023). *Why do Oaths Work? Image Concerns and Credibility in Promise Keeping.* Technical report, Tinbergen Institute Discussion Paper TI 2023-058/I. Available online at: https://papers.tinbergen.nl/23058.pdf (Accessed August 2, 2025).

Kriss, P. H., Weber, R. A., and Xiao, E. (2016). Turning a blind eye, but not the other cheek: an the robustness of costly punishment. *J. Econ. Behav. Organ.* 128, 159–177. doi: 10.1016/j.jebo.2016.05.017

Lang, M., and Schudy, S. (2023). (Dis)honesty and the value of transparency for campaign promises. *Eur. Econ. Rev.* 159:104560. doi: 10.1016/j.euroecorev.2023.104560

Linardi, S., and McConnell, M. A. (2011). No excuses for good behavior: volunteering and the social environment. *J. Public Econ.* 95, 445–454. doi: 10.1016/j.jpubeco.2010.06.020

Malmendier, U., te Velde, V. L., and Weber, R. A. (2014). Rethinking reciprocity. *Annu. Rev. Econom.* 6, 849–874. doi: 10.1146/annurev-economics-080213-041312

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: a theory of self-concept maintenance. *J. Mark. Res.* 45, 633–644. doi: 10.1509/jmkr.45.6.633

Mischkowski, D., Stone, R., and Stremitzer, A. (2019). Promises, expectations, and social cooperation. *J. Law Econ.* 62, 687–712. doi: 10.1086/706075

Regner, T. (2018). Reciprocity under moral wiggle room: is it a preference or a constraint? *Exp. Econ.* 21, 779–792. doi: 10.1007/s10683-017-9551-2

Robbett, A., Walsh, H., and Matthews, P. H. (2024). Moral wiggle room and group favoritism among political partisans. *PNAS* 3:307. doi: 10.1093/pnasnexus/pgae307

Saucet, C., and Villeval, M. C. (2019). Motivated memory in dictator games. *Games Econ. Behav.* 117, 250–275. doi: 10.1016/j.geb.2019.05.011

Schütte, M., and Thoma, C. (2014). *Promises and Image Concerns.* Technical report, Munich Discussion Paper No, 2014-18. Available online at: https://epub.ub.uni-muenchen.de/20861/ (Accesssed August 2, 2025).

Schwartz, S., Spires, E., and Young, R. (2019). Why do people keep their promises? A further investigation. *Exp. Econ.* 22, 530–551. doi: 10.1007/s10683-018-9567-2

Sengupta, A., and Vanberg, C. (2023). Promise keeping and reliance damage. *Eur. Econ. Rev.* 152:104344. doi: 10.1016/j.euroecorev.2022.104344

van der Weele, J. J., Kulisa, J., Kosfeld, M., and Friebel, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *Am. Econ. J. Microecon.* 6, 256–264. doi: 10.1257/mic.6.3.256

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations. *Econometrica* 76, 1467–1480. doi: 10.3982/ECTA7673

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Exp. Econ.* 13, 75–98. doi: 10.1007/s10683-009-9230-z