UCLA
UCLA Public Law & Legal Theory Series

Title
Norm-Based Enforcement of Promises

Permalink
https://escholarship.org/uc/item/89g1z89s

Authors
Stone, Rebecca
Stremitzer, Alexander
Atkinson, Nathan

Publication Date
2019

eScholarship.org Powered by the California Digital Library
University of California

UCLA School of Law

Law & Economics Research Paper No. 19-04

NORM-BASED ENFORCEMENT OF PROMISES

BY

REBECCA STONE
PROFESSOR OF LAW
UCLA SCHOOL OF LAW

ALEX STREMITZER
ETH ZURICH, CENTER FOR LAW & ECONOMICS

NATHAN ATKINSON
ASSISTANT PROFESSOR OF LAW
UNIVERSITY OF WISCONSIN LAW SCHOOL

Norm-Based Enforcement of Promises*

Nathan Atkinson[†] University of Wisconsin
Rebecca Stone[‡] UCLA
Alexander Stremitzer[§] ETH Zurich

February 4, 2025

Abstract

Considerable evidence suggests that people are internally motivated to keep their promises. But it is unclear whether such internal motivations create a meaningful level of commitment in economically relevant situations where promisors have strong self-interested reasons to break their promises. Self-interested incentives to break promises compete with—and so may overcome—the moral motivations of those who are internally motivated to keep them. We experimentally investigate the willingness of third parties to altruistically punish promise breaking. Participants observe a transaction between a potential promisor and promisee and can incur costs to punish the promisor for uncooperative actions. Our results suggest that the same moral reasons that motivate promisors to keep their promises make third-party observers more likely to punish promise breaking. This suggests that the same underlying promissory norms drive both (i) promise-keeping behavior in the absence of second- and third-party enforcement mechanisms and (ii) non-legal enforcement mechanisms that arise when third parties can punish potential promisors in a decentralized fashion. This makes it more likely that promissory norms support cooperative behavior.

Keywords: promises, norms, first-party enforcement, second-party enforcement, altruistic punishment.

JEL-Classification: K12, L14, D86, D91, C91.

*The authors are grateful to Gary Charness, Zeynep Gurguc, and William Hubbard for valuable comments. We are also thankful to seminar audiences at Alabama, ETH Zurich, Northwestern School of Law, University of Texas School of Law, University of Zurich, the Annual Conference of the Polish Law and Economics Society, and the Theory and Applications of Contracts Conference at Imperial College London. We thank Henry Kim and Hamid Shazad for their excellent research assistance. Declarations of interest: The authors used funding from the discretionary ETH Research budget allocated to Alexander Stremitzer's laboratory. We have no conflicts of interest to report.

[†]University of Wisconsin Law School, 975 Bascom Mall, Madison, WI 53706. natkinson@wisc.edu.

[‡]UCLA Law School, 385 Charles E. Young Drive, 1242 Law Building, Los Angeles, CA 90095, rebecca.stone@law.ucla.edu.

[§]ETH Zurich, Center for Law & Economics, IFW E49, Haldeneggsteig 4, 8092 Zurich, Switzerland, astremitzer@ethz.ch.

1 Introduction

A series of experimental studies find that in the absence of legal enforcement and reputational concerns promisors are motivated to keep their promises even when they have self-interested reasons to break them (Ellingsen and Johannesson 2004; Charness and Dufwenberg, 2006; Vanberg, 2008; Charness and Dufwenberg, 2010). Promisors seem to be motivated by moral reasons when deciding whether or not to keep their promises: they care about keeping promises for their own sake; they care about not disappointing others' expectations (Dufwenberg and Gneezy, 2000; Guerra and Zizzo, 2004; Reuben, Sapienza, and Zingales, 2009; Bellemare, Sebald, and Strobel, 2011; Regner and Harth, 2014; Khalmetski, Ockenfels, and Werner, 2015; Mischkowski, Stone, and Stremitzer, 2019),[1] especially when those others have altered their behavior in reliance on such expectations (Stone and Stremitzer, 2020), and they are especially motivated to meet expectations that they themselves have set through their promises (Ederer and Stremitzer, 2017; Mischkowski, Stone, and Stremitzer, 2019).

But such moral motivations may not guarantee performance of a promise—even one that has been relied upon extensively by the promisee—when, as will often be the case in commercial settings, significant self-interested incentives to break a promise compete with and may overpower the promisor's moral motivations to keep it. And, of course, they will have no effect on the behavior of self-interested actors who aren't motivated by moral reasons to keep their promises in the first place. If so, effective second- and third-party enforcement mechanisms will be needed to give such promisors incentives to perform their promises in many economically significant situations.

It does not, however, follow that promissory norms will therefore be irrelevant to the behavior of even purely self-interested promisors. Centralized enforcement systems like the legal system are or arguably should be designed to reflect or at least support underlying

[1] With the exception of Mischkowski, Stone, and Stremitzer (2019) these studies derived their findings in dictator, trust, and lost wallet games where there was no direct promissory link between the parties. However, some implicate social norms that might create an obligation to perform, which like obligations that arise from promises might exist independently of the recipient's expectations. For example, a lost wallet game implicates a social norm that requires one to return someone else's property.

moral practices of their subjects.[2] Decentralized mechanisms of enforcement by disinterested third-party observers may also be shaped by promissory norms, thus providing assurance to promisees when legal mechanisms of enforcement are unavailable or costly to the parties. This, in turn, may help to support the legal institution of contracting. Parties can rely on more informal mechanisms of enforcement resorting only to litigation when relationships break down completely.[3] If the decentralized enforcement regime tracks the underlying moral norms closely, that would also help to explain why parties prefer to rely on informal enforcement mechanisms than to litigate, insofar as our existing legal norms sometimes diverge significantly from the underlying moral practice.[4]

Thus, we seek to investigate whether the motivations of third-party observers to altruistically punish promise breaking track the motivations of morally-motivated promisors to keep their promises. Our hypothesis is that the same norms that deem it wrong to break a promise will motivate promisors to keep their promises and inform the moral judgments of third-party observers. In light of theory and evidence that suggests that persons are willing to altruistically enforce widely held norms, we should then expect that those third parties will punish the promisor for breaking her promise, even if such punishment comes at some cost to themselves.[5] Reputational concerns that track agents' moral beliefs would then make even self-interested promisors act *as if* they cared about the underlying moral norms that govern promising.

Taking these results seriously has the important implication that underlying promissory norms influence not only the internal motivations of promisors (Dixit, 2009; Krupka, Leider and Jiang, 2017) but also decentralized third-party enforcement mechanisms. This implica-

[2]See, for instance Restatement (Second) of Contracts, Section 1 (1981) (defining a contract as an enforceable promise or set of promises).

[3]A considerable body of scholarship on contract argues that contracting parties rely heavily on non-legal mechanisms and that their doing so may be less disruptive to their relationships than resorting to litigation. See, e.g. Bernstein (1993), Bernstein (1996), Scott (1990), Scott (2003), Macaulay (1963). (1993), Bernstein (1996), Scott (1990), Scott (2003), Macaulay (1963).

[4]See Shiffrin (2007) for various ways in which contract law fails to reflect promissory morality.

[5]Gintis et al. (2005) provide an overview of the theory and evidence that supports the existence of dispositions towards strong reciprocity. Fehr and Fischbacher (2004) provide experimental evidence of subjects' propensities to punish third parties at a cost to themselves.

tion is in line with Hart and Moore's (2008) theory that contracts set reference points or expectations of contracting parties and that promisees punish promisors by "shading" on their performance when those expectations are disappointed. Hart and Moore (2008) concerns second-party enforcement—that is, enforcement by the party who has been wronged by a breach of promise. Our findings suggest third-party enforcement behavior may similarly be shaped by underlying promissory norms. Our results are also consistent with the vast body of theoretical and empirical findings that suggest that altruistically enforced norms are central determinants of cooperative behavior (Gintis et al., 2005; Fehr and Fischbacher, 2004; Chaudhuri, 2011).

The remainder of the paper is organized as follows. Section 2 describes a theoretical model that makes precise the motivation for our experiments. Section 3 describes our experimental design and hypotheses. Section 4 describes the procedures we used to implement our experiments. Section 5 describes our results. Section 6 concludes.

2 Altruistic Enforcement of Promissory Norms

In this section, we develop a simple model of decentralized, third-party punishment for unkept promises. Our central idea is that third-party observers may be motivated to punish broken promises at a personal cost, which in turn can deter purely self-interested agents from breaching promises that others have relied upon.

We analyze a three-player game involving a potential promisor, *Player A*, who can (i) *promise* or not promise to cooperate with another player, and (ii) *cooperate* or not cooperate once that promise is made or withheld. A promisee, *Player B*, who invests resources in reliance on cooperation by A. A neutral third party, *Player C*, who can punish A at a personal cost if A chooses not to cooperate.

Denote by $\alpha \in \{0, 1\}$ the decision of A to make a promise ($\alpha = 1$) or not ($\alpha = 0$). Let $\kappa \in \{0, 1\}$ indicate whether A *cooperates* ($\kappa = 1$) or *does not cooperate* ($\kappa = 0$). Player B invests an amount $\iota \in [0, \bar{\iota}]$, anticipating whether A will cooperate. Finally, C observes A's behavior and chooses a costly punishment level $\rho \geq 0$, reducing A's payoff by some multiple

of ρ.

Let the material payoffs be given as follows. A's payoff is $\pi_A(\kappa) - \delta\,\rho$, where $\pi_A(0) > \pi_A(1)$ captures that A gains more by not cooperating. The term $\delta\rho$ reflects the reduction in A's payoff caused by C's punishment, with $\delta > 1$. B's payoff is $\pi_B(\iota, \kappa)$. Let $\Delta_B(\iota) = \pi_B(\iota, 1) - \pi_B(\iota, 0)$ be the difference in B's payoff if A cooperates and B's payoff if A does not cooperate for an investment level of ι, which we assume is strictly increasing in ι. C's material payoff is $\pi_C(\rho) = -\rho$, so each unit of punishment ρ costs C one unit of payoff.

Under the standard model of economic behavior, C is influenced only by monetary payoffs, and the game is easily solvable through backwards induction. Player C maximizes her payoff with $\rho^* = 0$, Player A chooses $\kappa^* = 0$, and Player B chooses $\iota = 0$. The promise decision is not payoff relevant, so Player A can choose either $\alpha \in \{0, 1\}$.

Prior evidence suggests, however that promisors are intrinsically motivated to keep their promises (Ellingsen and Johannesson 2004; Charness and Dufwenberg, 2006; Vanberg, 2008; Charness and Dufwenberg, 2010). If so, then A might keep any promise made to B notwithstanding A's incentive to defect. But prior evidence also suggests that intrinsic motivations to keep promises are variable. If A is self-interested or only weakly morally motivated to keep her promises, A has no intrinsic reason to keep any promise made to B. But that assumes that C acts in a self-interested fashion. What if a third party like C is a believer in promissory norms and so is motivated to punish non-cooperative behavior by A, especially when A made a promise to cooperate? There is evidence that some third parties are willing to engage in altruistic punishment of extant norms (Gintis et al., 2005; Fehr and Fischbacher, 2004; Chaudhuri, 2011). So perhaps this is also true in the promissory realm.

Thus, suppose that Player C also experiences a *non-monetary* or *moral* payoff term, $\lambda_i\,\phi(\alpha, \iota, \kappa, \rho)$, where $\lambda_i \geq 0$ is an individual-specific weight on norm-enforcement and $\phi(\cdot)$ measures how "wrong" C perceives A's conduct to be, thus capturing the utility C derives

from punishing norm violations.[6] Overall, C's payoff can be written as

$$u_C(\alpha, \iota, \kappa, \rho) = -\rho + \lambda_i \phi(\alpha, \iota, \kappa, \rho). \tag{1}$$

The term λ_i is a person-specific scalar that represents the relative weight of the non-monetary component of the third party's utility. We assume that $\phi = 0$ whenever $\kappa = 1$. That is, the third party has no preference to punish A when A cooperates with B. Let $\rho^*(\alpha, \iota, \kappa)$ indicate C's utility-maximizing level of punishment as a function of the actions chosen by A and B. C is motivated to engage in altruistic punishment of non-cooperation by A whenever $\lambda_i > 0$.

Our primary conjecture is that an altruistic third-party in C's position will exhibit a preference to punish that captures the structure of promissory norms that have been revealed to govern the behavior of first-party promisors who are intrinsically motivated to keep their promises and otherwise act cooperatively absent self-interested reasons to do so (Mischkowski, Stone and Stremitzer, 2019). First, just as people seem to be more willing to cooperate when they promised to do so, we hypothesize that third parties will be more willing to punish the non-cooperative behavior of one who promised another that she would cooperate than one who didn't make such a promise (the promise per se effect). Formally, $\frac{\partial \rho^*}{\partial \alpha} > 0$.

Second, just as people seem to be more willing to cooperate, the greater are recipients' reliance investments, we hypothesize that third parties will be more willing to punish non-cooperation as investment levels of recipients increase, even when the non-cooperative agent made no promise to cooperate (the reliance per se effect). Formally $\frac{\partial \rho^*}{\partial \iota} > 0$.[7]

Finally, just as motivations to cooperate seem to be more influenced by greater reliance investments by recipients when they promised recipients that they would cooperate, we hypothesize that the willingness of third parties to punish non-cooperation will be more influenced by increased reliance by recipients when that non-cooperation breaks a promise to

[6] Given that our focus is on the effects of third parties on promisors who may be insufficiently morally motivated to keep their promises, we ignore promisors' moral motivations to focus on the effects of C's on the equilibrium of the game.

[7] It may be that the willingness to punish depends not directly on the investment level, but instead only or also on the difference between the payoffs following cooperation and non-cooperation. This would change the setup of the model, but would not change the qualitative results.

cooperate than in the absence of such a promise (the interaction effect). That is, punishment following a broken promise coupled with reliance is greater than the sum of the punishments following just a broken promise with no reliance plus the punishment following no promise with positive reliance. Formally, $\frac{\partial^2 \rho^*}{\partial \alpha \partial \iota} > 0$

A simple functional form that captures these assumptions is:

$$\phi(\alpha, \iota, \kappa, \rho) = \begin{cases} (\beta_\alpha \alpha + \beta_\iota \iota + \beta_{\alpha\iota} \alpha\iota) \ln(\rho) & \text{if } \kappa = 0 \\ 0 & \text{if } \kappa = 1, \end{cases} \tag{2}$$

for some vector $\beta \geq 0$, and where $\lambda_i \geq 0$ is a parameter capturing the intensity of C's willingness to punish.[8] If A cooperates, C has no preference to punish A. However, if A did not cooperate, the third party has such a preference.[9]

This game can be solved through backwards induction. First, solving for the first order condition on the third party's optimal punishment decision yields

$$\rho^* = \lambda_i(\beta_\alpha \alpha + \beta_\iota \iota + \beta_{\alpha\iota} \alpha\iota), \tag{3}$$

where $\frac{\partial \rho^*}{\partial \alpha} = \lambda_i(\beta_\alpha + \beta_{\alpha\iota}\iota) > 0$, $\frac{\partial \rho^*}{\partial \iota} = \lambda_i(\beta_\iota + \beta_{\alpha\iota}\alpha) > 0$, and $\frac{\partial^2 \rho^*}{\partial \alpha \partial \iota} = \lambda_i \beta_{\alpha\iota} > 0$. Therefore, C's willingness to punish non-cooperation will be higher when A made a promise than when A made no promise holding B's investment constant. C's willingness to punish non-cooperation will increase with B's investment level whether or not A made a promise. And this increase will be greater if A made a promise than if A made no promise.

Given C's punishment behavior, we next solve for A's cooperation choice. A should take into account the expected punishment when deciding whether or not to cooperate. Using C's behavior in (3), coupled with the assumption that the third party will not punish A when A cooperates, A's expected utility gain from choosing noncooperation over cooperation is given by $[\pi_A(0) - \pi_A(1)] - \delta * \rho^*(\alpha, \iota)$. Therefore the promisor prefers noncooperation when this is greater than 0. Define $\Delta_A = \pi_A(0) - \pi_A(1)$ as A's material gain from defecting (absent

[8] Note that the log formulation implicitly assumes an interior solution in this setup.

[9] Note that under this utility function, the third party may actually prefer to observe non-cooperation and impose punishment than simply observing cooperation. However, this is without loss of generality, because a negative term could be added to the utility function.

punishment). A's optimal cooperation decision is therefore:

$$\kappa^* = \begin{cases} 1 & \text{if } \Delta_A \leq \delta\rho^*(\alpha, \iota) \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

That is, if the punishment is high enough to offeset Δ_A, then A cooperates. Otherwise A defects. Thus, A's behavior depends on $E[\lambda]$, a measure of how other-regarding the typical C is. If the average C is significantly other-regarding, then it makes material sense for A to keep her promise. On the other hand, if third parties are not significantly other-regarding on average, then A maximizes her material payoff by not cooperating.

Given A's cooperation decision, we solve for B's optimal investment choice, $\iota^*(\alpha)$. B's optimal investment decision depends on whether or not A will cooperate. Let $\iota^\dagger$ be the smallest investment that B can make such that the cooperation condition is satisfied. That is:

$$\iota^\dagger \equiv \arg\min_{\iota \geq 0} \left\{ \delta\rho^*(\alpha, \iota) \geq \Delta_A \right\}. \tag{5}$$

Note that if some choice $\iota' > 0$ leads to non cooperation, $\kappa(\alpha, \iota') = 0$, then B would do better by choosing $\iota = 0$. Moreover, if B's payoff is increasing in investment conditional on cooperation, that is $\pi_B(\iota, 1)$ is strictly increasing in ι, B will choose $\iota = \bar{\iota}$ whenever $\kappa = 1$. Taken together B's optimal investment is given by:

$$\iota^*(\alpha) = \begin{cases} 0 & \text{if } \iota^\dagger > \bar{\iota} \\ \bar{\iota} & \text{if } \iota^\dagger < \bar{\iota} \end{cases} \tag{6}$$

That is, B will strategically invest based on expectations about how that investment will affect both the third party's punishment decision, and thereby, the promisor's cooperation decision.

Finally we consider A's decision whether or not to make a promise. By making a promise, A may induce B to invest more (and thus raise the punishment from non-cooperation), which can in turn increase the joint surplus if A ultimately cooperates. However, if A breaks a promise, the expected punishment will be weakly higher, making defection more costly. A's expected payoff is given by:

$$U_A(\alpha) = \max \left\{ \pi_A(1), \pi_A(0) - \delta\rho^*(\alpha, \iota^*(\alpha)) \right\}, \tag{7}$$

where the max is effectively determined by κ^* above. A's optimal choice of promising is thus given by:

$$\alpha^* = \arg \max_{\alpha \in \{0,1\}} \{U_A(\alpha)\}. \tag{8}$$

Our aim in this paper is not to directly test whether the threat of punishment has an effect on promise keeping behavior or investment behavior. If potential promisors and their promisees can predict the behavior of third parties, it is straightforward for them to include expected punishment in their decision calculus. That is, the credible threat of altruistic punishment for norm violations should increase compliance with norms, and increase the willingness of promisee's to rely on promises they receive. The aim of our paper is instead to determine the degree to which third-party motivations to engage in altruistic punishment reflect shared promissory norms and track the first-party motivations of promisors to keep their promises, thus suggesting that the structure of the underlying norms–and the variables that the norm picks out as morally relevant, such as the degree of reliance of the promisee–drive the behavior of first-party potential promisors and third-party punishers in a similar way. If the motivations do arise from shared norms in this way, then the presence of C should alter the behavior of even a self-interested A and so increase B's willingness to invest in their joint project.

Consider the following numerical example. Both players A and B have endowments of 70, while C has an endowment of 110. B can choose investment $\iota \in [0, 20]$, where the investment returns 0 when A does not cooperate and returns 2ι when A cooperates. A gets a benefit of 40 from cooperating and 50 from not cooperating. The punishment multiplier is $\delta = 3$. The monetary payoffs of the players are:

$$\begin{aligned}
\pi_A(\kappa, \rho) &= 70 + 50 - 10\kappa - 3\rho = 120 - 10\kappa - 3\rho \\
\pi_B(\iota, \kappa) &= 70 - \iota + \mathbf{1}_{\kappa=1}(20 * \iota) \\
\pi_C(\rho) &= 110 - \rho
\end{aligned}$$

Promisors should take into account the expected punishment when deciding whether or not to cooperate. Using the functional form in (2), coupled with the assumption that the

third party will not punish following cooperation, the promisor's expected utility gain from choosing noncooperation over cooperation is given by $E[\pi_A\kappa = 0 - \pi_A\kappa = 1] = 10 - 3\overrightarrow{p}^*$. Therefore the promisor prefers noncooperation when this is greater than 0. Solving yields:

$$(E[\lambda_i])(\beta_\alpha\alpha + \beta_\iota\iota + \beta_{\alpha\iota}\alpha\iota) < \frac{10}{3}. \tag{9}$$

A does not know the distribution of λ. If the average player is significantly other-regarding, then it makes material sense for the promisor to keep her promise. On the other hand, if third parties are not significantly other-regarding, then the promisors maximizes her material payoff through noncooperation.

Given the expected behavior of the promisor and the third party, the promisee needs to choose an optimal investment level. Given that the promisor's optimal decision is based on her expectations about third party behavior as captured in (4), the promisee's optimal investment decision is given by maximizing ι subject to $(E[\lambda_i])(\beta_\alpha\alpha + \beta_\iota\iota + \beta_{\alpha\iota}\alpha\iota) \geq \frac{10}{3}$. That is, the promisee's return is increasing in investment conditional on cooperation, and decreasing in investment conditional on non-cooperation.[10]. Note that the promisee's choice of investment itself can affect the third party's punishment choice (through β_ι and $\beta_{\alpha\iota}$).

Figure 2 plots equilibrium behavior of the three players as a function of the expected punishment following non-cooperation, $(E[\lambda_i])(\beta_\alpha\alpha + \beta_\iota\iota + \beta_{\alpha\iota}\alpha\iota)$. The vertical dotted line is at $\frac{10}{3}$, which is the promisor's break even point following equation (4). For expected punishment to the left of the dotted line, the promisor chooses non-cooperation, and for expected punishment to the right of the dotted line, the promisor chooses cooperation. The solid blue line thus illustrates C's expected punishment. To the left of the vertical line, A chooses non-cooperation and C imposes positive punishment. To the right of the vertical line, A chooses cooperation, so C does not impose any punishment. The dotted blue 45-degree line illustrates C's punishment decision off of the equilibrium path. The dashed red line shows B's investment decision. When expected punishment is low, B does not invest.

[10]Note that with the linear functional form, the solution is bang-bang, and the promisee never chooses $\iota = 10$. However, in a more general functional form, or in the case of a risk averse player, $\iota = 10$ would be supported in equilibrium

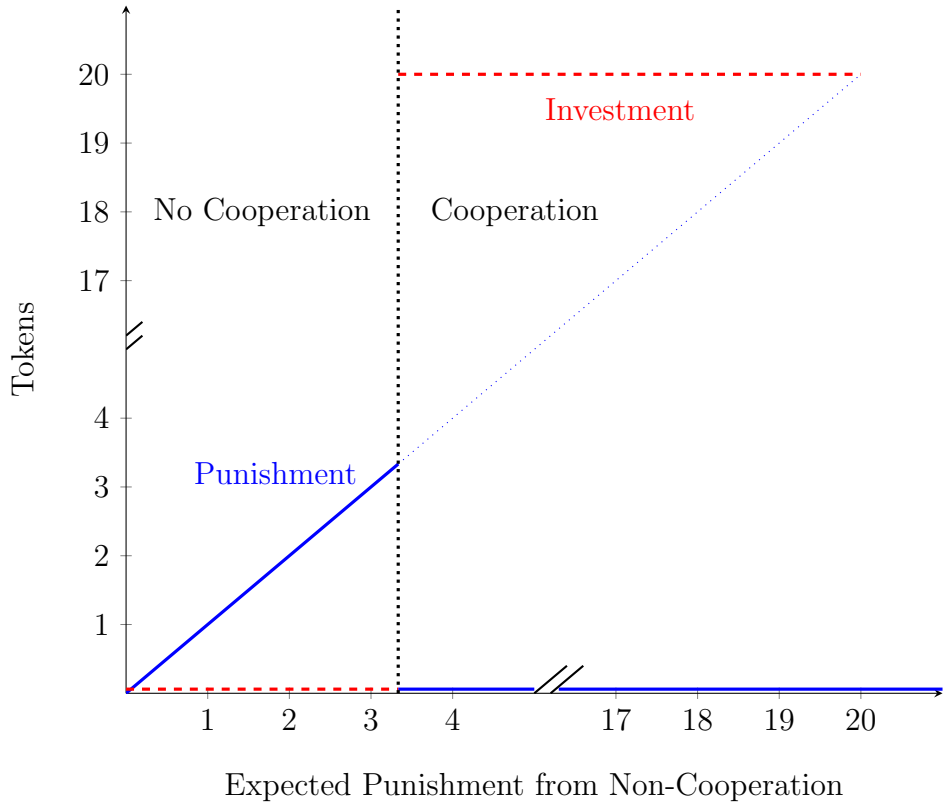


Figure 2: Equilibrium Behavior

However, when expected punishment is greater than $\frac{10}{3}$, B chooses an investment of 20.[11]

This model shows how a third-party observer's willingness to punish non-cooperation—particularly when a promise has been made and the promisee has relied—can deter a purely self-interested promisor from breaking that promise. In our main experiment (Section 3.1), we implement exactly the numerical example from Section 2 (where non-cooperation yields a 10-token advantage for the promisor, and each token of punishment imposed by the third party reduces the promisor's payoff by a factor of three) as an incentivized study. We present a variation on this model as a vignette study in Section 3.2.

3 Design and Hypotheses

In this section, we describe the games we used in our two experiments and the hypotheses we tested in the light of the model of the punisher's preferences that we developed in the

[11] In this setup, B never chooses an investment of 10, but if we introduced risk aversion, there would be a range of values for which B chooses an investment of 10.

previous section.

3.1 Main Experiment

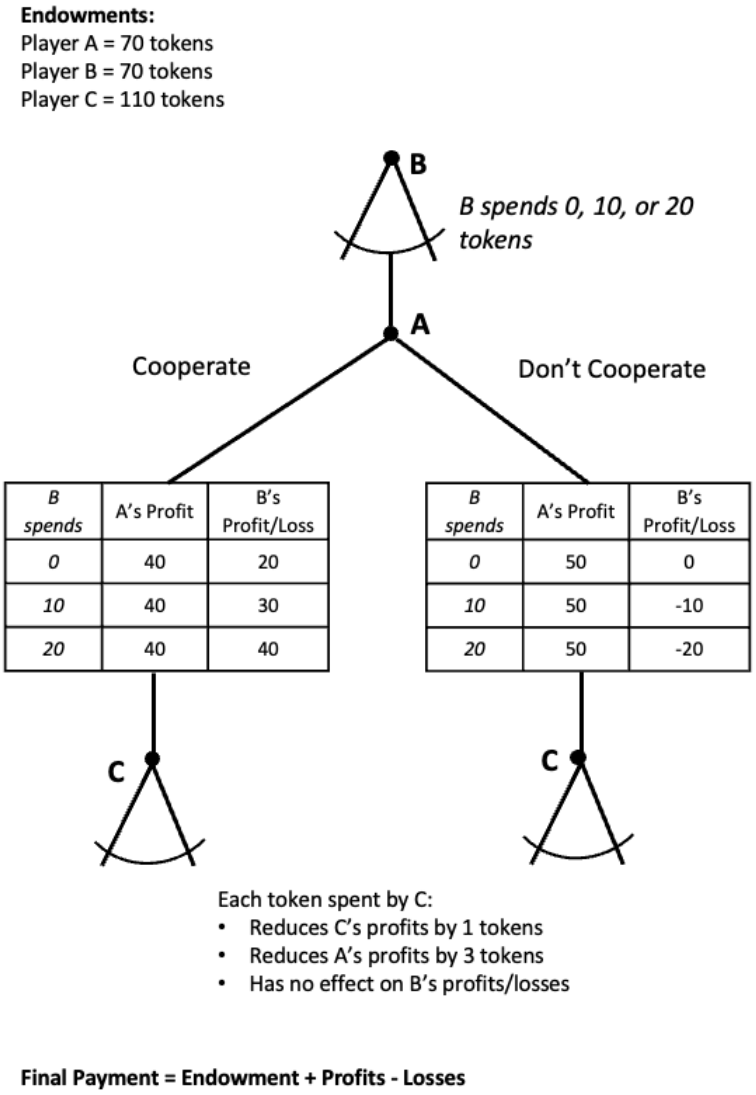


Figure 1: Game Witnessed by Third Party (Incentivized)

Our main experiment is based on the three-player game depicted in Figure 1. Each subject in a randomly selected triad of subjects are randomly assigned one of three roles: A, B, or C. The three subjects are then given endowments of tokens based on their roles: Players A and B were given endowments of 70 tokens, and Player C 110 tokens. Any profits or losses they made in the subsequent game are then added to or subtracted from these

endowments. Each token is worth 0.10 CHF for the purposes of determining subjects' final payoffs.[12]

The main game begins with B's decision to invest 0, 10, or 20 tokens. A then decides whether or not to cooperate with B. A's decision to cooperate increases B's payoff and the joint surplus but decreases A's payoff. B's payoff also depends on the amount B chose to invest: it is strictly increasing in this amount if A cooperated with B, and strictly decreasing if A did not. A's payoff is unaffected by B's investment.

The main game is preceded by a communication stage in which, following Vanberg (2008), the two players from the triad who were randomly selected to be A or B but have yet to learn which of A or B they have been selected to be, have the opportunity to communicate with one another. They are told whether they are A or B prior to playing the main game but only once the communication stage is completed. Behind this veil of ignorance, each can choose to make a promise to cooperate with the other ("I promise to choose Cooperate if I am chosen to be Player A") in the event that they turn out to be Player A. The veil gives them an incentive to make promises to one another if they believe that doing so will be perceived by the other as a friendly act and makes the other accordingly more inclined to cooperate with them in the event that they find themselves in the more vulnerable role of Player B.[13]

An advantage of structuring the players' communications in this way is that, unlike in a standard trust game (such as the game we use in our second experiment), the promisee does not explicitly rely by making a promise (though, of course, they may be hoping that their decision to be promise will be reciprocated in the next stage of the game as just described). Thus, we can isolate the pure effect of promising—that is, the effect of a promise that hasn't been relied upon—on the willingness of the third party to punish non-cooperation, what

[12]At the time of the experiment the exchange rate between Swiss Francs and US Dollars was 1 CHF = \$1.09.

[13]The main difference between our design of the promise-generating mechanism and Vanberg's (2008), is that Vanberg's subjects sent free-form messages rather than precoded messages. Our subjects could only send precoded messages. This has the disadvantage of being less natural. But it enabled us to keep control of the communications subjects made to one another and to avoid the need to code potentially ambiguous communications as involving promises or not after the fact.

Mischkowski, Stone and Stremitzer (2019) refer to as the "promise per se effect." In a standard trust game, by contrast, where the trustee is given the opportunity to promise to cooperate with the trustor at the outset of the game, a subsequent decision by the trustor to opt-in to the game constitutes reliance on the promise to cooperate (indeed, a likely motivation for the trustee to make such a promise is to induce this type of reliance), making it impossible to study the effect of promising when reliance is zero.[14]

If one of the players decided against making a promise during the communication stage, no message was conveyed to the other. Thus, when a subject made no promise, they didn't communicate with the other player at all (unlike in our second experiment, where non promises arise from statements of an intent to cooperate that explicitly disclaim any promise).[15]

The main experiment is followed by a punishment stage in which the third subject in the triad, Player C, who observes the interaction between A and B, but has no personal stake in it, has the opportunity to punish A by spending tokens to reduce A's payoff: for every token C spends punishing A, A's payoff is reduced by 3 tokens. Thus, the decision to punish does entail a personal cost to C.[16] A and B knew from the outset that A could be punished by C

[14]It is an interesting question whether the mechanism used to induce a promise matters. The literature has tended to assume that all promises are created equal—that no matter how a promise came into existence, it is a promise, and can be studied indiscriminately as such. But in reality promises can vary across multiple dimensions: the degree of the formality of the promise, the mutuality or unilaterality of the promise (was it part of an exchange of promises), the nature of the relationship between the parties exchanging the promise, the likelihood of future reciprocication either by return promise or performance, and so forth. Only a few recent papers have started to unpack the potential behavioral implications of such differences (Charness and Dufwenberg, 2008; Ederer and Stremitzer, 2018; Bartolomeo, Dufwenberg, and Papa, 2022 Bartolomeo, Dufwenberg, Papa, and Passarelli, 2023), casting doubt on the idea that all promises are created equal. We therefore regard it as a virtue of our two experiments that each employs a different promise-generating mechanism. To the extent that results are similar across the experiments, we show that our results are robust to a change in the promise-generating mechanism.

[15]A limitation of this design choice is that a subject only communicates with the other when they make a promise. Thus, it's possible that any observed effects of promising are produced by promising plus communication rather than promising alone. We designed the communication stage in this way to track the most natural way in which subjects don't make promises to one another–namely, by remaining silent. In our vignette experiment, by contrast, an A who doesn't promise does so expressly by explicitly disclaiming the promise. That design choice, though less natural, has the advantage of ensuring that A communicates with B regardless of whether A makes a promise to cooperate. We thought that the conjunction of results from the two experiments would be more convincing if we could confirm our hypotheses using two different constructions of the message space.

[16]C's choice was not presented as a choice to punish A. The prompt C saw was as follows: "You can now decide how much to spend on reducing A's payoff. Each token that you spend will reduce your payoff by 1 token and A's payoff by 3 tokens. If [randomly inserted option], how much would you like to spend reducing

in this way.[17]

Our focus is on C's decision during this punishment phase, and whether it is influenced by the same promissory norms that motivate many promisors to keep their promises. In line with the simple theoretical model of C's preferences described in the previous section, we make the following hypotheses about C's behavior:

Hypothesis 1 *C's willingness to punish non-cooperation will be higher when A made a promise than when A made no promise for all levels of B's investment (H1.1). C's willingness to punish non-cooperation will increase with B's investment level whether or not A made a promise. (H1.2) This increase will be greater if A made a promise than if A made no promise (H1.3).*

Given that Hypothesis 1 arises from our conjecture that C's willingness to punish non-cooperation is driven by shared norms, we also hypothesize that perceptions of appropriateness of A's actions of subjects assigned roles A or B will exhibit a parallel structure. Hypothesis 2 summarizes.

Hypothesis 2 *A's and B's perceptions of the inappropriateness of A's non-cooperation will be higher when A made a promise than when A made no promise for all levels of B's investment (H2.1). Their perceptions of the inappropriateness of A's non-cooperation will be increasing in the level of B's investment whether or not A made a promise (H2.2). This increase will be greater if A made a promise than if A made no promise (H2.3).*

3.2 Vignette Experiment

Our main experiment used real interactions and incentives and thus directly models and tests whether and how third parties will punish non-cooperation at a cost to themselves. Here

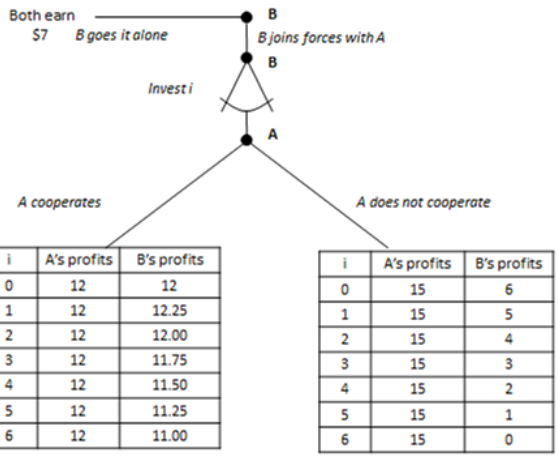
A's payoff?"

[17]We chose not to allow C to see whether B made a promise during the communication stage. This was not made explicit to A and B. Their beliefs about what C observes could influence their behavior by altering their expectations about the expected punishment. Given, however, that our focus is on C's punishment decision rather than the decisions of A and B, we thought it best to ensure that C's information set was independent of B's decision whether to make a promise or not during the communication stage. Otherwise, C might have adjusted his punishment in response to his perceptions of B's worthiness (given that those perceptions might have been influenced by whether B made a promise).

we report the results of a vignette study that we conducted prior to the main experiment. Although it didn't involve real interactions and incentives, we think the results contribute to our understanding of the dynamics of third-party enforcement of promissory norms. We demonstrate the robustness of our results insofar as they are consistent across experiments. In addition, the vignette study has some different features from our main experiment, allowing us to examine the robustness of our results across different contexts. First, while we tried to make our main experiment as abstract as possible to avoid potential demand effects, we supplied participants in the vignette study with more context and a more relatable scenario. Second, and relatedly, we used an easier to motivate trust game for the second experiment instead of generating promises using the more complex and obscure device of the veil of ignorance. Thus, the setup of the vignette experiment was arguably easier for participants to understand, which is particularly important when they are being asked hypothetically how they would act rather than actually making choices that might cost or benefit them as our subjects in the main experiment did. Finally, we used a different payoff structure that enabled us to study altruistic punishment in a context in which reliance by the promisee can be unproductive.

The trust game that we used for this study is depicted in Figure 2. B begins by choosing either to join forces with A in a cooperative project or to go it alone. In the cooperative venture, B selects an investment level, i. A then decides whether or not to cooperate. Joining forces with A presents risk for B, as it is a costly decision that A can exploit by choosing not to cooperate with B. But so long as A cooperates, both A and B do better when B joins forces with A instead of going it alone. B's payoff conditional on cooperation is tied to the investment level, increasing to start with diminishing returns beyond an optimal point. If A chooses not to cooperate, higher investment by B decreases B's payoff.

We add to the trust game an initial communication stage that occurs before B makes his initial decision whether to join forces with A and gives A the opportunity to promise B that he will cooperate if B joins forces with him. Specifically, A chooses whether to promise B to cooperate with B if B joins forces with them ("I promise that I will cooperate with you") or



i	A's profits	B's profits
0	12	12
1	12	12.25
2	12	12.00
3	12	11.75
4	12	11.50
5	12	11.25
6	12	11.00

i	A's profits	B's profits
0	15	6
1	15	5
2	15	4
3	15	3
4	15	2
5	15	1
6	15	0

Figure 2: Game Witnessed by Third Party (Vignette)

to tell B that they planned on cooperating with B without making any promises ("All I can say is that I plan to cooperate with you, though I can't promise that I will do so").

After the trust game has been played there is a punishment stage in which C chooses how much to punish A given what transpired between A and B. More specifically, we ask our subjects to imagine they are in the position of C having observed various different iterations of the game between A and B and to imagine how much they would punish A were that to involve them incurring a cost themselves. Given the possible messages sent during the communication stage, we are able to compare the effect of a promise by A on C's willingness to punish A with the effect of a statement of intention that explicitly disclaims any promise. Because communication between A and B occurs even when A makes no promise to B, we can isolate the effect of a promise on the willingness to punish from the effect of communication (with or without a promise).

In line with our theoretical model described in the previous section, we hypothesized that the third party's willingness to punish non-cooperation tracks that of the third party in our main experiment, as described in Hypothesis 1 above.

Unlike the game employed in our main experiment, the trust game allows for inefficient overinvestment, which enables us to study how unproductive reliance motivates the punishment of non-cooperation. There is a large contract theory and experimental literature on overinvestment. Of particular relevance here, Stone and Stremizter (2020) find experimen-

tal evidence that overinvestment makes promisors more likely to keep their promises. The proposed explanation is that breaking a relied-upon promise is perceived as morally worse than breaking a promise that hasn't been relied upon even when the reliance is unproductive and that promisees accordingly strategically overinvest to encourage promise keeping. If unproductive reliance motivates C to punish non-cooperation by A more, then B has a further reason to overinvest to enhance the likelihood of punishment and thus enhance A's self-interested reasons to cooperate.

4 Procedure

In this section, we describe the procedures employed in our two experiments. In our main experiment, subjects played the incentivized trust game, where the decisions of all subjects, including subjects in the role of third-party observer, had monetary consequences corresponding to the payoffs of the game. In the second experiment, we asked subjects to report the likelihood that they would punish, at a cost to themselves, non-cooperation by A in various iterations of this game. The scenarios they were reacting to were hypothetical and their responses had no effect on anyone's payoff including their own.

4.1 Main Experiment

In our main experiment, which we conducted in January 2022, groups of three subjects were anonymously matched with one another to play the game described in Section 3.1 and depicted in Figure 1. We elicited the decisions of subjects assigned to be A and B using the direct response method, while we used the strategy method to elicit the punishment decisions of subjects assigned to be C. Specifically, without knowing the actual decisions of A and B, C made punishment decisions for all possible choices that A might have made during the communication stage (promise or no promise) and all possible configurations of A's cooperation decision and B's investment decision. Thus, C made punishment decisions for six scenarios where A didn't cooperate, three where A made a promise for each possible investment level of B, and three where A didn't promise for each investment level, and then

six more scenarios where A cooperated. In this way, we generated within-subject data that showed each C's reactions to all relevant scenarios in the game. The punishment that was actually inflicted on A and the subjects' final payoffs were determined by the punishment that C chose for the scenario that corresponded to the actual decisions that A and B made.

To minimize order effects and generate between-subject data, the first of the non-cooperation scenarios C was shown was randomly selected and shown to subjects on a separate screen before they were asked to fill in their responses for the remaining scenarios on a second screen.[18] That first scenario was shown to subjects before they were informed that they would be asked to react to the five other possible scenarios.[19]

At the end of the experiment, all subjects who had been assigned to be A or B were asked to report their perceptions of the appropriateness of various actions. We elicited those perceptions by employing the method developed by Krupka and Weber (2013) to elicit subjects' beliefs about relevant shared norms (as opposed to their personal attitudes). We asked subjects to answer on a four-point scale (very inappropriate, moderately inappropriate, somewhat inappropriate, or appropriate) while also telling them that they would receive CHF 0.10 every time they selected the *modal* response given by subjects who had been assigned the same role in the game.[20] Thus, subjects maximized their monetary payoff by selecting

[18]Because of the large number of options, we split the table over two pages. Subjects were first asked to tell us how much they wanted to punish A in each scenario in which A didn't cooperate and then how much they wanted to punish A in each scenario in which A cooperated. Given that we did not expect many participants to impose punishment in cases where there was cooperation, we asked about punishment following cooperation second. It is possible that this choice introduced order effects. That is, a player's decision about how much to punish cooperation might be a function of the choices they made about how to punish non-cooperation. While this may have affected players' payoffs, it doesn't affect our main results because we are interested in (and our pre-registered hypotheses are about) punishment following non-cooperation not punishment following cooperation.

[19]Our aim was to simulate a direct response to the scenario that they were first randomly presented with. A concern about the strategy method is that subjects may experience it as artificial as they don't know what the others actually chose when reacting to it and therefore also don't know which of their responses will end up being implemented. Brandts and Charness (2011) survey studies that employ both the direct response method and the strategy method. Although they find that "there are significantly more studies that find no difference across elicitation methods than studies that find a difference" (p. 387), they also note that there is some evidence that emotions run higher when decisions are elicited under the direct-response method. Perhaps this is because those decisions feel more real to subjects. We find that order of presentation matters. Subjects' responses change depending on the scenario to which they were first exposed to. We discuss this in detail in Section 5.1.

[20]More specifically, for each scenario, each subject was asked to indicate whether she believed "that A's choice of "Don't cooperate" would be very inappropriate, moderately inappropriate, somewhat inappropriate,

the action that the greatest number of similarly situated subjects also selected rather than by offering their own views about the appropriateness of A's action. If a shared norm governing the appropriateness of A's actions exists, it should provide a focal point that (in the absence of any other focal point) subjects can use to solve this coordination game, in which case subjects' responses should track the structure of any such norm.

The experiment was programmed in oTree and conducted in January 2022 after obtaining IRB approval. We preregistered our hypotheses (https://osf.io/z2dm3) and recruited 914 subjects from the UAST subject pool run by ETH Zurich and the University of Zurich.[21]

Subjects were asked detailed control questions to determine whether they had understood the scenario, and they were not allowed to proceed until they had answered those questions correctly.[22] Before subjects began the experiment, they were informed that the task would take approximately 20 minutes and that they would earn money from the study. We did not announce an hourly wage.[23] Average earnings before any bonuses were CHF 10.42, ranging from a low of CHF 4.5 to a high of CHF 12. On average, subjects took 15.2 minutes to

or appropriate." She was then told that we would "determine which response was selected by the most participants assigned the role of Player A," and that if she gave "the response that is most frequently given by other people" she would "receive a bonus of \$0.10." The modal response was calculated for each group. So subjects in the role of Player A were rewarded for giving the modal response of others in the same role and subjects in the role of Player B were rewarded for giving the modal response of others in the role of Player B. The study was conducted in groups of 90, so the norm matching was conducted for each group of 90 separately so that they could be paid within 24 hours through bank transfer, as is the norm for the UAST. While there is an active debate on whether including a "neutral" option in questionnaires (see e.g. Nadler, Weston and Voyles (2015)), we chose a four-point scale rather than a five-point scale because we were concerned that the neutral option would provide too strong of a focal point for the participants to coordinate on.

[21]We aimed to run the experiment on 900 participants (50 per cell). However, because of subjects not showing up and dropping out, the lab ran the study iteratively in order to recruit 900. We ended up with data on 914 subjects: 304 subjects in role C, 305 in role A, and 305 in role B. The numbers are uneven because subjects can drop out. 40 subjects timed out during the experiment. Our preregistration inadvertently included an inconsistency. In item 12, we noted that we would "include all participants in the analysis" but in item 22 we wrote that "exclude outliers that outside of 3 standard deviations of the mean". This second item was a mistake. We do not exclude any data in our analysis, as doing so does not make sense for the statistical tests that we conduct.

[22]At the end of the experiment subjects were asked to leave comments for the researchers. 210 subjects left comments of some form. Of those, 29 participants reported that something about the experiment was unclear to them (most of these indicated that they were unclear about whether there would be more than one round). More subjects used their free response to indicated that the instructions were clear, with a particular emphasis on the diagrams.

[23]Most studies at UAST do not announce a wage or expected payments. However, participants know to expect a wage between CHF 20-30 per hour.

complete the experiment so that the effective hourly wage was CHF 41.13.

Subjects' earnings depended on the total number of tokens they had at the end of the game. Each subject started with an endowment of tokens: 70 if assigned to be A; 70 if assigned to be B; and 110 if assigned to be C. The total number of tokens increased or decreased depending on the outcome of the game. One token was worth CHF 0.10.

The game tree and a link to the full instructions was included at the top of every slide that subjects saw. We reproduce screenshots from the experiment (without repeating the game tree on each slide) in Appendix D.

4.2 Vignette Experiment

In our vignette experiment, which was conducted in October 2016, subjects were presented with the game described in Section 3.2 and instructed to imagine that, like Player C, they had observed as third parties six iterations of the trust game in which B joined forces with A but A subsequently decided not to cooperate with B. There were three "Promise conditions" and three "No Promise conditions", each one characterized by a particular investment level chosen by B: 0, 3, and 6.[24] Following the presentation of each scenario, subjects were asked to report the likelihood that they would choose to inflict a punishment on A for not cooperating with B at some small but not insignificant cost to themselves.[25] Subjects responded on a five-point scale.[26]

Subjects saw six scenarios in a randomized sequence.[27] This allowed us to generate

[24]The game tree lists investment levels of 0 through 6, but we only ask about these three levels. We presented the vignette's full investment structure, including the initial increase from 0 to 1, to establish context for participants, demonstrating the baseline rationality of investment. This contextual information allows potential punishers to assess promise-breaking across investment levels, even when the optimal level was not chosen.

[25]We told subjects that inflicting punishment on A might consist of "boycotting A's business which results in a monetary loss for A." We also told them that "[i]nflicting this punishment, however, is not costless to you" as "boycotting a business that you have been buying from forces you to switch to another product requiring you to change your habits, pay more for the other product, consume a product that you like less, etc...)."

[26]Across all of the within-subject responses, 17.5% of subjects reported that they were "very unlikely" to punish, 26.4% reported "unlikely", 13.4% reported "undecided", 35.9% reported "likely", and 16.8% reported "very likely."

[27]The order was not completely random, though the first scenario each subject saw was. Subjects either saw all three promise conditions first in a randomized order and then all three no promise conditions in a

both between- and within-subject data. Subjects' responses in their entirety constitute within-subject data, while their responses to the first condition they saw constitute our between-subject data.

We programmed the vignettes using Qualtrics and recruited 1200 subjects from Amazon Mechanical Turk's pool of MTurk workers who had a HIT approval rate of 95% or greater.[28] We determined our sample size using a simulation based on pilot data on 300 subjects with the goal of 80% statistical power (rejecting the null with 80% or greater probability if the hypothesized effect exists).[29]

Subjects were asked control questions to ascertain whether they had read and understood the scenario, and they were not allowed to proceed until they answered those questions correctly. At the end of the survey subjects were asked several other questions to assess how carefully and honestly they responded to the questions. We also elicited subjects' demographic characteristics.

Before subjects were presented with the scenarios, they were informed that they would be paid $1.50 for participating and that the task would take approximately 10 to 15 minutes. The announced hourly wage was therefore $6 to $9 per hour. Thus, on average subjects could expect to receive more than the federal minimum wage at the time ($7.25 per hour) and expected payments were much higher than the wages MTurk workers typically earn.[30] On average our subjects took 6.8 minutes to complete the task and so the effective average hourly wage was $13.24. We reproduce screenshots from the experiment in Appendix C.[31]

randomized order or vice versa.

[28] A HIT, or Human Intelligence Task, is a task assigned to a worker on Mechanical Turk. To be paid, the person assigning the work must approve the task. The 95% HIT rate indicates that these workers have had their work approved at least 95% of the time, which we use as a proxy for worker quality.

[29] We excluded the pilot data from our subsequent analysis. However, including the data does not change the statistical significance of any of our results.

[30] Studies have found a median hourly wage of $1.38 (Horton & Chilton, 2010) and a typical payment of $0.01-$0.10 per HIT (Mason & Watts, 2010).

[31] We did not pre-register this experiment.

5 Results

In this section, we describe in detail the results from both experiments. When testing our hypotheses on the within-subject data we use the Wilcoxon sign-rank test. When testing our hypotheses on the between-subject data we primarily use the Wilcoxon ranksum test. However, the Wilcoxon ranksum test is unavailable for testing H1.3 and H2.3.[32] We therefore employ a bootstrapping procedure for a non-parametric statistical test of H1.3 and H2.3 described in Appendix B (Efron and Tibsharani, 1993). We report results from linear regression analysis in Appendix A.

5.1 Main Experiment

Table 2 and Figure 4 summarize the average punishment by treatment condition in experiment 1 for both within-subject and between-subject punishment data following non-cooperation. Approximately 70 percent of subjects assigned to role C inflict a positive amount of punishment on A for non-cooperation in at least one condition.[33]

Table 1: Mean Punishment Imposed Following Non-Cooperation

	Observed Investments in the Relationship		
	0	10	20
Between-Subject			
No Promise	2.14 (N=50)	4.89 (N=49)	5.04 (N=48)
Promise	3.83 (N=52)	3.45 (N=53)	4.5 (N=52)
Within-Subject (N=304)			
No Promise	1.96	3.08	4.03
Promise	3.61	5.22	6.95

[32]This is because testing for an interaction effect requires us to compare the difference in the reported willingness to punish or amount of punishment (or perceptions of appropriateness) for different reliance levels in the Promise and the No Promise conditions. The Wilcoxon ranksum test allows us to test for differences between unmatched data but not differences in differences.

[33]Approximately 77% of subjects assigned to role A make a promise, 90% of those who make a promise subsequently cooperate, and 49% of those who do not make a promise subsequently cooperate. Of subjects assigned to role B 13% invest zero, 32% invest 10, 11% invest 2, and 55% invest 20.

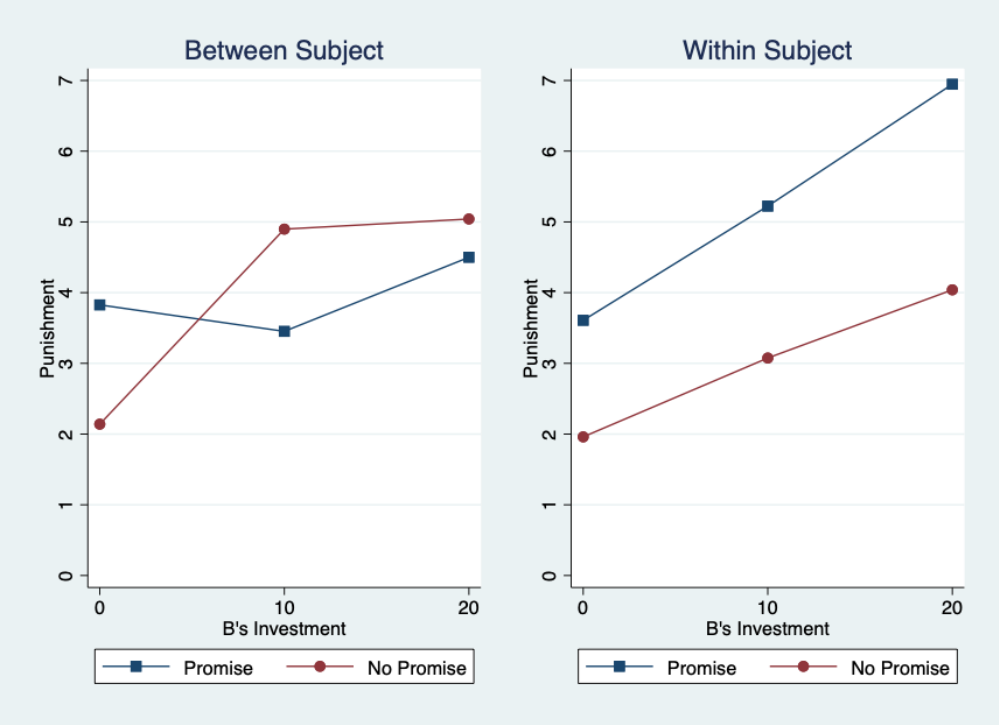


Figure 4: Average Reported Likelihood of Punishment Following Non-Cooperation Across Conditions (Incentivized)

5.1.1 Third Party: Within-Subject Data

Descriptively the within-subject punishment data support our hypotheses. The Promise line is above the No Promise line, suggesting that subjects punish non-cooperation more harshly when a promise was made regardless of the investment level (H1.1). Both lines are increasing as investment increases from 0 to 10, and 10 to 20, and from 0 to 20 suggesting that higher investment increases their willingness to punish non-cooperation (H1.2). Finally, the Promise line is rising more steeply over each interval and the entire range when there was a Promise in support of the interaction effect (H1.3).[34]

Tests on the within-subject data confirm the promise per se effect (H1.1) for three invest-

[34]Subjects assigned to be C punish A much less in scenarios where A cooperated, as we would expect. Still, some subjects punish A even when A cooperated. 28% inflicted some punishment in at least one scenario where A cooperated compared to 70% when A did not cooperate. In the case where A makes a promise and cooperates, 12.8% of participants inflict punishment in one of the conditions. Summing over the six possible scenarios, the average subject punishes 0.76 when A cooperated compared to 4.14 when A didn't cooperate. Descriptively the average punishment increases very slightly as investment increases consistent with a reliance per se effect. Moreover, for each level of investment, subjects punish A for cooperation *more* if A did not promise than if A did promise. However, we can think of this pattern as suggesting that Player A gets punished for not making a promise.

ment levels suggesting that subjects are willing to inflict a higher punishment on A when he made a promise regardless of the investment level.[35] Tests also confirm the reliance per se effect (H1.2) for both promises and no promises made for all investment intervals.[36] Finally, the interaction effect (H1.3) is confirmed for all investment intervals.[37]. Linear regressions reported in Appendix A support all three hypotheses for the within-subject data.

5.1.2 Third Party: Between-Subject Data

Descriptively the between-subject data seems to support the existence of a promising per see effect for zero investment but not for investment of 10 and 20 (H1.1). However, none of these differences are statistically significant (Wilcoxon ranksum yields, $p = .12$, $p = .43$, $p = .75$). There seems to be some support for the reliance per se effect but the effect is only significant for the (0 10) and (0 20) intervals in the no promise condition.[38]. Appendix B describes the bootstrapping procedure that we performed, with no evidence supporting the interaction effect (H1.3). Linear regressions reported in Appendix A provide weak support for the reliance effect (H1.2), but no support for the promising per se effect (H1.1) or the interaction effect (H1.3) for the between-subject data.[39]

We thus conducted an exploratory analysis (which we did not preregister) by breaking down the within-subject data into six between-subject groups, each consisting of the within-subject data generated by those who saw a particular scenario first. Figure 5 shows that the decisions of each group in isolation largely conform to our hypotheses. The lines depicting punishment when there was a promise are all above the no-promise lines consistent with the promise per se effect. All lines are increasing consistent with the reliance per se effect. And in most (though not all) of the graphs the promise lines are steeper than the no-promise lines consistent with the interaction effect. This suggests that the departures of our between-subject results from our hypotheses may have been driven by anchoring effects

[35]Wilcoxon signrank yields $p = .0000$ for all three tests.
[36]Wilcoxon signrank yields $p = .0000$ for all three tests.
[37]Wilcoxon signrank yields $p = .002$, $p = .0000$, $p = .0000$
[38]Wilcoxon ranksum, No Promise: (0 10, $p < .01$), (10 20, $p = .96$), (0 20, $p = .01$); Promise: (0 10, $p = .84$), (10 20, $p = .63$), (0 20, $p = .38$)
[39]Restricting attention to participants that impose positive punishment yields qualitatively similar results.

arising from the order of presentation. That is, subjects initial punishment decisions are somewhat arbitrary, but their subsequent choices are less arbitrary in relative terms, and accordingly they behave consistently with our hypotheses relative to the (more arbitrary) anchor that they set by their first choice.

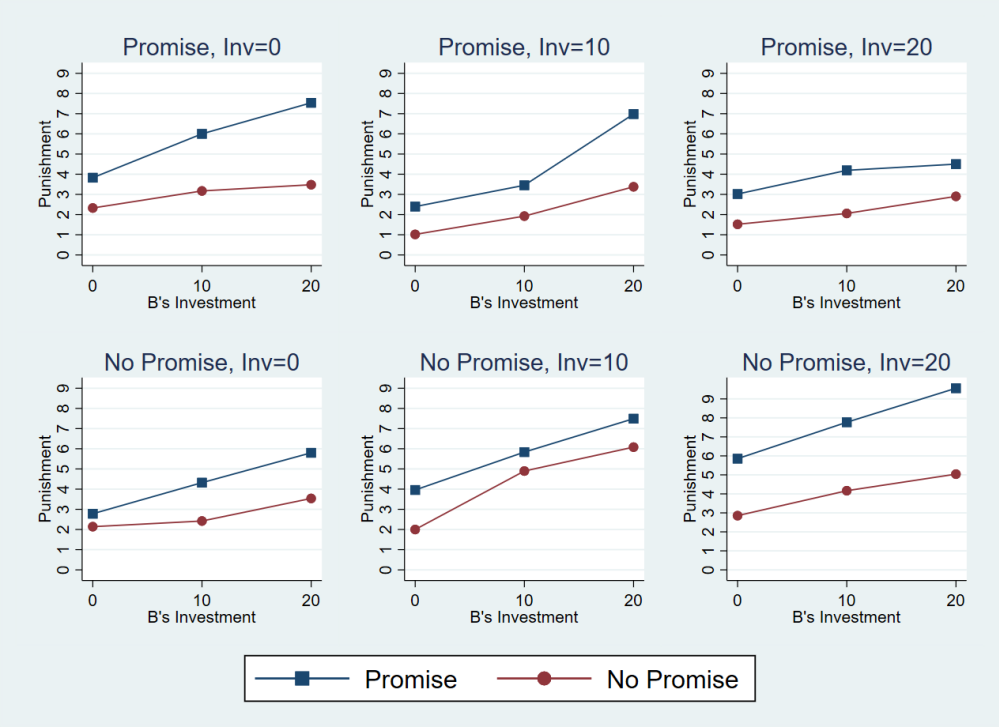


Figure 5: Within-subject Third-Party Punishment Decisions Following Non-Cooperation Broken Down By between-subject Group (Incentivized)

Figures 6 illustrate perceptions of social appropriateness reported by subjects in roles A and B respectively. Unfortunately, there was a technical glitch with the randomization of B's inappropriateness decision. All subjects were assigned to the same treatment, so we don't have any between-subject data for B.[40]

5.1.3 Perceptions of Promissory Norms

We next consider players' perceptions of promissory norms. The perceptions are largely consistent across those cast in the roles of A and B. This is consistent with findings of Erkut, Nosenzo and Sefton (2015) that elicited norms in dictator games do not vary between

[40] Also see Tables 3 and 4 in Appendix A.

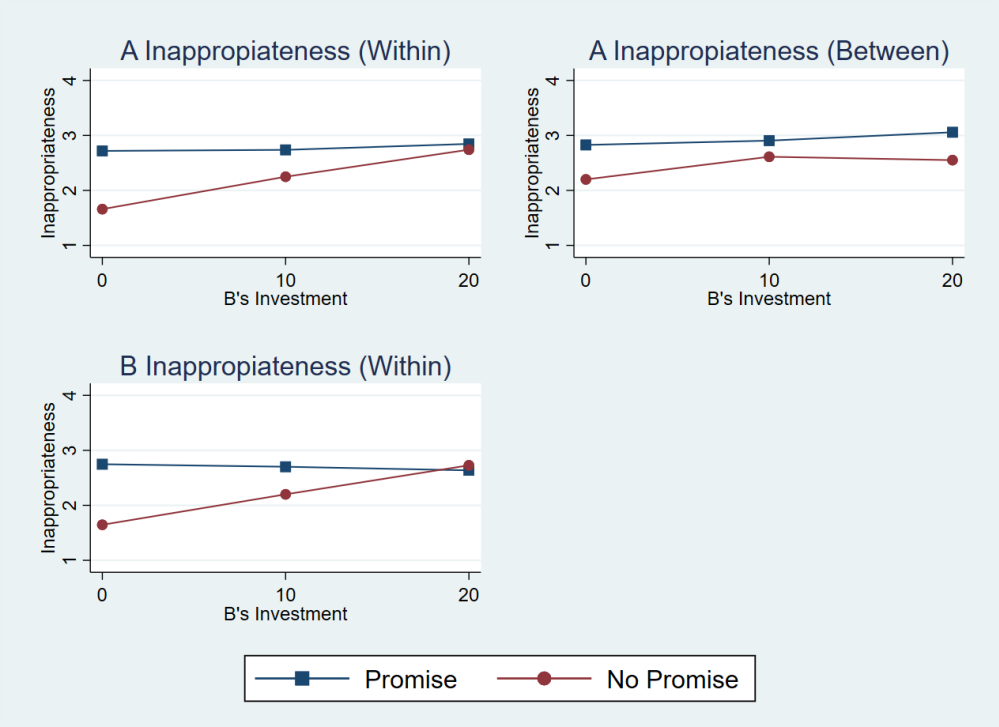


Figure 6: Player A's and B's Perceptions of Inappropriateness Across Conditions (Incentivized)

dictators, recipients, and third parties.

Descriptively, these data exhibit similar patterns to the punishment data consistent with our Hypothesis 2 that propensities to punish are driven by the same underlying promissory norms. We largely find support for the promising per se effect (H2.1).[41] We also find support for the reliance per se effect (H2.2) in the no promise conditions.[42] As for the interaction effect, we find support for the opposite result, which we might term a "reverse interaction effect."[43]

[41]The promise per se effect (H2.1) is significant at the 1% level for the within-subject appropriateness data elicited from subjects in role A for investment levels 0 and 10, and significant at the 10% level for an investment level of 20. For the between-subject data, the promise per se effect is significant at the 1% level for investment levels 0 and 20 and significant at the 10% level for an investment level of 10. For role B, the within-subject data is significant at investment levels 0 and 10. $p < 0.01$ in both cases. As mentioned, a coding error meant that all subjects assigned to the role of B were exposed to no promise and an investment of 10, so there are no between-subject results.

[42]We don't find it in the promise conditions, but the purest form of the reliance per se effect is seen, because we see it in the no promise conditions which isolate the effects of reliance from the effects of the promise. The absence of an effect for the promise conditions may be the result of a ceiling effect discussed above.

[43]The "reverse interaction effect" is significant at $p < 0.01$ between all levels for both role A and role B.

As much as these results seem mixed as measured against Hypothesis 2, they are consistent with the patterns found in Di Bartolomeo et al. (2023), which also exhibit a reverse interaction effect. Both those findings and ours likely resulted from a ceiling effect. Recall that we asked subjects to choose between 4 levels in judging the appropriateness of a decision not to cooperate (appropriate, somewhat inappropriate, moderately inappropriate, very inappropriate). Our subjects assigned a relatively high level of inappropriateness to the actions of a player who breaks her promise even if the recipient had not relied on it. Consequently, there is not much room to express harsher disapproval as reliance increases. By contrast, subjects' lower assignments of disapproval to noncooperation by a player who made no promise leave more room for upward adjustment of the inappropriateness rating as reliance increases. Similarly, Di Bartolomeo, Dufwenberg and Papa (2023) ask subjects to choose between keeping or breaking a promise. If subjects think that promises generate sufficient reasons for them to keep them, then the impact of higher expectations on their willingness to keep their promises will be obscured because they cannot do more than perform. The data will then show a reverse interaction effect even if promises interact positively with expectations in motivating subjects to perform.

5.2 Vignette Experiment

Table 1 and Figure 3 summarize the mean reported likelihood of punishment by treatment condition in experiment 2 for both our within-subject and our between-subject punishment data.[44] Descriptively, both sets of data are largely in line with our hypotheses. In line with hypothesis H1.1, subjects reported a higher average likelihood of punishing A when they

[44]We excluded the responses of the 126 subjects who reported in the post-experiment survey that they did not or only "kind of" understood how A's and B's actions affected their payoffs. We also excluded the responses of four subjects who reported that they had done the study before. We have no good explanation as to why four subjects reported having taken the survey before. We provided links to participants which were only good for a single log in. We implemented filters preventing subjects (as identified by their MTurk IDs) from participating who had participated in pilots of our experiment or similar experiments we had run in the past. So the only explanation for the four self-reported repeat takers could be that subjects have multiple MTurk IDs or mistakenly checked the wrong box. Including these data does not, however, change the results. Similarly, results do not qualitatively change when we exclude subjects who report that they did not "take the scenario seriously," that they did not "carefully read the instructions," and that they chose their answers to make themselves "seem like a good person."

were told that A made a promise to cooperate with B. Thus, the Promise lines in Figure 3 lie above the No Promise lines. In line with hypothesis H1.2, subjects also reported a higher average likelihood of punishing A the greater was B's investment level in both the Promise and No Promise conditions (with the exception of the (3-6) interval in the No Promise conditions in the between-subject data). Thus, both lines in Figure 3 are upward sloping (at least over the entire range). Finally, in line with hypothesis H1.3, the increase in the average reported likelihood of punishing A caused by higher investment is larger when subjects were told that A made a promise to cooperate with B. Thus, the Promise lines in Figure 3 are steeper than the No Promise lines over each interval and over the entire range.

Table 2: Mean Reported Willingess to Punish

	Observed Investments in the Relationship		
	0	3	6
Between-Subject			
No Promise	2.47 (N=173)	2.90 (N=183)	2.84 (N=183)
Promise	2.85 (N=156)	3.73 (N=194)	3.89 (N=186)
Within-Subject (N=1075)			
No Promise	1.92	2.55	2.76
Promise	2.75	3.81	4.11

Tests on the within-subject data yield support for all of our hypotheses.[45] Tests on the between-subject data find support for a promise per se effect (H1.1) for each level of investment.[46] Subjects report that they are more willing to punish non-cooperation when it breaks a promise to cooperate. Tests on the between-subject data also suggest that reliance matters to subjects' reported willingness to punish if there was a promise,[47] and for the (0-3) and (0-6) investment intervals if there was no promise.[48] Thus, the between-subject data support the reliance per se effect (H1.2) for the most part. Finally, we perform a

[45]We generally obtain significance at the 1% level. Only under a very restrictive definition of our data pool (eliminating all observations that might be invalid in the light of our post-experiment survey) do we find $p = 0.02$.

[46]Each is significant at the 1% level.

[47]The effect is significant at the 1% level for the (0-3) and (0-6) intervals, and at the 5% level ($p = 0.02$) for the (3-6) interval.

[48]The effect is significant at the 1% level. For the (3-6) interval, mean effort goes down, contrary to H1.2, though the effect is not significant at any level (p=0.63).

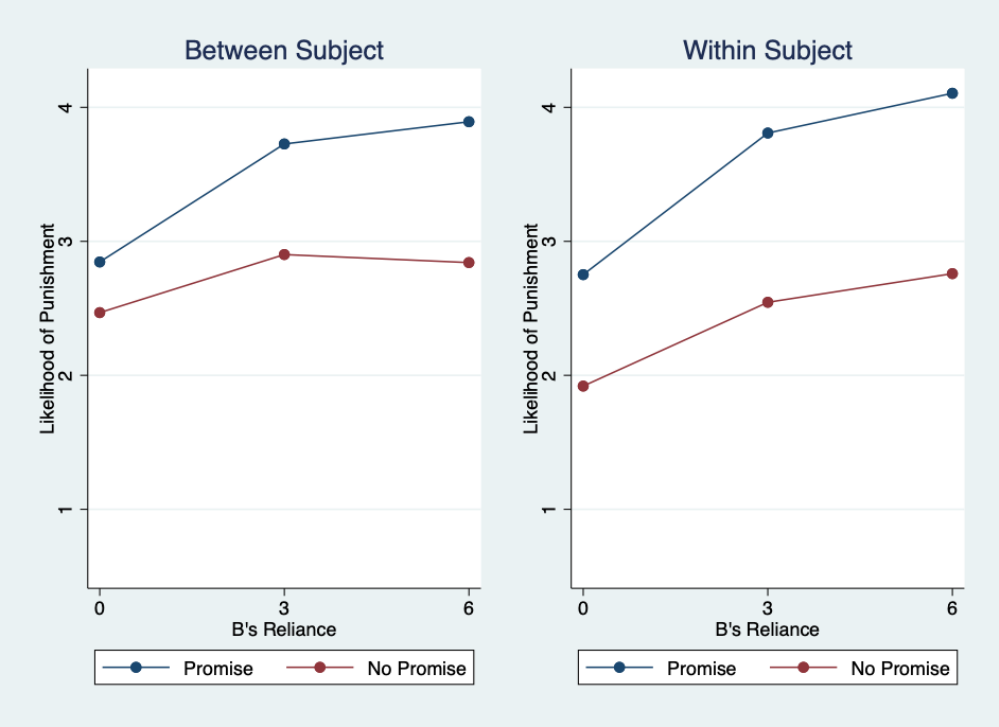


Figure 3: Average Reported Likelihood of Punishment Across Conditions (Vignette)

bootstrapping procedure to see whether the between-subject data exhibit an interaction effect (H1.3). This yields significance at the 1% level for the (0-6) interval.[49] This suggests that increased reliance has a greater effect on the reported willingness of third parties to punish non-cooperation when non-cooperation breaks a promise to cooperate.

6 Conclusion

We set out in this paper to investigate whether the moral motivations that lead many promisors to keep their promises in the absence of formal sanctions also shape the willingness of disinterested third parties to punish promise-breakers at a cost to themselves. Drawing on two experiments—a fully incentivized study and a vignette study—our findings indeed suggest that promise-breaking triggers third party punishment in line with the moral motivations that lead individuals to keep their promises. We find evidence that third parties are more willing to punish non-cooperative behavior of one who promised than one

[49]The bootstrapping procedure simulates synthetic samples and is described in detail in Appendix B.

who did not make such a promise (the promise per se effect). We further observe that reliance by the promisee increases the third party's willingness to punish non-cooperation (the reliance effect), and that combining a promise with reliance triggers still more punishment (the interaction effect). This pattern mirrors earlier experimental work on first-party promise keeping and is consistent with the idea that the same promissory norm underpins both individuals' internal motivations to keep promises and bystanders' propensities to punish promise-breakers.

Despite their novelty, our results are subject to important caveats. First, while the results of the vignette experiment are clear, the results of the incentivized experiment differ based on the unit of analysis. When we look at the within-subject data from the incentivized experiment, we find strong support for all three hypothesized effects. However, when we look at the between-subject data, we do not find the same effects. Indeed, further analysis shows an anchoring effect where third-party punishment decisions seem partly driven by whichever scenario subjects happen to see first. This discrepancy highlights the well-known challenge of choosing among within-subject and between-subject experimental designs.

Second, we note that our vignette study was not incentivized; while the instructions stressed that punishment would come at some personal cost, participants faced no actual reduction in earnings. Vignette studies cannot fully substitute for real stakes in identifying how people trade off their moral convictions against monetary payoffs. Yet focusing on a purely hypothetical scenario allowed us to reduce extraneous complexity and helps demonstrate how third-party punishment might operate in a more contextualized setting. We view the similarity of the core findings in the vignette study and the within-subject incentivized study as evidence that shared promissory norms may be robust to differences in context.

Third, like most laboratory experiments, ours remains limited in external generalizability. We used subject pools comprised of university students (for the incentivized experiment) and Amazon Mechanical Turk workers (for the vignette experiment), and we collected data in a controlled environment. Of course, many real-world relationships evolve over longer time periods and may feature repeated and more complex interactions. However, the fact

that complete strangers were willing to sacrifice their own monetary payoffs in order to altruistically punish promise-breakers indicates the existence of a behavioral effect even if it does not allow us to determine its real-world magnitude.

With these limitations in mind, we see two main contributions of our paper. First, our results reinforce an important lesson from the literature on promise-keeping: promises themselves matter. We add to that lesson by showing that a similar structure governs the willingness of a third party to punish promise breakers. Second, our findings suggest that purely decentralized, norm-based enforcement —as embodied by bystanders who bear a personal cost to punish—has the same form that determines whether a promise maker will or will not keep a promise. Although how much these moral sentiments matter in real-world markets will depend on many institutional details, our experiments highlight that the norms of promissory commitment and reliance sensitivity can extend beyond two-party relationships.

Combining the moral motivations to keep promises with a willingness by impartial observers to punish promise-breakers at their own expense paints a more complete picture of how decentralized social sanctions can support contractual cooperation. The possibility of third-party enforcement through moral or reputational levers may help promisees feel secure enough to invest, *even if* a promisor's self-interest might otherwise tempt her to break the promise, and even if a promisee believes that a particular promisor is devoid of any intrinsic motivation to keep her promises. Going forward, we hope that our results will inspire further research on the links between relational contracting, altruistic punishment, and voluntary promise keeping.

References

Bellemare, Charles, Alexander Sebald, and Martin Strobel. 2011. "Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models." *Journal of Applied Econometrics*, 26: 437–453.

Bernstein, Lisa. 1993. "Social norms and default rules analysis." *S. Cal. Interdisc. LJ*, 3: 59.

Bernstein, Lisa. 1996. "Merchant law in a merchant court: Rethinking the code's search for immanent business norms." *University of Pennsylvania Law Review*, 144: 1765–1821.

Brandts, J.;, and Gary Charness. 2011. "The strategy versus the direct-response method: a first survey of experimental comparisons." *Experimental Economics*, 14(3)): 375–398.

Charness, Gary, and Martin Dufwenberg. 2006. "Promises and Partnership." *Econometrica*, 74: 1579–1601.

Charness, Gary, and Martin Dufwenberg. 2010. "Bare Promises: An Experiment." *Economics Letters*, 107: 281–83.

Chaudhuri, Ananish. 2011. "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature." *Experimental economics*, 14: 47–83.

Di Bartolomeo, Giovanni, Martin Dufwenberg, and Stefano Papa. 2023. "Promises and partner-switch." *Journal of the Economic Science Association*, 9: 77–89.

Di Bartolomeo, Giovanni, Martin Dufwenberg, Stefano Papa, and Francesco Passarelli. 2023. "Promises or agreements? Moral commitments in bilateral communication." *Economics Letters*, 222: 110931.

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

Dufwenberg, Martin, and Uri Gneezy. 2000. "Measuring Beliefs in and Experimental Lost Wallet Game." *Games and Economics Behavior*, 30: 163–182.

Ederer, Florian, and Alexander Stremitzer. 2017. "Promises and Expectations." *Games and Economic Behavior*, 106: 161–178.

Ellingsen, Tore, and Magnus Johannesson. 2004. "Promises, Threats and Fairness." *Economic Journal*, 114: 397–420.

Erkut, Hande, Daniele Nosenzo, and Martin Sefton. 2015. "Identifying social norms using coordination games: Spectators vs. stakeholders." *Economics Letters*, 130: 28–31.

Fehr, Ernst, and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and human behavior*, 25: 63–87.

Gintis, Herbert, Samuel Bowles, Robert Boyd, and Ernst Fehr. 2005. "1 Moral Sentiments and Material Interests: Origins, Evidence, and Consequences." *Moral sentiments and material interests*, 1.

Guerra, Gerardo, and Daniel John Zizzo. 2004. "Trust responsiveness and beliefs." *Journal of Economic Behavior & Organization*, 55: 25–30.

Hart, Oliver, and John Moore. 2008. "Contracts as Reference Points." *The Quarterly Journal of Eeconomics*, 123: 1–48.

Horton, John J., and Lydia B. Chilton. 2010. "The Labor Economics of Paid Crowdsourcing." *EC '10*, 209–218. New York, NY, USA: ACM.

Krupka, Erin L., and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association*, 11: 495–524.

Krupka, Erin L, Stephen Leider, and Ming Jiang. 2017. "A meeting of the minds: informal agreements and social norms." *Management Science*, 63: 1708–1729.

Macaulay, Stewart. 1963. "Non-Contractual Relations in Business: A Preliminary Study." *American Sociological Review*, 55–67.

Mason, Winter, and Duncan J. Watts. 2010. "Financial Incentives and the Performance of Crowds." *ACM SigKDD Explorations Newsletter*, 11: 100–108.

Mischkowski, Dorothee, Rebecca Stone, and Alexander Stremitzer. 2016. "Promises, Expectations, and Social Cooperation." *Harvard Law School John M. Olin Center Discussion Paper No. 887*.

Mischkowski, Dorothee, Rebecca Stone, and Alexander Stremitzer. 2019. "Promises, expectations, and social cooperation." *The Journal of Law and Economics*, 62: 687–712.

Nadler, Joel T, Rebecca Weston, and Elora C Voyles. 2015. "Stuck in the middle: the use and interpretation of mid-points in items on questionnaires." *The Journal of general psychology*, 142: 71–89.

Regner, Tobias, and Nicole S Harth. 2014. "Testing belief-dependent models." *Max-Planck Institute of Economics Jena Working Paper*.

Scott, Robert E. 1990. "A relational theory of default rules for commercial contracts." *The Journal of Legal Studies*, 19: 597–616.

Scott, Robert E. 2003. "A theory of self-enforcing indefinite agreements." *Colum. L. Rev.*, 103: 1641.

Selten, R. 1967. "Die Strategiemethode zur Erforschung des eingeschraenkt rationalen Verhaltens im Rahmen eines Oligopolexperiments." 136–168. in *Beitraege zur experimentellen Wirtschaftsforschung*, edited by H. Sauermann., Tuebingen: Mohr.

Shiffrin, Seana Valentine. 2007. "The Divergence of Contract and Promise, 120 Harv." *L. Rev*, 708: 710.

Stone, Rebecca, and Alexander Stremitzer. 2020. "Promises, Reliance, and Psychological Lock-In." *The Journal of Legal Studies*, 49: 33–72.

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

APPENDIX A: Tables for the Incentivized Eperiment

Table 3: Perceptions of Inappropriateness (Player A)

	Observed Investments in the Relationship		
	0	10	20
Between-Subject			
No Promise	2.2 (N=50)	2.61 (N=49)	2.55 (N=49)
Promise	2.83 (N=52)	2.91 (N=53)	3.06 (N=52)
Within-Subject (N=305)			
No Promise	1.66	2.25	2.74
Promise	2.72	2.74	2.85

Perceptions of inappropriateness of non-cooperative behavior by Player A as determined by those in the role of Player A. Inappropriateness was asked on a four point scale: appropriate (1), somewhat inappropriate (2), moderately inappropriate (3), and very inappropriate (4).

Table 4: Perceptions of Inappropriateness (Player B)

	Observed Investments in the Relationship		
	0	10	20
Between-Subject			
No Promise		2.2 (N=305)	
Promise			
Within-Subject (N=305)			
No Promise	1.65	2.2	2.73
Promise	2.75	2.70	2.64

Perceptions of inappropriateness of non-cooperative behavior by Player A as determined by those in the role of Player B. Inappropriateness was asked on a four point scale: appropriate (1), somewhat inappropriate (2), moderately inappropriate (3), and very inappropriate (4). Note that due to a coding error in the experiment, all subjects were assigned to the same treatment, so we do not have between-subject data.

APPENDIX B: BOOTSTRAP

Experiment 1

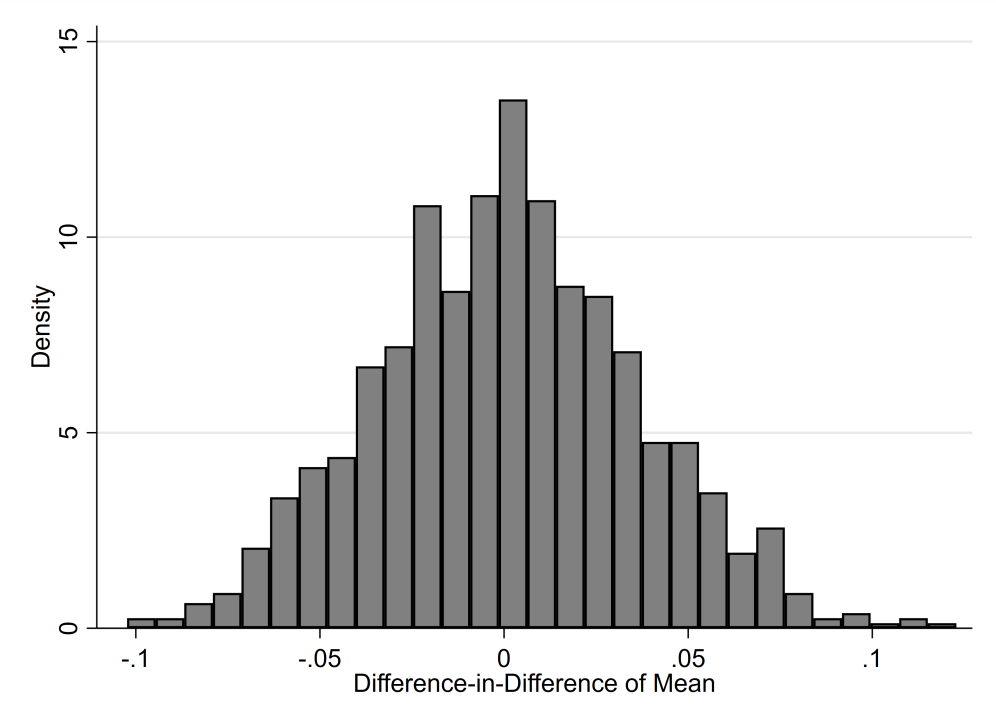


Figure 7: Simulated Distributions for the Experiment 1 (0-6 Gap)

Table 5: Linear Regression of Player C' Punishment

	within-subject	within-subject	between-subject	between-subject
Promise	2.235***	1.605***	-0.095	1.020
	(0.27)	(0.26)	(0.69)	(0.97)
Investment	0.135***	0.104***	0.088*	0.146*
	(0.01)	(0.01)	(0.04)	(0.06)
Interaction		0.063***		-0.112
		(0.02)		(0.08)
Constant	1.671***	1.986***	3.137***	2.567***
	(0.27)	(0.26)	(0.58)	(0.62)
R^2	0.054	0.056	0.014	0.020
N	1824	1824	304	304

Note: Each specification is a linear regression where the outcome variable is the punishment choice of C. Standard errors in parentheses are clustered at the level of the individual. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 6: Linear Regression of Inappropriateness (Player A)

	within-subject	within-subject	between-subject	between-subject
Promise	0.551***	1.028***	0.476***	0.537**
	(0.06)	(0.07)	(0.11)	(0.19)
Investment	0.030***	0.054***	0.015*	0.018
	(0.00)	(0.00)	(0.01)	(0.01)
Interaction		-0.048***		-0.006
		(0.01)		(0.01)
Constant	1.914***	1.675***	2.309***	2.278***
	(0.05)	(0.05)	(0.11)	(0.14)
R^2	0.104	0.133	0.066	0.067
N	1830	1830	305	305

Note: Each specification is a linear regression where the outcome variable is the perception of inappropriateness of non-cooperative behavior by Player A as determined by those in the role of Player A. Inappropriateness was asked on a four point scale: appropriate (1), somewhat inappropriate (2), moderately inappropriate (3), and very inappropriate (4).. Standard errors in parentheses are clustered at the level of the individual. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 7: Linear Regression of Inappropriateness (Player B)

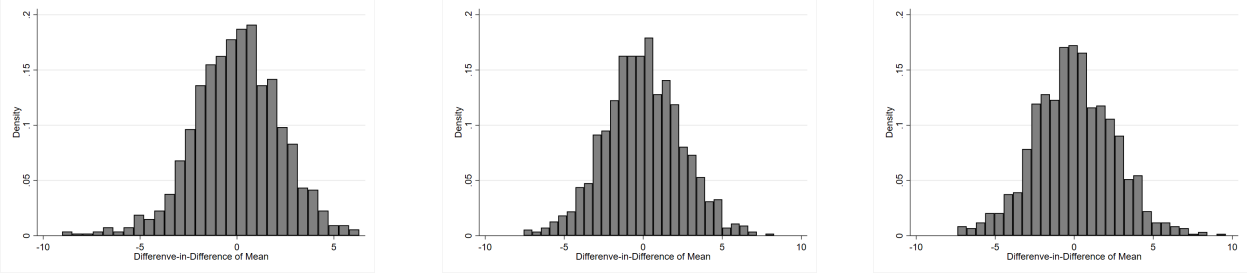
	within-subject	within-subject
Promise	0.504***	1.101***
	(0.06)	(0.07)
Investment	0.024***	0.054***
	(0.00)	(0.00)
Interaction		-0.060***
		(0.01)
Constant	1.949***	1.650***
	(0.05)	(0.05)
R^2	0.078	0.124
N	1830	1830

Note: Each specification is a linear regression where the outcome variable is the perception of inappropriateness of non-cooperative behavior by Player A as determined by those in the role of Player B. Inappropriateness was asked on a four point scale: appropriate (1), somewhat inappropriate (2), moderately inappropriate (3), and very inappropriate (4).. Standard errors in parentheses are clustered at the level of the individual. $^{*}\, p < 0.1$; $^{**}\, p < 0.05$; $^{***}\, p < 0.01$

We performed a two-step bootstrapping procedure in order to generate a non-parametric test of H1.3 on the (0-6) gap using the between-subject data from the first experiment. (We will eventually repeat this exercise using the between-subject data from the second experiment.) Let $\hat{\theta}_r$ and θ_r be the estimators of the mean reported likelihood of punishment in the Promise and No Promise conditions respectively when the promisee's reliance investments are r. We observe that $\left(\hat{\theta}_6 - \theta_6\right) - \left(\hat{\theta}_0 - \theta_0\right) = 0.097$. That is, the difference in means between the Promise and the No Promise samples is higher if promisees' reliance investments are 6 than if they are 0. We want to know the probability with which we would observe this positive difference-in-difference of means by chance. In other words, we want to test the null hypothesis that $\left(\hat{\theta}'_6 - \theta'_6\right) - \left(\hat{\theta}'_0 - \theta'_0\right) = 0$, where $\hat{\theta}'_1, \theta'_1, \hat{\theta}'_0, \theta'_0$ are the means of the underlying distributions from which our samples are drawn.

We tested the null hypothesis in two steps (see, e.g., Efron and Tibshrani, 1993, pp. 220-223). First, we recentered the original samples to conform with the null hypothesis. Specifically, we subtracted from each observation in each of the four samples the respective sample means and then added to each observation in the two Promise samples the mean effect of promising. In other words, if the mean for the combined No Promise samples is $\bar{x}$ and the mean for the combined Promise samples is $\bar{y}$, we added $(\bar{y} - \bar{x})$ to each observation in the two Promise samples.[50]

[50]By subtracting the sample means, we made our data conform to the hypothesis $\hat{\theta}'_6 = \theta'_6 = \hat{\theta}'_0 = \theta'_0 = 0$. In doing so, we eliminated all of our hypothesized effects from our data. By adding back $(\bar{y} - \bar{x})$ to the observations in the promise samples, we effectively added back in the expectations-independent effect of promising, so that our data ended up conforming to our less restrictive null hypothesis $\left(\hat{\theta}'_6 - \theta'_6\right) - \left(\hat{\theta}'_0 - \theta'_0\right) = 0$. We didn't add back in the effect of reliance alone, because doing so would leave this hypothesis unchanged.



(a) 0-10 (b) 0-20 (c) 10-20

Figure 8: Simulated Distributions for Experiment 2

We then created four synthetic samples with sample sizes equal to our real samples by randomly drawing with replacement from each of the four samples that were constructed above. We then calculated the difference-in-difference $\left(\hat{\theta}''_6 - \theta''_6\right) - \left(\hat{\theta}''_0 - \theta''_0\right)$, where $\hat{\theta}''_1, \theta''_1, \hat{\theta}''_0, \theta''_0$ are the means of these synthetic samples. After repeating this procedure 10,000 times, we obtained a simulated distribution of the differences-in-differences of the means that would arise if the null hypothesis were true (that is, if the difference of means between the Promise and the No Promise conditions was equal across different reliance levels).

The area under the curve to the right of the observed estimator 0.097 corresponds to the probability that a greater or equal difference-in-difference would have been observed if the null hypothesis were true. This value, 0.004, is small enough for us to reject the null hypothesis at the 1% level.

Experiment 2

We performed a two-step bootstrapping procedure in order to generate a non-parametric test of H1.3 on the between-subject data from the second experiment. Let $\hat{\theta}_r$ and θ_r be the estimators of the mean reported likelihood of punishment in the Promise and No Promise conditions respectively when the promisee's reliance investments are r. Our hypothesis is that $\left(\hat{\theta}_h - \theta_h\right) - \left(\hat{\theta}_l - \theta_l\right) > 0$ when $h > l$. That is, the difference in means between the Promise and the No Promise samples is higher if promisees' reliance investments are higher. We want to know the probability with which we would observe this positive difference-in-difference of means by chance. In other words, we want to test the null hypothesis that $\left(\hat{\theta}'_h - \theta'_h\right) - \left(\hat{\theta}'_l - \theta'_l\right) = 0$, where $\hat{\theta}'_h, \theta'_h, \hat{\theta}'_l, \theta'_l$ are the means of the underlying distributions from which our samples are drawn. Given that there are three investment levels, we will test this for (h, l) equal to each of $(10, 0)$, $(20, 0)$, and $(20, 10)$. We tested the null hypothesis in the same manner as described in Experiment 1 above.

The area under the curve to the right of the observed estimators corresponds to the probability that a greater or equal difference-in-difference would have been observed if the null hypothesis were true. This value is 0.929 for the 0-10 gap, 0.822 for the 0-20 gap, and 0.344 for the 10-20 gap. This means that we do not reject the null hypothesis for any of the gaps.

APPENDIX C: INSTRUCTIONS (VIGNETTE)

UNIVERSITY OF CALIFORNIA LOS ANGELES
STUDY INFORMATION SHEET

Professors Alexander Stremitzer (PhD) and Rebecca Stone (PhD), from the School of Law at the University of California, Los Angeles (UCLA) are conducting a survey.
You were selected as a possible participant because you are subscribed to Amazon Mechanical Turk as an MTurk Worker. Your participation is voluntary.

Why is this survey being done?
We are conducting this survey in order to investigate how people make decisions.

What will happen if I participate?
If you volunteer to participate, the researcher will ask you to do the following:

- Read a case study and make decisions based on the given situation
- Provide demographical data

How long will the survey take?
The survey will take a total of about 10 to 15 minutes.

Are there any potential risks or discomforts that I can expect from participating?
There are no anticipated risks or discomforts.

Are there any potential benefits if I participate?
You will not directly benefit from your participation.

The results of the research may improve our general understanding of which factors are important when people make decision in their everyday life.

What other choices do I have if I choose not to participate?
You are free to choose any other HIT ("Human Intelligence Task") on Amazon Mechanical Turk or refrain from any participation on any task.

Will I be paid for participating?
You will receive $1.50 for completing the survey. The payment will be subscribed to your account a few days after you complete it.

Will information about me and my participation be kept confidential?
Any information that is obtained in connection with this survey that can identify you will remain confidential. It will be disclosed only with your permission or as required by law. Confidentiality will be maintained by means of limiting the access to the data only to the investigators, using it only for research purposes and collecting only information that does not allow anyone to draw conclusions about your identity.

What are my rights if I participate?
You can choose whether or not you want to participate, and you may withdraw your consent and discontinue your participation at any time. Whatever decision you make, there will be no penalty to you, and no loss of any benefits to which you were otherwise entitled. You may refuse to answer any questions that you do not want to answer.

Who can I contact if I have questions about this study?
The research team:
If you have any questions, comments or concerns about the research, you can talk to the one of the researchers. Please contact:

Alexander Stremitzer: alexander.stremitzer@law.ucla.edu, phone: +1 (310) 267-4583

UCLA Office of the Human Research Protection Program (OHRPP):
If you have questions about your rights while taking part in this survey, or you have concerns or suggestions and you want to talk to someone other than the researchers, please call the OHRPP at (310) 825-7122 or write to:

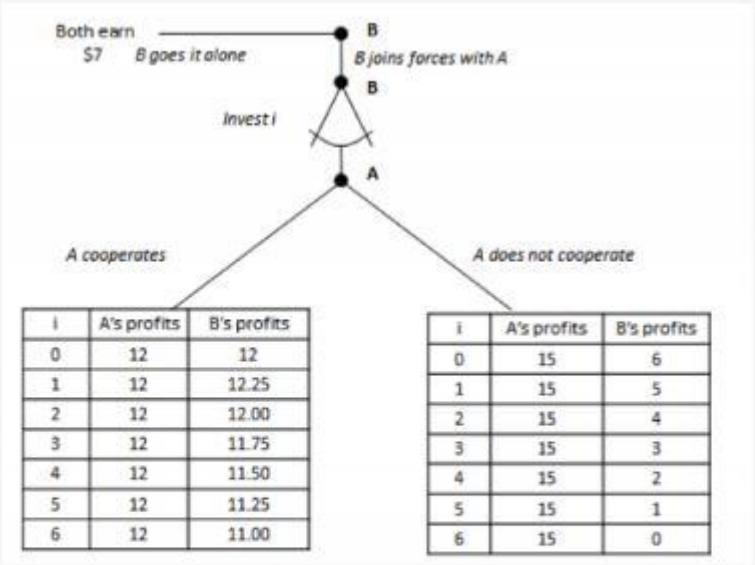
UCLA Office of the Human Research Protection Program
11000 Kinross Avenue, Suite 211, Box 951694
Los Angeles, CA 90095-1694

○ **I have read and agree to the terms and conditions**

>>

Thank you for participating in this survey. The purpose of this survey is to study how people make decisions in certain situations. You will earn $1.50 for completing the survey. The survey will take approximately 10-15 minutes.

>>



i	A's profits	B's profits
0	12	12
1	12	12.25
2	12	12.00
3	12	11.75
4	12	11.50
5	12	11.25
6	12	11.00

i	A's profits	B's profits
0	15	6
1	15	5
2	15	4
3	15	3
4	15	2
5	15	1
6	15	0

Imagine you witness the following:

Two people, A and B, have an opportunity to cooperate with each other.

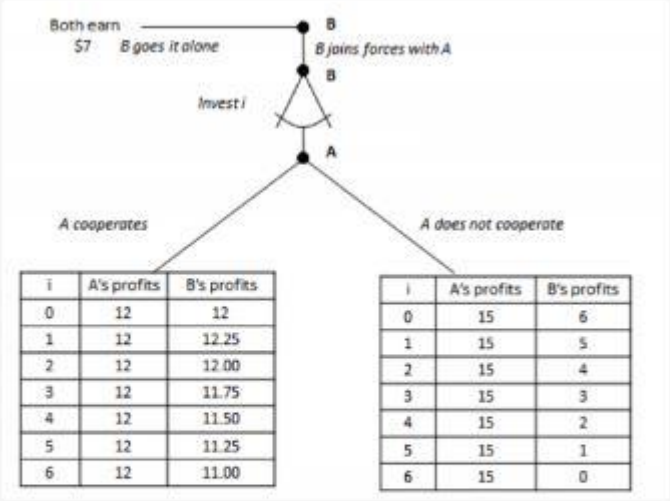
B must first choose between two alternatives (see diagram above): going it alone, or joining forces with A.

The upside of joining forces with A is that the combined monetary payoff of the two parties is much higher than if B were to go it alone.

The downside is that by joining forces with A, B puts himself at the mercy of A. This is because once B has decided to join forces with A, A can unilaterally decide not to cooperate with B in which case A gets all the profit from the venture. In other words, if B thinks that A is not going to cooperate, B is better off going it alone. However, that would leave both A and B worse off than if B had joined forces with A and A had cooperated.

If B decides to join forces with A, then, before A makes the decision whether to cooperate or not, B may invest in their cooperative venture. B's investment is lost if A does not cooperate. The higher B's level of investment, the higher will be B's loss (see right payoff table). However, if A cooperates with B, B's profits initially increase with his investment but then decrease once his investment rises beyond a certain point (see left payoff table).

>>



i	A's profits	B's profits
0	12	12
1	12	12.25
2	12	12.00
3	12	11.75
4	12	11.50
5	12	11.25
6	12	11.00

i	A's profits	B's profits
0	15	6
1	15	5
2	15	4
3	15	3
4	15	2
5	15	1
6	15	0

Preliminary Questions

1) What are the parties' profits when A cooperates and B has invested 3?

Profit of A:

Profit of B:

2) What are the parties' profits when A doesn't cooperate and B has invested 3?

Profit of A:

Profit of B:

3) If B thinks that A won't cooperate, which action will give B a higher profit?

- ○ Going it alone
- ○ Joining forces with A

>>

Imagine, you witness the following conversation:

B says to A:

"I would really like to join forces with you, but how can I be sure that you will cooperate with me rather than going it alone and taking all the profit for yourself?"

A responds:

"All I can say is that I plan to cooperate with you, though I can't promise that I will do so."

B listens carefully and decides to join forces with A.

A ***decides not to cooperate*** with B. Remember: A ***made no promise to cooperate*** with B. You observe this.

Imagine you can inflict a monetary punishment on A for not cooperating (e.g., by boycotting A's business which results in a monetary loss for A). Inflicting this punishment, however, is not costless to you (boycotting a business that you have been buying from forces you to switch to another product requiring you to change your habits, pay more for the other product, consume a product you like less, etc…).

>>

◯ Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 3 knowing that A did not make a promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
◯	◯	◯	◯	◯

>>

◯ Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 0 knowing that A did not make a promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
◯	◯	◯	◯	◯

>>

◯ Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 6 knowing that A did not make a promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
◯	◯	◯	◯	◯

>>

Imagine, you witness the following conversation:

B says to A:

"I would really like to join forces with you, but how can I be sure that you will cooperate with me rather than going it alone and taking all the profit for yourself?"

A responds:

"I understand that you are worried I could take advantage of you, but I promise I will cooperate with you."

B says:

"Okay, if you promise to cooperate with me, let's work together."

B subsequently decides to join forces with A.

A ***decides not to cooperate*** with B. Remember: A ***promised to cooperate*** with B. You observe this.

Imagine you have the power to inflict a punishment on A for breaking the promise to B (e.g., by boycotting A's business which results in a monetary loss for A). Inflicting this punishment, however, is not costless to you (boycotting a business that you have been buying from forces you to switch to another product requiring you to change your habits, pay more for the other product, consume a product you like less, etc…).

>>

Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 6 in reliance on A's promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
○	○	○	○	○

>>

○ Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 3 in reliance on A's promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
○	○	○	○	○

>>

◯ Go back to display scenario instructions again

What is the likelihood you would choose to inflict the punishment on A ***if B has invested 0 in reliance on A's promise***?

Very Unlikely	Unlikely	Undecided	Likely	Very Likely
◯	◯	◯	◯	◯

>>

Did you understand how A's and B's actions affect their profits?

- ◯ Yes
- ◯ No
- ◯ Kind of

Finally, please answer the following questions:

	Yes	No
I didn't take the scenario seriously. I just wanted to earn the $1.50 fee as quickly as possible.	◯	◯
I carefully read the instructions.	◯	◯
I chose my answers in order to make myself seem like a good person.	◯	◯
This is the first time I have completed this survey.	◯	◯

>>

Please provide some demographical information.

What is your age?

What is your gender?

- ○ Female
- ○ Male

Is English your first language?

- ○ Yes
- ○ No

What is your highest level of schooling?

- ○ Master's, doctoral, or professional degree such as medicine or law
- ○ Bachelor's degree
- ○ Associate's degree
- ○ Vocational or technical certificate / diploma after high school (such as cosmetics)
- ○ High school diploma
- ○ I did not complete high school

Are you an mTurk Master Worker (your response to this questions will have no effect on your payout)?

- ○ Yes
- ○ No
- ○ I do not know what an mTurk Master Worker is.

>>

Thank you for participating in the survey.

Here is your MTurk Code: 752796

To receive payment for participating, click "Accept HIT" in the Mechanical Turk window, enter this code, and then click "Submit".

APPENDIX D: INSTRUCTIONS (INCENTIVIZED)

Instructions Page

Instructions

Welcome to this study. After carefully reading the instructions and answering questions to check your understanding, you can begin. During the study, you can earn money based on your decisions and the decisions of the other two participants.

Two participants will make decisions that affect each other. Before the game starts, these participants can communicate with each other if they decide to do so. After these two participants have had the opportunity to communicate with one another, they are randomly assigned to the roles of A and B. The third participant is assigned to the role of C.

The game proceeds as follows (see figure). At the beginning of the game, players receive endowments of tokens, which will be converted to 0.10 CHF at the end of the game. Players A and B receive an endowment of 70 tokens each, while player C receives an endowment of 110 tokens.

B then decides whether to spend 0, 10, or 20 tokens from his or her endowment after which A chooses between *Cooperate* and *Don't Cooperate*.

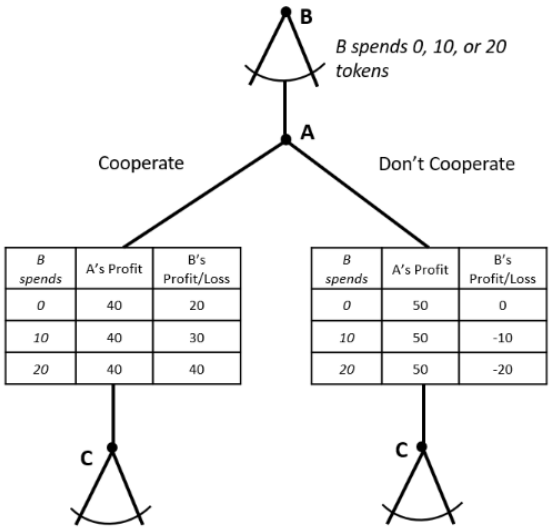
A's choice affects both A's and B's payoffs. If A chooses Don't Cooperate, A receives a higher payoff and B receives a lower payoff. If A chooses Cooperate, A receives a lower payoff and B receives a higher payoff. This is true regardless of B's level of investment. In addition, B's payoff is higher the more B spent if A chooses Cooperate and B's payoff is lower the more B spent if A chooses Don't Cooperate. The amount B spent doesn't affect A's payoff.

Finally C, has the opportunity to reduce A's payoff after observing A's and B's decisions. Doing so is costly to C but is even more costly for A. For every token spent by C, A's payoff is reduced by 3 tokens.

In the end, players receive payouts that are equal to their endowments plus or minus any profits or losses, modified by the effect of C's decision to reduce A's payoff.

More detailed instructions are below.

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by **1** token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

Detailed Instructions

Detailed Instructions

This study involves three participants: Player A, Player B, and Player C. After reading the instructions and answering some questions to check your understanding, you will be randomly assigned to one of the three roles and matched with two other players. Your identity and the identities of the other two participants are anonymous.

Step 1: Communication Phase. The computer randomly selects two players who may communicate with each other. After they are finished communicating, they are randomly assigned to the roles of A and B. The remaining player is assigned to the role of C.

Step 2: Player B's Decision. Player B can choose to spend **0, 10, or 20** tokens from her endowment, generating a return as explained below. Player B's decision has no effect on Player A, but affects her return depending on Player A's decision.

If B spends 0 tokens, B's profit is 20 if A chooses Cooperate, and 0 if A chooses Don't Cooperate, which will be added to her endowment. If B spends 10 tokens, her profit will be 30 if A chooses Cooperate, but she will have a loss of 10 tokens if A chooses Don't Cooperate. If B spends 20 tokens, her profit will be 40 if A chooses Cooperate, but she will lose 20 if A chooses Don't Cooperate.

Step 3: Player A's Decision. Player A observes how many tokens B spends and then Player A must decide between "Cooperate" and "Don't Cooperate". Player A gets 40 tokens from choosing Cooperate and 50 tokens from choosing Don't Cooperate.

- If Player A chooses "Cooperate", the preliminary profits / losses of the players are:

	A	B			C
B spends	**A's profit**	**B's gross return**	**B's spent tokens**	**B's profit/loss**	**C's profit**
0	40	20	0	20 - 0 = 20	0
10	40	40	10	40 - 10 = 30	0
20	40	60	20	60 - 20 = 40	0

- If Player A chooses "Don't Cooperate", the preliminary profits / losses of the players are:

	A	B			C
B spends	**A's profit**	**B's gross return**	**B's spent tokens**	**B's profit/loss**	**C's profit**
0	50	0	0	0 - 0 = 0	0
10	50	0	10	0 - 10 = -10	0
20	50	0	20	0 - 20 = -20	0

Step 4: Player C's Decision. Player C can spend money to reduce Player A's payoff. Each token spent reduces Player C's profits by 1 token and reduces Player A's profits by 3 tokens, while leaving B's profits / losses unchanged. Player C can choose to spend between 0 and 30 tokens.

Step 5: Final Payoffs. Each player's profits / losses are added to their initial endowment. Players are then paid 0.10 CHF for each token.

Next

Preliminary Questions

Preliminary Questions:

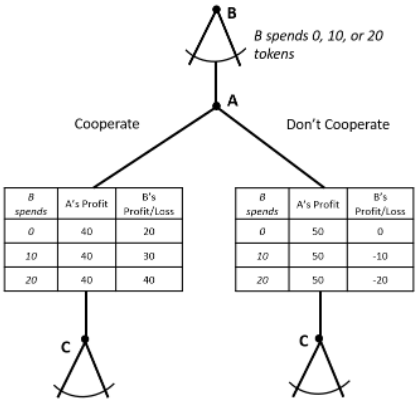
Before the study begins, you will answer some preliminary questions to check your understanding of the instructions. You must get the answers correct to proceed.

Next

Preliminary Questions

Preliminary Questions

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by 1 token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

1) What are the players' profits / losses if:
- Player B spends 10 tokens
- Player A chooses Cooperate
- Player C spends 0 tokens on reducing A's payoff?

Player A's profits / losses:

Player B's profits / losses:

Player C's profits / losses:

2) What are the players' profits / losses if:
- Player B spends 10 tokens
- Player A chooses Cooperate
- Player C spends 5 tokens on reducing A's payoff?

Player A's profits / losses:

Player B's profits / losses:

Player C's profits / losses:

3) What are the players' profits / losses if:
- Player B spends 10 tokens
- Player A chooses Don't Cooperate
- Player C spends 5 tokens on reducing A's payoff?

Player A's profits / losses:

Player B's profits / losses:

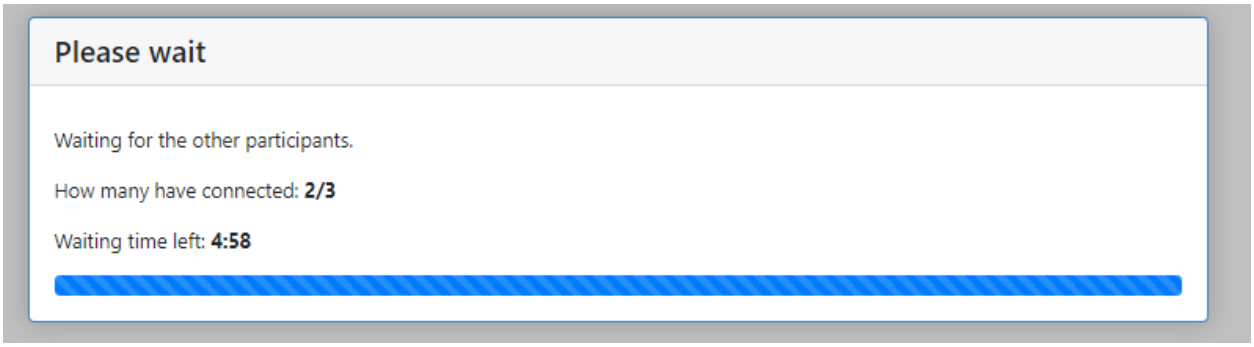
Player C's profits / losses:

4. Which decision takes place first?

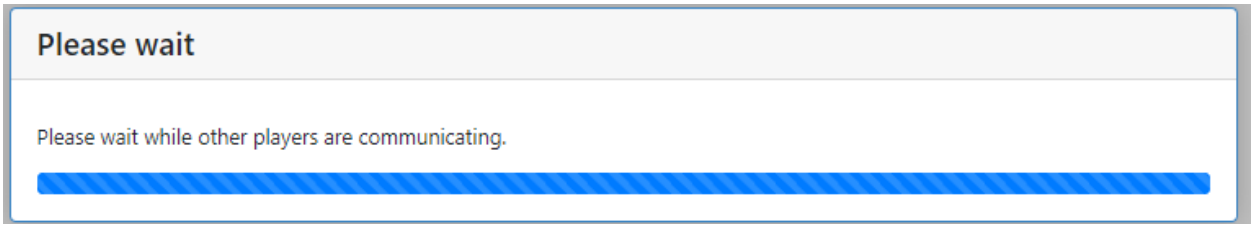
- ○ A's choice of Cooperate or Don't Cooperate.
- ○ B's choice to spend 0, 10, or 20 tokens.

Next

Waiting screen while being matched



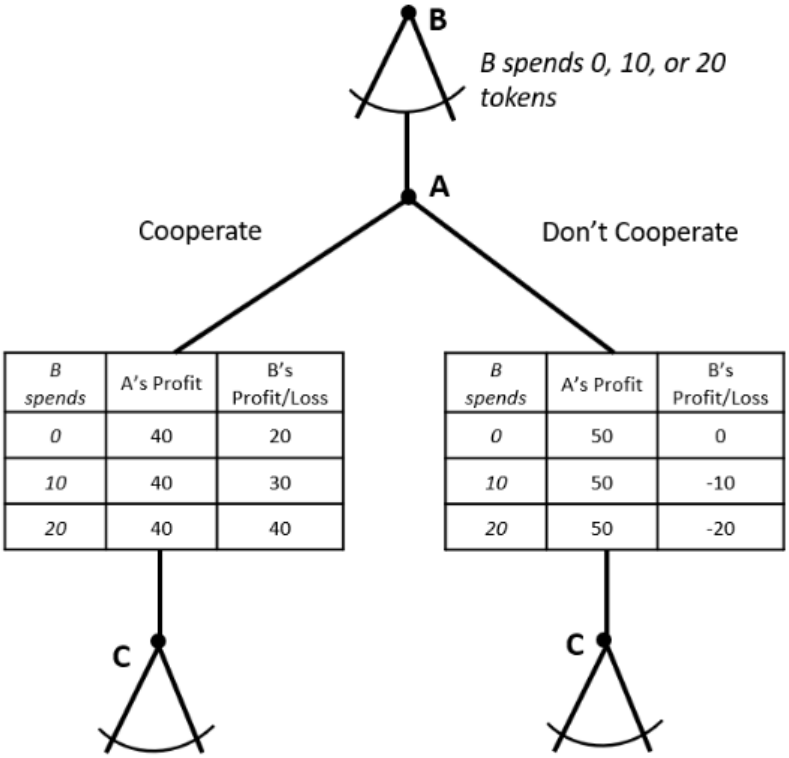
Waiting screen for player C while A and B are communicating



Player A and B communication screen

Make your decision within 120 seconds **1:45**

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by 1 token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

You can now choose to send a message to the other player.

○ I promise to choose Cooperate if I am chosen to be Player A.

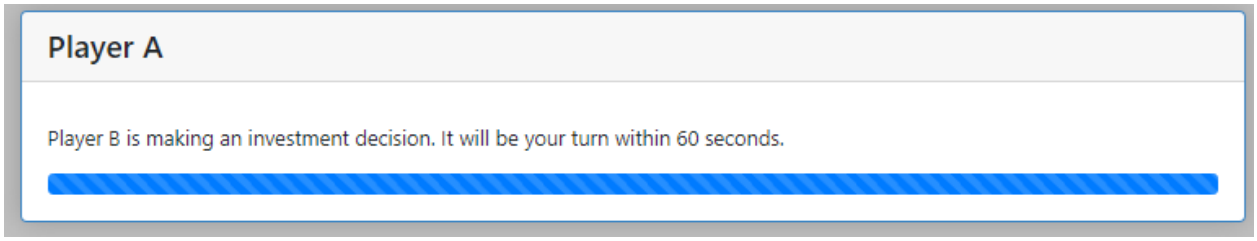
○ Do not send a message.

Select your message and then hit the button.

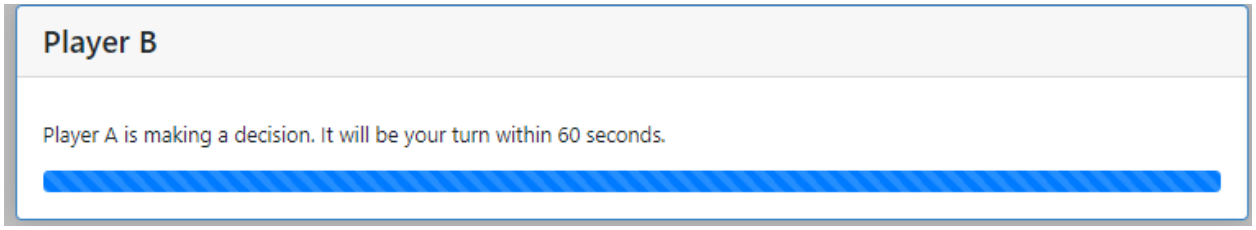
Next

Review Instructions

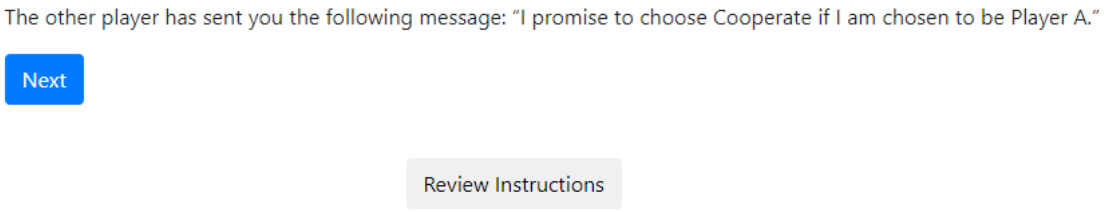
Player A waiting screen while B decides on investment



Player B waiting screen while A decides on cooperation



Player A and B Cooperation Screen:

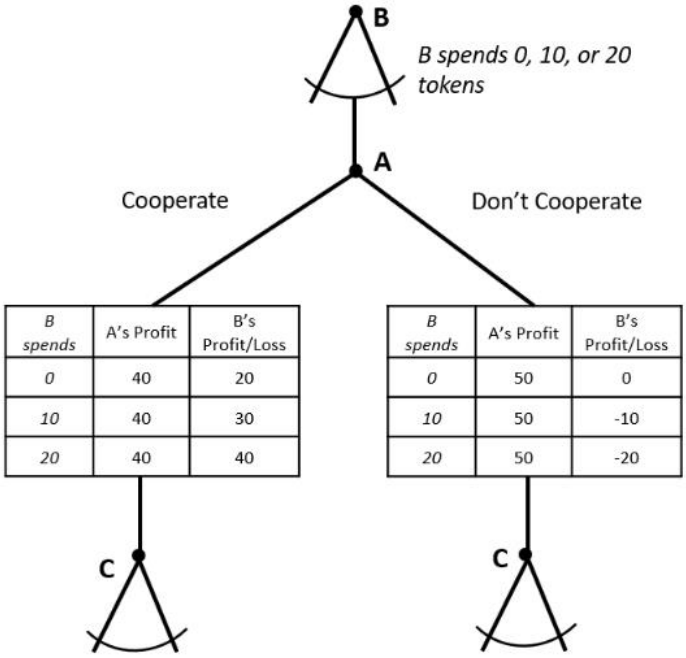


Player B investment decision

Player B

Make your decision within 60 Seconds. **0:45**

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by 1 token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

You can now choose to spend 0, 10 or 20 tokens. Remember, that both your choice and Player A's decision will affect your final payoff.

Please make your choice: :

10

Next

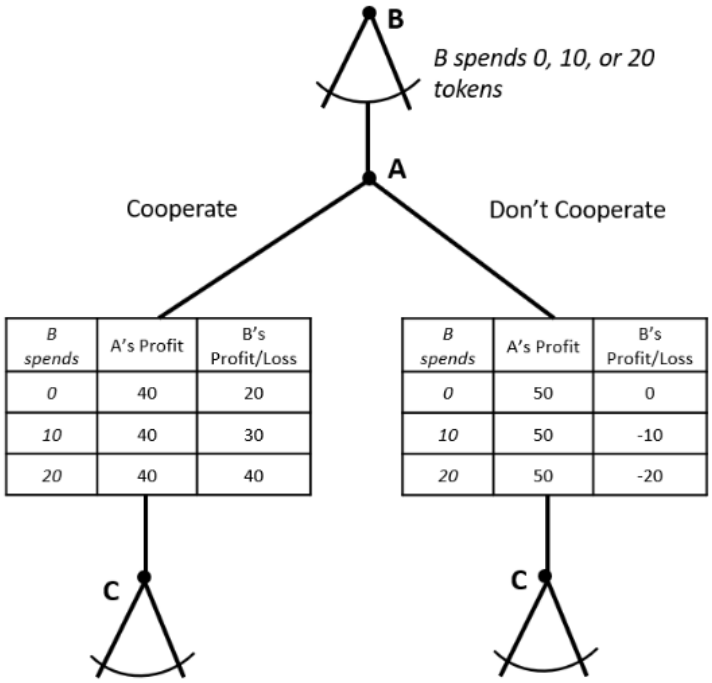
Review Instructions

Player A Cooperation Decision

Player A

Make your decision within 60 seconds. **0:43**

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by 1 token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

Player B has spent 10 tokens.

You can now decide between "Don't Cooperate" and "Cooperate". Remember, your payoff will depend on your choice and whether C spends any money reducing your payoff. Please make your decision.

Your decision?

- (•) Cooperate
- ( ) Don't Cooperate

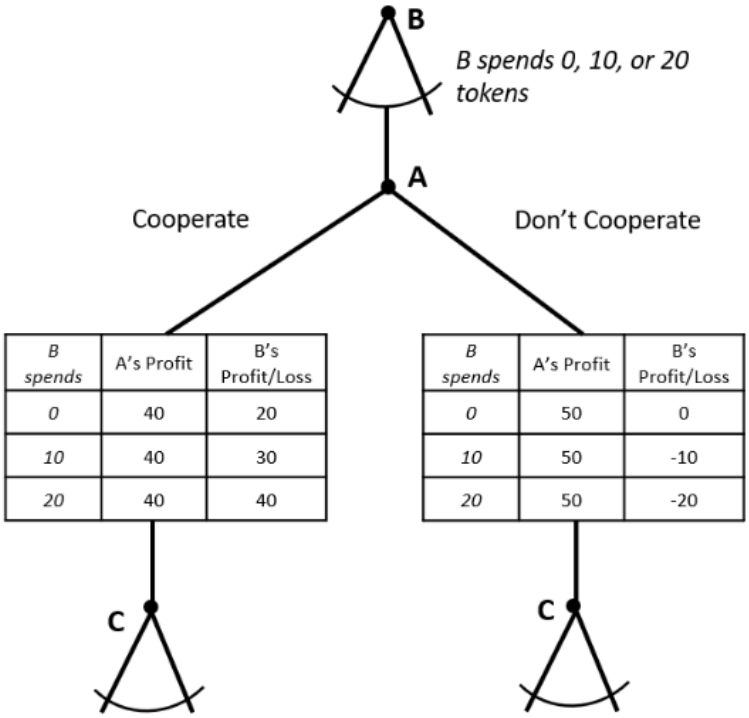
Next

Review Instructions

Player C punishment screen – Between Subject

Player C

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by 1 token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

You can now decide how much to spend on reducing A's payoff. Each token that you spend will reduce your profits by 1 token and A's profits by 3 tokens. You can spend up to 30 tokens.

Player A **promised to choose Cooperate during the communication phase**, Player B **spent 0**, and Player A **chose Don't Cooperate**

Next

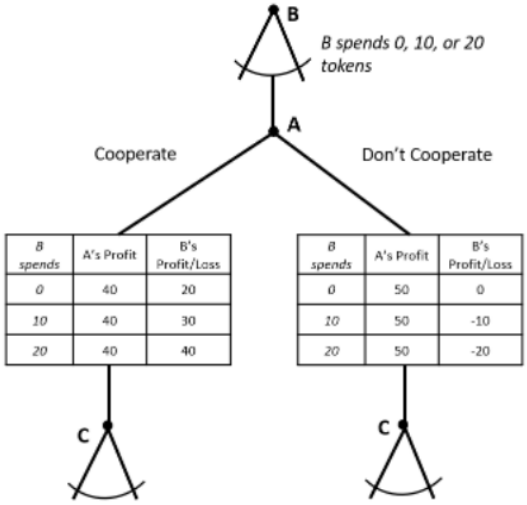
Review Instructions

Player C punishment screen – within subject – No cooperation

Player C

Time left to complete this page: **4:25**

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by 1 token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

The table below illustrates all of the possible scenarios in which A may have chosen **Don't Cooperate.** For each scenario, **please indicate how much you would like to spend on reducing A's payoff.** Your choice will be matched with the actual choices that A made in response to B's spending to determine the actual reduction of A's payoff.

Each token that you spend will reduce your profits by 1 token and A's profits by 3 tokens. You can spend up to 30 tokens.

Scenario	Punishment
Player A **decided not to communicate with Player B during the communication phase**, Player B **spent 0**, and Player A **chose Don't Cooperate**	
Player A **decided not to communicate with Player B during the communication phase**, Player B **spent 10**, and Player A **chose Don't Cooperate**	
Player A **decided not to communicate with Player B during the communication phase**, Player B **spent 20**, and Player A **chose Don't Cooperate**	
Player A **promised to choose Cooperate during the communication phase**, Player B **spent 0**, and Player A **chose Don't Cooperate**	2
Player A **promised to choose Cooperate during the communication phase**, Player B **spent 10**, and Player A **chose Don't Cooperate**	
Player A **promised to choose Cooperate during the communication phase**, Player B **spent 20**, and Player A **chose Don't Cooperate**	

Next

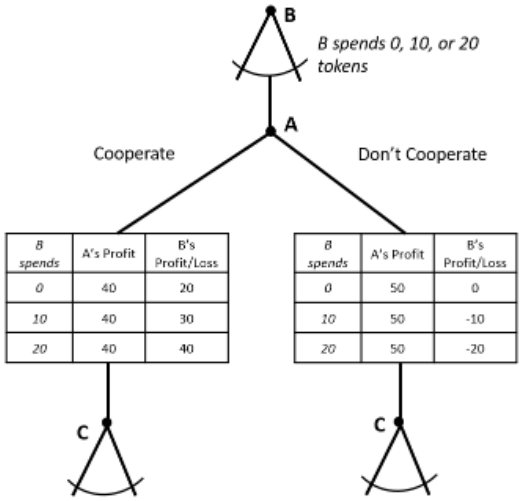
Review Instructions

Player C punishment screen – within subject – Cooperation

Player C

Time left to complete this page: **4:35**

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by 1 token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

The table below illustrates all of the possible choices made by A and B for which you can reduce A's payoffs, given that A **chose Cooperate**. For each option, please indicate how much you would like to spend on reducing A's payoff. Your choice will be matched with the actual choices made by A and B to determine the punishment.

Each token that you spend will reduce your profits by 1 token and A's profits by 3 tokens. You can spend up to 30 tokens.

Scenario	Punishment
Player A **decided not to communicate with Player B during the communication phase**, Player B **spent 0**, and Player A **chose Cooperate**	
Player A **decided not to communicate with Player B during the communication phase**, Player B **spent 10**, and Player A **chose Cooperate**	
Player A **decided not to communicate with Player B during the communication phase**, Player B **spent 20**, and Player A **chose Cooperate**	
Player A **promised to choose Cooperate during the communication phase**, Player B **spent 0**, and Player A **chose Cooperate**	
Player A **promised to choose Cooperate during the communication phase**, Player B **spent 10**, and Player A **chose Cooperate**	
Player A **promised to choose Cooperate during the communication phase**, Player B **spent 20**, and Player A **chose Cooperate**	

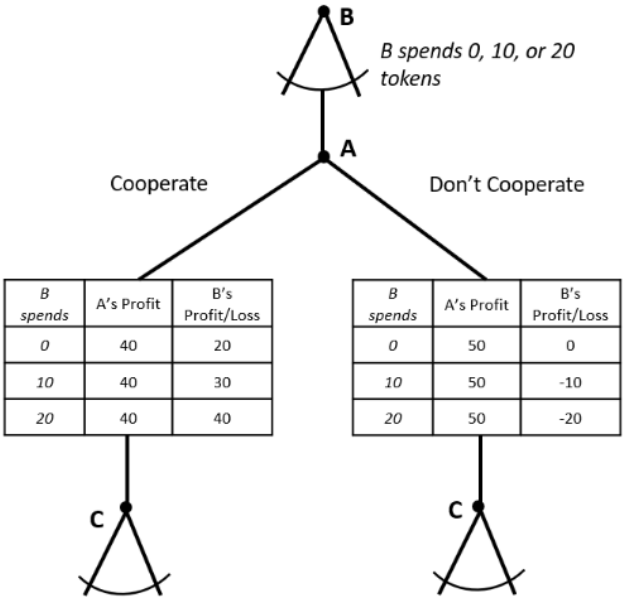
Next

Review Instructions

Player A Appropriateness screen – Between Subject

Player A

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by 1 token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

You now have an opportunity to earn additional bonus money based on how you think that others perceive particular actions.

For the scenario below, please indicate whether you believe that A's choice of "Don't Cooperate" would be very inappropriate, moderately inappropriate, somewhat inappropriate, or appropriate.

We will determine which response was selected by the most participants assigned the role of Player A. If you give the response that is most frequently given by other people, you will receive a bonus of 0.20 CHF.

For the scenario below indicate how appropriate or inappropriate it would be for Player A to choose "Don't Cooperate."

Scenario	Very inappropriate	Moderately inappropriate	Somewhat inappropriate	Appropriate
A made a promise to choose Cooperate during the communication phase and B spent 0	○	○	○	○

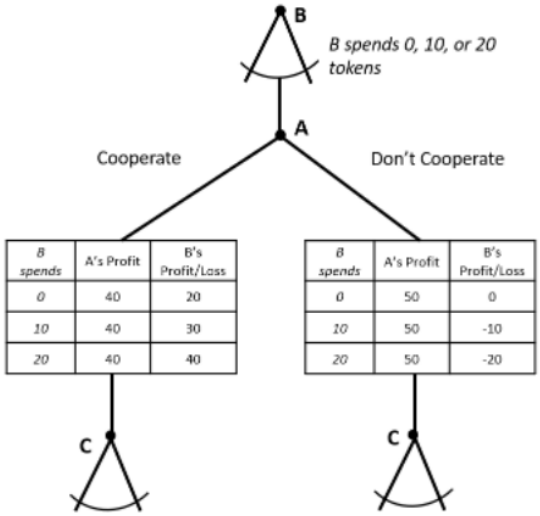
Next

Review Instructions

Player A Appropriateness – Within Subject

Player A

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by 1 token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

You now have an opportunity to earn additional bonus money based on how you think that others perceive particular actions.

The tables below give a list of possible scenarios. For each of the scenarios, please indicate whether you believe that A's choice of "Don't Cooperate" would be very inappropriate, moderately inappropriate, somewhat inappropriate, or appropriate.

For each of the choices, we will determine which response was selected by the most participants assigned the role of Player A. If you give the response that is most frequently given by other people, you will receive a bonus of 0.20 CHF for each correct answer.

For the scenarios below indicate how appropriate or inappropriate it would be for Player A to choose "Don't Cooperate."

Scenario	Very inappropriate	Moderately inappropriate	Somewhat inappropriate	Appropriate
A decided not to communicate with B during the communication phase and B spent 0	○	○	○	○
A decided not to communicate with B during the communication phase and B spent 10	○	○	○	○
A decided not to communicate with B during the communication phase and B spent 20	○	○	○	○

Player A's Choice	Very inappropriate	Moderately inappropriate	Somewhat inappropriate	Appropriate
A made a promise during the communication phase and B spent 0	○	◉	○	○
A made a promise during the communication phase and B spent 10	○	○	○	○
A made a promise during the communication phase and B spent 20	○	○	○	○

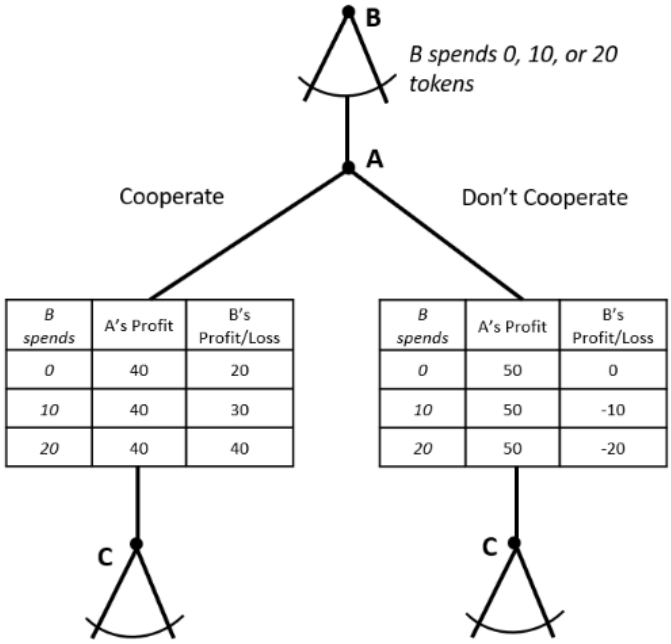
Next

Review Instructions

Player B Appropriateness screen – Between Subject

Player B

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by 1 token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

You now have an opportunity to earn additional bonus money based on how you think that others perceive particular actions.

For the scenario below, please indicate whether you believe that A's choice of "Don't Cooperate" would be very inappropriate, moderately inappropriate, somewhat inappropriate, or appropriate.

We will determine which response was selected by the most participants assigned the role of Player B. If you give the response that is most frequently given by other people, you will receive a bonus of 0.20 CHF.

For the scenario below indicate how appropriate or inappropriate it would be for Player A to choose "Don't Cooperate."

Scenario	Very inappropriate	Moderately inappropriate	Somewhat inappropriate	Appropriate
A made a promise to choose Cooperate during the communication phase and B spent 0	○	○	○	○

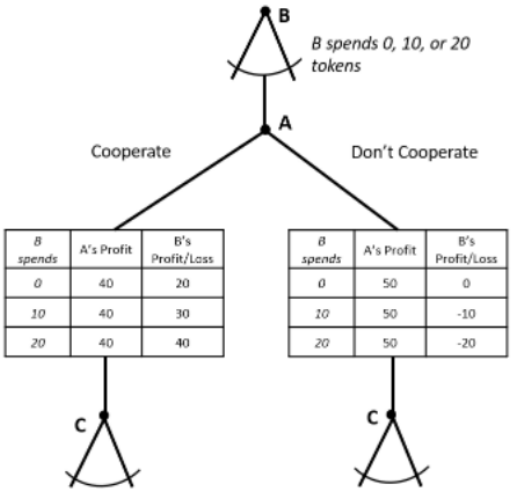
Next

Review Instructions

Player B Appropriateness screen – Within Subject

Player B

Endowments:
Player A = 70 tokens
Player B = 70 tokens
Player C = 110 tokens



B spends	A's Profit	B's Profit/Loss
0	40	20
10	40	30
20	40	40

B spends	A's Profit	B's Profit/Loss
0	50	0
10	50	-10
20	50	-20

Each token spent by C:
- Reduces C's profits by 1 token
- Reduces A's profits by 3 tokens
- Has no effect on B's profits/losses

Final Payment = Endowment + Profits - Losses

You now have an opportunity to earn additional bonus money based on how you think that others perceive particular actions.

The tables below give a list of possible scenarios. For each of the scenarios, please indicate whether you believe that A's choice of "Don't Cooperate" would be very inappropriate, moderately inappropriate, somewhat inappropriate, or appropriate.

For each of the choices, we will determine which response was selected by the most participants assigned the role of Player B. If you give the response that is most frequently given by other people, you will receive a bonus of 0.20 CHF for each correct answer.

For the scenarios below indicate how appropriate or inappropriate it would be for Player A to choose "Don't Cooperate."

Scenario	Very inappropriate	Moderately inappropriate	Somewhat inappropriate	Appropriate
A decided not to communicate with B during the communication phase and B spent 0	○	○	○	○
A decided not to communicate with B during the communication phase and B spent 10	○	○	○	○
A decided not to communicate with B during the communication phase and B spent 20	○	○	○	○

Player A's Choice	Very inappropriate	Moderately inappropriate	Somewhat inappropriate	Appropriate
A made a promise during the communication phase and B spent 0		◉		
A made a promise during the communication phase and B spent 10	○	○	○	○
A made a promise during the communication phase and B spent 20	○	○	○	○

Next

Review Instructions

Player C Punishment – Free Response

Please briefly describe what factors influenced your decisions about reducing Player A's payoff:

Next

Player A – Free Response

Please briefly describe what factors influenced your decision about making a promise:

Please briefly describe what factors influenced your decision about choosing Cooperate or Don't Cooperate:

Next

Player B – Free Response

Please briefly describe what factors influenced your decision about making a promise:

Please briefly describe what factors influenced your decision to spend 0, 10, or 20 tokens:

Next

Survey Questions 1

Survey Questions (1 of 2)

Please answer the following questions (your response to these questions will have no effect on your payout)

	Yes	No
I did not take the scenario seriously. I just wanted to earn the fee as quickly as possible	○	○
I carefully read the instructions	○	○
I chose my answers to make myself seem like a good person	○	○
This is the first time that I completed this survey	○	○

Next

Review Instructions

Survey Questions 2

Survey Questions (2 of 2)

Please answer the following questions. These will have no effect on your payout.

What is your age?

What is your gender?

- ○ Male
- ○ Female
- ○ Non-binary
- ○ Prefer not to answer

What is your first language?

- ○ German
- ○ French
- ○ Italian
- ○ English
- ○ Other

What is your highest education level?

- ○ Some high school
- ○ Completed high school
- ○ Some college
- ○ Completed college
- ○ Post-college (i.e. graduate) degree (e.g. Masters, Phd)

Do you have any other comments for the researchers?

Next

Review Instructions