Journal of the Economic Science Association (2023) 9:77–89
https://doi.org/10.1007/s40881-023-00128-4

ORIGINAL PAPER

Promises and partner-switch

Giovanni Di Bartolomeo[1,2] · **Martin Dufwenberg**[3,4,5] · **Stefano Papa**[6,7]

Received: 4 September 2022 / Revised: 18 December 2022 / Accepted: 21 February 2023 /
Published online: 25 March 2023

Abstract

Building on a partner-switching mechanism, we experimentally test two theories that posit different reasons why promises breed trust and cooperation. The expectation-based explanation (EBE) operates via belief-dependent guilt aversion, while the commitment-based explanation (CBE) suggests that promises offer commitment power via a (belief-independent) preference to keep one's word. Previous research performed a similar test, which we argue should be interpreted as concerning informal agreements rather than (unilateral) promises.

Keywords Promises · Partner-switching · Expectations · Commitment · Guilt · Informal agreements

JEL Classification A13 · C91 · D01 · D64

✉ Giovanni Di Bartolomeo
giovanni.dibartolomeo@uniroma1.it

Martin Dufwenberg
martind@eller.arizona.edu

Stefano Papa
stefano.papa@uniroma2.it

[1] Department of Economics and Law and CIMEO, Sapienza University of Rome, Rome, Italy

[2] Department of Economics, University of Antwerp, Antwerp, Belgium

[3] Department of Economics, University of Arizona, Tucson, USA

[4] Department of Economics, University of Gothenburg, Gothenburg, Sweden

[5] CESifo, Munich, Germany

[6] Department of Economics and Finance, University of Rome Tor Vergata, Rome, Italy

[7] CIMEO, Sapienza University of Rome, Rome, Italy

1 Introduction

Promises may foster trust and cooperation. A literature explores why. Charness and Dufwenberg (2006) (C&D) propose and report experimental support for an *expectation-based explanation* (EBE): A promise feeds a self-fulfilling circle of beliefs about beliefs. Promises are honored because if a person breaks his promise, then he would experience guilt for letting down the co-player's expectation.[1] Therefore, the co-player trusts the promisor. Vanberg (2008) proposes an alternative *commitment-based explanation* (CBE): People like to keep their word.[2] To experimentally test CBE it is crucial to develop a design that *exogenously* varies whether a player sends a promise to another. Vanberg ran an experiment that achieved that by relying on an ingenious "partner-switching" feature. His results support CBE.[3]

While the title of Vanberg's paper includes the question, "Why do people keep their promises?" his approach to CBE is broader as he also refers to obligations "based on agreements or contracts" (p. 1467). His experiment reflects this too. Let us highlight two differences between C&D's and Vanberg's designs. First, C&D focus on a binary trust game, where two players move in sequence. Vanberg instead explores a symmetrized dictator game, where only one player is active along any play path and where players initially do not know their role (dictator or recipient). Second, C&D and Vanberg explore different communication protocols. C&D study a single pre-play message that cannot be responded to. Vanberg instead allows subjects to send messages back and forth. If they then reciprocate each other's promises, their exchange may have the flavor of a conversation that generates an informal agreement.

A summary of Vanberg's contributions reveals some interesting remaining unchartered research territory: First, Vanberg identified a potential confound to C&D's result, namely CBE. Second, he also developed a design-tool – partner-switching – that allows testing for CBE. Third, he found support for CBE in a relevant context (with messages back-and-forth). However, he did not run a test of the relevance of CBE in C&D's context (with unilateral messages). While he identified a potential confound to C&D's result, he did not test its relevance in C&D's setting. Since C&D's study has garnered much interest, and given that the difference between messages back-and-forth and unilateral messages may be psychologically relevant, we propose that running such a test is of interest. In this paper, we report results from a design that accomplishes this.

We did not enter this research exercise with strong prior ideas as to how and why promises and informal agreements might trigger different forms of motivation and behavior. Nevertheless, interest in exploring related issues is enhanced by noting that several papers have documented that sometimes the nature of a communication protocol matters to behavior. For example, Brandts et al. (2019) survey how

[1] EBE is grounded in the theory of guilt aversion. See Battigalli & Dufwenberg (2007) for a general approach based on psychological game theory (compare Battigalli & Dufwenberg 2009, 2022; Geanakoplos et al., 1989).

[2] Ostrom et al. (1992), Ellingsen & Johannesson (2004), Charness and Dufwenberg (2010: Sect. 5.2), and Di Bartolomeo et al. (2019a) discuss similar ideas.

[3] Di Bartolomeo et al. (2019b) report similar results from a related design; we elaborate on nuances in Sect. 2.

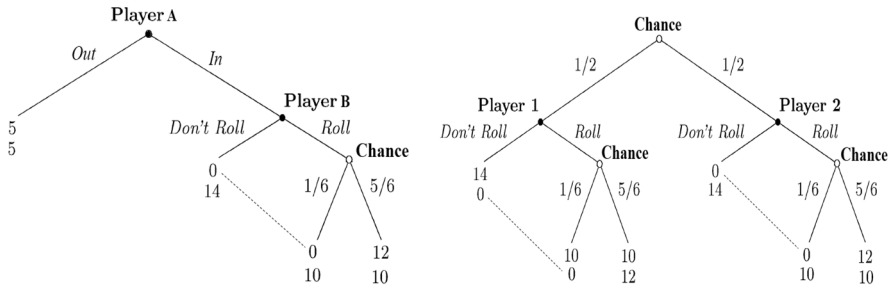

Fig. 1 The game trees: C&D's (to the left) and Vanberg's (to the right)

different communication structures have different impacts. In particular, they highlight dimensions that refer to messages' order, direction, and frequency—all elements that differ in the communication protocols associated with the two designs described in Fig. 1. The classification of communication contents also discovers different channels through which communication affects choices, potentially driven by different rationales.[4] To give an example, Krupka et al. (2017), investigating informal agreements, emphasize the social norm channel for their efficacy.[5] Relatedly, it may seem reasonable that communication that goes back and forth through multiple messages could create personal bonds and support the maintenance of agreements through moral commitment mechanisms, as effectively supported by Vanberg's results. However, the possibility of creating two-way links is nil with unilateral messages, suggesting that the psychology of promise-keeping mechanisms may be different in such context.

Besides evaluating CBE in C&D's trust game setting with unilateral promises, we also report additional results regarding EBE. Vanberg's switching feature, in fact, allows a relevant test, although there are limits to which extent this can be done (as noted by Ederer & Stremitzer, 2017 and Di Bartolomeo et al., 2019b.) We postpone a discussion of details until Sect. 2.

The rest of the paper is organized as follows. Section 2 provides in-depth scientific background: hypotheses, designs (C&D's, Vanberg's, ours, and also that of Di Bartolomeo et al.,'s 2019b, which helps add perspective), and other related literature. Section 3 describes our procedures. Section 4 explains what we found. Section 5 concludes.

2 Scientific background

We first recall what C&D and Vanberg did, then explain what we added. Figure 1 depicts C&D's game (form) to the left and Vanberg's to the right. Note how they differ, as indicated above.

[4] Other aspects than classification can be relevant. For instance, Ederer & Schneider (2020) investigate the relationship between time and trust with and without pre-play communication (mainly promises). Nielsen et al. (2019) compare the behavior of individuals and teams in trust games with pre-play communication.

[5] See also Kessler & Leider (2012) and Dufwenberg et al. (2017), Di Bartolomeo et al. (2023).

2.1 EBE

C&D explored experimental treatments with and without pre-play communication. In one treatment, B could send a single pre-play message to A. Suppose that B experiences guilt if he chooses *Don't*, and that the guilt increases the more strongly B believes that A believes B will choose *Roll*.[6] A promise from B to A may then feed a self-fulfilling circle of beliefs about beliefs that B will *Roll*, and, therefore, A will choose *In*. C&D articulated this idea – aka EBE – tested it, and they found support.

2.2 CBE

Vanberg points out that C&D's story is confounded. Suppose B has an innate preference for keeping his word: If B promises to *Roll* then he will prefer not to renege. If A anticipates this, he will choose *In*. This idea – aka CBE – generates the exact prediction as EBE.

2.3 Partner-switching

Vanberg came up with a clever experimental device to test the empirical relevance of CBE, enabling him to draw robust causal inferences regarding the impact of a promise. Namely, he proposed that if subjects i and j formed a chatting pair and then i was chosen to be the dictator, then with 50% probability, j would be "switched" and replaced by another subject k who previously chatted with subject l. Moreover, if there were a switch, i, but not k, would be informed of this. For cases where l sent a similar message to k as i sent to j (note: i could read l's message to k) EBE suggests that i would behave the same way with or without a switch. CBE, by contrast, implies that i will fulfill the promise if and only if there were no switch.[7]

2.4 Vanberg's results

Vanberg did not base his test of CBE on C&D's game but on the game to the right in Fig. 1. That is, if a subject i were selected to be the dictator (by the initial chance move), then subject j, with whom i had initially communicated, would be switched to another subject k who had initially communicated with a fourth subject l. Moreover, instead of using C&D's single-shot messages from one player to the other, Vanberg allowed the two players to engage in four rounds of back-and-forth messaging. Based on this design, Vanberg reported support for CBE.

[6] Note the reference to guilt aversion (compare with footnote 1). Several other experiments, starting with Dufwenberg & Gneezy (2000), tested hypotheses related to guilt aversion, often without communication in the picture; among others, Engler et al. (2018). See Cartwright (2019) and Rimbaud (2021) for surveys.

[7] A methodologically attractive feature of Vanberg's switching methodology is that randomization is really at the individual level, not at the session level (which is otherwise common in many experiments, although results are at times unjustifiably interpreted as if the randomization was at the individual level).

Vanberg, furthermore, elicited subjects' (first- and second-order) beliefs that a dictator would *Roll*, and he documented that these beliefs increased when subjects that received a promise were involved. Given this data pattern, Vanberg's design admits the following clean test of EBE: Consider *i*, *j*, *k*, *l* as described in the previous paragraph, and suppose that *i* made a promise to *Roll* to *j* and is switched. EBE implies that *i* should be more likely to *Roll* if *l* made a promise to *Roll* to *k* than if *l* did not make such a promise. However, Vanberg did not find support for this prediction.

2.5 Unchartered territory

It is natural to wonder whether Vanberg's key results – support for CBE, not support for EBE – would also be obtained in C&D's setting. The back-and-forth nature of Vanberg's communication protocol may generate experiences that look like informal agreements and have a different flavor than one-sided promises. While informal agreements and one-sided promises both evoke issues of keeping-one's-word, the modes of exchange are conceptually distinct.They may relate differently to how and why trust and cooperation may be induced.

2.6 Our design

We apply Vanberg's partner-switching feature to C&D's original game, thus providing new independent tests of CBE and EBE in C&D's setting. A subject in B's position can send a single written free-form message to a subject in A's position. Subsequently, there was a 50% probability that the A-subject would be switched and replaced by another subject (also in the position of A) who previously received a message from yet another subject *l* (in the position of player B). If there was a switch, only the B-subject was informed.

2.7 Di Bartolomeo et al. (2019b) (DDPP)

To clarify our contribution further, it is helpful to compare it with our previous study DDPP (the extra P is Francesco Passarelli's). DDPP tweak Vanberg's design so that the switching probability is not 50% but rather (depending on treatment) 25% or 75%. This induces exogenous variation in whether a player who sent a promise was paired with whoever received that promise (as needed for Vanberg's test of CBE) and in players' beliefs.[8] This new feature allows DDPP to run several new tests of EBE and guilt aversion and CBE. Their main results align with

[8] DDPP explain that "in light of the relevance of CBE (as documented by Vanberg), it is plausible that people expect dictators to be more inclined to keep their own promise than a promise made by someone else. Hence, recipients who received a promise should expect it to be kept with a higher probability if the switching probability is low (i.e., 25%) rather than high (i.e., 75%). And, if dictators understand that, their second-order beliefs should vary by switching probability in the same direction as recipients' first-order beliefs.".

Vanberg's, supporting CBE but not EBE. Our current design differs in two respects from DDPP's. First, we work with C&D's game rather than Vanberg's. Second, we use Vanberg's switching probability rather than those of DDPP. In other words, we conduct tests a la Vanberg in C&D's game, whereas DDPP modify Vanberg's tests while staying close to his game.

2.8 Less closely related literature

So far, in this section, we referenced the studies that most closely relate to (and directly motivate) us. All this work is part of broader experimental literature on communication in various trust games. We offer brief comments to give readers a richer context and backdrop.

Ellingsen and Johannesson (2004) are pioneers in studies that considered preferences for promise-keeping in trust games. Much of their focus, unlike ours, is on theories that refer to fair distributions of material rewards. They find that promises to behave reasonably mitigate the hold-up problem as investors rely on their trading partner's promises.

Ederer and Stremitzer (2017) propose a novel design that includes an "unreliable random device," creating exogenous variation in players' expectations and allowing clean tests of EBE. Their exercise may be seen as parallel to Vanberg's, except that the primary focus is on EBE rather than CBE. Their paper relates more closely to DDPP (as discussed above) than our current work. They provide evidence that promisors' aversion to disappointing others' expectations leads them to behave more generously.

Ismayilov and Potters (2016, 2017) stress the endogeneity nature of promises in exploring alternative motivations for promise-keeping. They explore "internal consistency" (desired by the party issuing a promise) and "social obligation" (felt because someone received one's promise). Investigating causality, in the end, they argue that promises do not induce trustworthiness. Instead, cooperators are more likely to send promises than non-cooperators.

As mentioned also above, differences in communication aspects matter (Brandts et al., 2019). In a context like ours, Charness and Dufwenberg (2010) and Di Bartolomeo et al. (2019a) allowed Bs to send or not send a prefabricated bare promise-to-Roll to A. Charness and Dufwenberg (2010) found that these promises did not affect beliefs and *Roll* rates. Instead, Di Bartolomeo et al. (2019a) focused on silent agents, finding that passive communication (silence) matters as it signals and somehow justifies selfish behavior.

Several recent studies have found that informal agreements may significantly affect subsequent play. For example, Krupka et al. (2017), who elicit social norms in a context with or without them, estimate that honoring an informal agreement in the double dictator game is worth giving up approximately 10% of total earnings. They also compare social norms to guilt aversion and lying aversion. And report that informal agreements affect behavior through their direct effect on social norms as well as through an indirect effect on beliefs. For more work on formal

agreements, see Miettinen (2013) and Dufwenberg et al. (2017) for theory and Kessler and Leider (2012), Dufwenberg et al. (2017), and Di Bartolomeo et al. (2023) for experiments.

3 Procedures

Our experiment was conducted at the CIMEO Lab of Sapienza University of Rome in May 2019. It involved 226 undergraduate students (8 sessions) recruited using an online recruitment system.[9] Upon arrival at the lab, subjects were randomly assigned to isolated computer terminals. Three assistants handed out instructions and checked that participants correctly followed the procedures. Before playing any game, subjects completed a short questionnaire testing their comprehension.

Each session consisted of 10 rounds, with perfect stranger matching. At the end of each session, one of the rounds was randomly chosen for payment. All subjects received a fixed show-up fee of 2.50 tokens, where 1 token = 0.5 euro.

Each round implemented the following sequence of six stages.[10]

1. Role assignment: Player positions B and A are randomly assigned, and pairs are formed.
2. Communication: B can send a free-form message to A ($\leq$90 characters).
3. A's action: A reads B's message, and then A has to choose *In* or *Out*.
4. Switching: Some As were switched with a 50% probability. Only Bs were informed whether or not a switch occurred. Bs with switched As were allowed to read the message previously received by the new A's pre-switch B.[11]
5. Belief elicitation: This stage has two parts: (a) First-order beliefs: each A was asked to guess if his/her unknown B would choose to *Roll* or *Don't*; (b) Second-order beliefs: each B was asked to guess the guess of the A with whom they would play after the switching stage occurred.[12]

[9] Across the 8 sessions, there were 28 subjects in session 1; 22 in 2; 30 in 3; 30 in 4; 26 in 5; 30 in 6; 30 in 7; 30 in 8, for a total of 226 participants. Compared to Vanberg (2008), we added two sessions and two rounds as, in our setup, potential useful observations used in the test fall as some As can choose to exit. In this way, assuming an *Out* rate of about 40%, which is a realistic value, we predicted obtaining about the same number of useful observations (i.e., dictator involved in the second stage of the game). The *Out*-rate (observed) was 39%, with 691 useful observations against 768 in Vanberg. Note that the participants were 226 in our experiment against 192 in Vanberg's. The difference with the prediction is that sessions effectively ran did not always involve 30 participants.

[10] Note that our design does not involve deception of As as all rules are known from the begging to all players.

[11] Only in pairs where A chose *In*, switches were possible (with 50% probability). In other pairs, where A chose *Out*, the game finished in that round. Note that when pairs choosing *In* were odd-numbered, half-plus-one pairs were switched, so the probability of being switched was slightly higher than 0.5.

[12] Specifically, if B's partner is switched, then B must guess the guess of the new partner after she reads the message that the new partner has received from his old partner during communication. Conversely, if partner of B is not switched, B must guess the guess of the partner with whom she communicated before.

6. B's action: B chooses between *Roll* or *Don't*. Then all subjects are informed about their payoff in that round. As are neither informed whether they had been switched nor about B's choice; only payoffs are revealed.[13]

Eliciting first- and second-order beliefs is common in the experimental literature on guilt aversion. Doing so here allows us to compare our findings regarding beliefs to C&D's. Incentives for beliefs elicitation were provided for all rounds except the one chosen for payment, implying that subjects had no incentive to hedge against bad outcomes and thus to misreport their beliefs.[14]

4 Results

We report our main findings related to CBE and EBE in Sect. 4.2. However, for the relevance of those tests, it is first critical to document that, in our design, promises on balance have the effect of raising subjects' expectations and *Roll* rates. We establish that in Sect. 4.1.

4.1 Promises, second-order beliefs, and roll rates

Our sample consists of 1130 pairs of subjects and 1130 messages from B to A. We code these messages according to whether or not they conveyed a promise to *Roll* (or a similar-in-spirit clear statement of intent to *Roll*).[15] This way, we obtained 527 promises out of 1130 messages (47%). The proportion of As who choose *In* is 61% (691 out of 1130). The proportion of As who choose *In* after receiving a promise is 76% (398 out of 527). The proportion of As who chose *In* when Bs did not promise is 49% (293 out of 603). As expected, the proportion of As who choose *In* after receiving a promise is significantly higher than that of As who do not receive a promise (76% vs. 49%: $Z=2.52$, $p=0.012$).

Let us now focus on second-order beliefs.[16] As said before, we observe that As chooses *In* in 691 cases, so we also have 691 Bs who choose between *Roll* and *Don't*. However, due to the partner-switching feature, these Bs are not necessarily those who sent a message to the As choosing *In*. Table 1 reports their second-order beliefs (in bold), the Bs' beliefs about the probability that an A subject believed B would roll the die, and reports standard deviations and the number of observations (in brackets). The first two columns (a and b) refer to the non-switched cases, while the second and third columns (c and d) refer to Bs whose partner was switched. Odd

[13] Participants A could obtain a zero payoff either because B chose to *Roll* or because the outcome of the die-roll was #1 when B chose to *Roll*.

[14] Our elicitation procedure is described in detail in the online Appendix.

[15] Following Vanberg, we asked two assistants to code the messages, having decided ex-ante to use the code of only one of them. Assistants were unaware of our choice. This way, we were able to check the robustness of the codification. The correlation between the two codes of messages is 0.89.

[16] First-order beliefs display similar patterns as the second-order beliefs and are reported in the online Appendix.

Table 1 Second-order beliefs of B's

	No switch		Switch	
	Promise read	No promise read	Promise read	No promise read
	(a)	(b)	(c)	(d)
(1) B makes a PROMISE	**70%** (0.29/204)		**67%**	**54%**
(2) B does not make a PROMISE		**55%** (0.31/120)	**59%** (0.31/90)	**57%** (0.30/83)

Pooled data from all sessions and all rounds. Standard errors and observations are in parentheses. The *Z*-statistic reflects Wilcoxon signed-rank tests using session-level data and average second-order beliefs

columns (a and c) refer to the case where Bs read a promise, while even ones (b and d) refer to the case where Bs did not. Rows indicate whether Bs made a promise (1) or not (2).

Looking at Bs with non-switched partners, as in C&D, we find that the second-order beliefs of Bs who made a promise are significantly different from those of Bs who did not send a promise (70% vs. 55%: $Z=2.38$, $p=0.017$).[17] It shows the correlation between promises and second-order beliefs, also found by C&D. Among the Bs who made a promise, the average second-order belief of those who read a promise is independent of the switch (70% vs. 67%: $Z=0.00$ and $p=1.000$), i.e., second-order beliefs of Bs with non-switched partners who made a promise are not significantly different from those of other Bs who made a promise and were re-matched with As who received a promise by someone else. Therefore, like Vanberg, we obtain exogenous variation in promises.

The second-order beliefs of switched promisors who are re-matched with an A who received a promise are higher than those of switched promisors who are re-matched with an A who did not receive a promise by someone else (67% vs. 54%: $Z=2.10$, $p=0.036$). Similarly, the second-order beliefs of Bs with non-switched partners who made a promise are higher than those of Bs who made a promise and are re-matched with an A who did not receive a promise by someone else (70% vs. 54%: $Z=2.52$, $p=0.012$).

The second-order beliefs of Bs who did not send a promise (all those in row (2)) are not statistically different.[18] This is suggestive of heterogeneity between agents

[17] The statistics reported are obtained from the Wilcoxon signed rank test, which compares averages at the session level. Our data are independent at the session level but not at the individual level. The Wilcoxon signed rank tests account for such structure in the data. Following Vanberg's, we also test the robustness of our results by using GLLAMM regressions on individual data. Estimation results are in line with the conclusions derived in the paper. Results are reported in the online Appendix.

[18] These beliefs are also not significantly different from the second-order beliefs of Bs with switched partners who sent a promise and were re-matched with As who did not receive a promise (row (1) column (d)).

Table 2 B's Roll rates

	No Switch		Switch	
	Promise read	No Promise read	Promise read	No Promise read
	(a)	(b)	(c)	(d)
(1) B makes a PROMISE	**74%** (0.44/204)		**70%** (0.46/104)	**59%** (0.49/90)
(2) B does not make a PROMISE		**29%** (0.46/120)	**31%** (0.47/90)	**39%** (0.49/83)

Pooled data from all sessions and all rounds. Standard errors and observations are in parentheses. The *Z*-statistic reflects Wilcoxon signed-rank tests using session-level data roll rates and average second-order beliefs

who self-select to send promises and those who do not, as reported by Ismayilov and Potters (2016).

4.2 Main results: CBE & EBE

Table 2 reports the average *Roll* rates of Bs (in bold), standard deviations, and the number of observations (between brackets). The structure is otherwise like that of Table 1. We distinguish the average *Roll* rates of Bs who promise and Bs who do not promise, by rows. Columns refer to the message they read and indicate whether a switch occurred.

Focusing on columns (a) and (b) of Tables 1 and 2, we observe a correlation between promise-keeping and second-order beliefs, as in C&D. Second-order beliefs of promisors are higher in Table 1 (70% vs. 55%: $Z=2.38$ and $p=0.017$) as are average *Roll* rates (74% vs. 29%: $Z=2.52$, $p=0.012$) in Table 2. Our results show that people are likelier to keep promises, which correlates with high second-order beliefs. However, as argued by Vanberg, correlation does not necessarily imply causation. We need to further investigate the issue by using our exogenous variation in promises.

Among Bs who made a promise, the average *Roll* rate of Bs with non-switched partners is not statistically different from that of Bs who read a promise made by someone else (74% vs. 70%: $Z=0.14$, $p=0.889$). Thus, we do not find support for CBE. The behavior of Bs with non-switched partners who keep their own promises is not different from that of Bs who keep promises done by another.

Our result here differs from that of Vanberg, who found support for CBE when he ran an analogous test. Instead, the result is consistent with the idea that people keep promises made by others since those are associated with higher second-order beliefs, as predicted by EBE.

Two direct tests for EBE are obtained by comparing promisor Bs (row (1) and column (a)) who read a promise whether they are switched (row (1) and column (c).) with those who did not (row (1) and column (d)). Remember that they have different second-order beliefs (columns (a) and (c) vs. (d).) Looking at Table 2, the average

Roll rate of Bs with non-switched partners and that of other Bs who read a promise made by someone else are both higher than that of Bs with switched partners who did not read a promise made by someone else (74% vs. 59%: $Z=2.52$, $p=0.012$; 70% vs. 59%: $Z=2.10$, $p=0.036$). This finding supports EBE. Again, our results differ from Vanberg's; he found no statistically significant difference and no support for EBE.

5 Conclusions

Humans are social animals who communicate. There is interest in exploring the psychological mechanisms that matter for different forms of communication. We distinguish between unilateral and bilateral messages. The former variety may produce a unilateral promise, while the latter variety may generate an informal agreement. The causal impact of promises or agreements may then be studied using Vanberg's (2008) partner-switching technique, the central idea being that a switch anullates either form of covenant.

C&D proposed a theory – EBE – that may explain why promises foster trust and cooperation. Vanberg pointed out that C&D's results are confounded: an alternative explanation – CBE – is conceivable. Using his partner-switching technique, he reported support for CBE and a lack of support for EBE. As Vanberg deviated from C&D's design as regards the choice of the game and communication protocol, it is natural to wonder if his main result also extends to a setting that uses C&Ds game and communication protocol. This is what we have explored. Our results are the opposite of Vanberg's, supporting EBE and not supporting CBE.

The findings can be linked to the distinction between promises and informal agreements. C&D's game is *asymmetric*. At the root, both players know that player A has to trust player B to *Roll*, not the other way around. By contrast, Vanberg's game is *symmetric*. At the root, both players know that either may have to trust the other to *Roll*. Moreover, the communication protocols differ, with a one-sided message from B to A in C&D's case and a conversation-like exchange in Vanberg's case. The symmetry of Vanberg's game, and the back-and-forth nature of his communication protocol, invites the reflection that players may be inclined and able to strike a deal of conditional cooperation: "I'll promise to *Roll* if you promise to *Roll*." And if both players do promise to *Roll*, their exchange has the semblance of an informal agreement. Vanberg's results are consistent with and supportive of the idea that players have a belief-independent preference to honor such agreements.

As regards CBE, there is no tension between our results and Vanberg's. If his study is interpreted as documenting evidence for a preference for honoring informal agreements, then this has no counterpart in our (or C&D's) design. A preference for keeping a unilateral promise may be a somewhat different animal than a preference for honoring a gentleman's agreement. Different forms of CBE are considered by Vanberg and by us.

Regarding EBE, it may seem puzzling that this theory is supported in C&D's setting but not Vanberg's. Data is what it is though, and, after observing the results,

we have the following reflections: Different games may trigger different thinking. Humans are motivated in many ways.[19] However, perhaps humans cannot consider more than a few such motivations at a time, and perhaps Vanberg's setting, relative to ours (and C&D's), triggers other motivations that may crowd out the belief-dependent feelings built into EBE? To mention two such potential motivations, consider the preference for honoring an informal agreement described above. Second, consider reciprocity, such that a player would wish to choose *Don't* [*Roll*] if and only if he or she believed that the other player would have done likewise had he or she been designated to choose whether or not to *Roll*. In Vanberg's game, this motivation would be potent (since there is a node where the other player may choose *Roll*), whereas, in C&D's game, it would be muted (since the other player has no *Roll* choice).

We hope that future research may take inspiration from these speculative remarks and develop new designs that may be useful for testing the empirical relevance of our ideas.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40881-023-00128-4.

Acknowledgements We are grateful to the Journal Editor Lionel Page and two referees for their very helpful comments. We also thank Francesco Bloise and Nicolò Ottaviano for their comments on previous versions.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Data availability Data are available from the authors upon request.

References

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review, 7*, 170–176.

Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory, 144*, 1–35.

Battigalli, P., & Dufwenberg, M. (2022). Belief-dependent motivations and psychological game theory. *Journal of Economic Literature, 60*, 882–883.

[19] By, e.g., reciprocity, emotions, and image concerns, on top of more classical items like income or concern for fair distributions. See Battigalli & Dufwenberg (2020) for a systematic discussion.

Brandts, J., Cooper, D. J., & Rott, C. (2019). Communication in laboratory experiments. In A. Schram & A. Ule (Eds.), *Handbook of research methods and applications in experimental economics* (21st ed., pp. 401–418). Elgar Publishing.

Cartwright, E. (2019). A survey of belief-based guilt aversion in trust and dictator games. *Journal of Economic Behavior and Organization, 167*, 430–444.

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

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

Di Bartolomeo, G., Dufwenberg, M., & Papa, S. (2019a). The sound of silence: A license to be selfish. *Economics Letters, 182*, 68–70.

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

Di Bartolomeo, G., Dufwenberg, M., Papa, S., & Passarelli, F. (2023). Promises or agreements? Moral commitments in bilateral communication. *Economics Letters, 222*, 110931.

Dufwenberg, M., & Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior, 30*, 163–182.

Dufwenberg, M., Servátka, M., & Vadovič, R. (2017). Honesty and informal agreements. *Games and Economic Behavior, 102*, 269–285.

Ederer, F., & Schneider, F. (2020). Trust and promises over time. *The American Economic Journal, 14*, 304–320.

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

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

Engler, Y., Kerschbamer, R., & Page, L. (2018). Guilt averse or reciprocal? Looking at behavioral motivations in the trust game. *Journal of the Economic Science Association, 4*(1), 1–14.

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

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.

Kessler, J., & Leider, S. (2012). Norms and contracting. *Management Science, 58*, 62–77.

Krupka, E. L., Leider, S., & Jiang, M. (2017). A meeting of the minds: Contracts and social norms. *Management Science, 63*, 1708–1729.

Miettinen, T. (2013). Promises and conventions: An approach to pre-play agreements. *Games and Economic Behavior, 80*, 68–84.

Nielsen, K., Bhattacharya, P., Kagel, J., & Sengupta, A. (2019). Teams promise but do not deliver. *Games and Economic Behavior, 117*, 420–432.

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

Rimbaud, C. (2021), "Introduction," in *Three Essays on Guilt Aversion: Theory and Experiments*. Ph.D. Thesis, GATE-LAB, University of Lyon.

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