Economic Theory
https://doi.org/10.1007/s00199-024-01563-y

RESEARCH ARTICLE

Voluntary versus mandatory information disclosure in the sequential prisoner's dilemma

Georg Kirchsteiger[1] · Tom Lenaerts[2,3,4] · Rémi Suchon[5]

Received: 25 July 2022 / Accepted: 23 February 2024

Abstract

In sequential social dilemmas with stranger matching, initiating cooperation is inherently risky for the first mover. The disclosure of the second mover's past actions may be necessary to instigate cooperation. We experimentally compare the effect of mandatory and voluntary disclosure with non-disclosure in a sequential prisoner's dilemma situation. Our results confirm the positive effects of disclosure on cooperation. We also find that voluntary disclosure is as effective as mandatory disclosure, which runs counter to the results of existing literature on this topic. With voluntary disclosure, second movers who have a good track record chose to disclose, suggesting that they anticipate non-disclosure would signal non-cooperativeness. First movers interpret non-disclosure correctly as a signal of non-cooperativeness. Therefore, they cooperate less than half as often when the second mover decides not to disclose.

Keywords Information disclosure · Sequential social dilemma · Laboratory experiment

We thank Elias Fernandez and Antoine Deplancke for helping with the experiment in Brussels and Lille. This project received financial support from Fonds de la Recherche Scientifique (FNRS, PDR T014318F). The data used in this manuscript can be found at https://doi.org/10.5281/zenodo.10691582.

✉ Rémi Suchon
remi.suchon@univ-catholille.fr

Georg Kirchsteiger
georg.kirchsteiger@ulb.be

Tom Lenaerts
tom.lenaerts@ulb.be

[1] CEPR - CESifo, ECARES, Université Libre de Bruxelles, Brussels, Belgium

[2] Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium

[3] Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, Belgium

[4] Center for Human Compatible AI, UC Berkeley, CA, USA

[5] ANTHROPO LAB - ETHICS EA 7446, Université Catholique de Lille, 59000 Lille, France

Published online: 06 April 2024

JEL Classification C7 · C92 · D8 · D9

1 Introduction

A lot of economic interactions, in markets or elsewhere, resemble sequential social dilemmas: they require trust between strangers. Therefore, a lot of marketplaces have implemented mechanisms of information disclosure to improve trust (e.g. of a buyer or an employer) and trustworthiness (of a seller or a job candidate). In this paper, we experimentally compare the effectiveness of mandatory versus voluntary disclosure system in sustaining cooperation in a sequential social dilemma.

Voluntary and mandatory information disclosure both exist in the field. In some markets, sellers may hide some relevant information about the quality of their products. For instance, food producers can choose (not) to display the labels revealing the nutritional values of their products, or in some localities, restaurants can freely choose to reveal their hygiene scores (see e.g. Dranove and Jin 2010, for additional examples). In some situations, market participants can voluntarily reveal information about their past behavior: job candidates can choose to complement their applications with recommendation letters from past employers, and prospective tenants can complement their files with rental receipts issued by former landlords. Moreover, many reputation mechanisms have some discretion about the revelation of past information. For instance, in online markets with information about past behavior, it is often possible for participants to manipulate what to disclose from their past records by creating a new alias after a history of uncooperative behavior.

A number of lab experiments have confirmed that information disclosure helps sustain cooperation in sequential social dilemma (Bolton et al. 2004; Bohnet and Huck 2004; Charness et al. 2011; Duffy et al. 2013). These papers focus on mandatory disclosure, whereas we compare mandatory with voluntary information disclosure mechanisms. Mandatory disclosure maybe opposed by individuals, who may value privacy and control over their data.[1] Moreover, enforcing mandatory disclosure might be costly, as a central authority would have to allocate resources to monitor disclosure and sanction non-compliers. For these reasons, reputation systems based on voluntary information disclosure may be more desirable.[2]

In theory, systems with voluntary disclosure should be as effective as mandatory disclosure due to the unraveling principle (Milgrom 1981). Those who choose not to disclose their record will all be treated the same, so there is an incentive for those with a good record to disclose their records. This leads to unraveling: only those with bad records will withhold information. Since this is anticipated, "non-disclosers" will be treated with skepticism, i.e. one does not cooperate with them or refrain from interacting with them at all. This mechanism provides an incentive to build a good record, i.e. to cooperate. However, evidence from sender–receiver experiments has demonstrated

[1] See e.g. Varian (2009), Acquisti et al. (2016) for reviews on the economics of privacy and e.g. Benndorf et al. (2015), Benndorf and Normann (2018), Schudy and Utikal (2017) for recent experimental evidence on privacy preferences.

[2] In this line, Bertomeu and Cianciaruso (2018) propose a theoretical appraisal of voluntary disclosure and show that the welfare effect of mandatory disclosure may be ambiguous.

the limits of the unraveling principle. A common observation is that senders with bad private information exploit receivers who are not skeptical enough (see e.g. Jin et al. 2021; Montero and Sheth 2021; Sheth 2021, for recent examples).[3] This evidence was derived in the context of sender–receiver games, where the information refers to the realization of a random variable, and where the sender has an incentive to hide information that would induce the receiver to make a choice disadvantageous for the sender. Such a sender–receiver game is very different from the cooperation context we are investigating, where the information refers to players' past cooperation choices. Specifically, in sender–receiver games, the underlying information is exogenous and does not depend on the senders' past choices. This may be a good approximation of e.g. information disclosure of a product's quality when this quality cannot be influenced by the seller. However, in many important real-life situations, one must decide on whether to disclose information about one's own past decisions, for instance information on whether one betrayed trust in the past. We thus investigate whether the limits of the unraveling principle observed in sender–receiver games are also present when the information disclosure refers to own choices by testing whether voluntary information disclosure is less effective than the mandatory one in the context of a sequential social dilemma with the possibility of partner avoidance.

To investigate this question, we set up a sequential prisoner's dilemma experiment (SPD), where the first player 1 (P1—by convention female) decides whether to cooperate (trust) or defect. After being informed about P1's decision, player 2 (P2—by convention male) takes his cooperation decision (reciprocate trust or renege). Before the SPD takes place, P1 decides whether the game should be played or whether she takes an outside option instead. The outside option gives both players higher payments than what they get in the SPD if both defect, i.e. if they play the unique Nash equilibrium, but less than what they get when both players cooperate. Hence, if P1 does not trust that P2 would cooperate, she should take the outside option, while if she trusts him, she should opt into the SPD and choose cooperation.

Each subject plays this game repeatedly with fixed roles but with changing partners with exogenous matching (stranger matching). Before P1 decides whether to pick the outside option or the SPD, she might get informed about her prospective P2's past cooperation choices. If information about P2's past choices is disclosed, it is correct and complete. The disclosure of information is either mandatory ("Mand" treatments), or it is voluntary, i.e. only disclosed if P2 wants so ("Vol" treatments). The information might be disclosed with 100% probability, (no noise, "NN" treatments), or it is disclosed only with 90% probability (low noise, "LN" treatments). This implies four treatments: MandNN, MandLN, VolNN, and VolLN. Besides, in a baseline treatment, no information gets disclosed. In this treatment, P1 has to choose between the SPD and the outside option without having any information about the previous choices of her prospective P2. The noise treatments allow us to measure the robustness of voluntary disclosure when seeing no information might be plausibly blamed on bad luck. They also allow us to explore skepticism, i.e. the extent to which facing non-disclosure deters P1s from cooperating. We do so by comparing the behavior of P1 when the lack

[3] There is also evidence of failures of the unraveling principle outside the lab. For instance, Brown et al. (2012) show that film studios exploit the fact that moviegoers fail to anticipate that movies that are not reviewed before release tend to be of low quality.

of information is due to P2's choice (in the VolNN treatment) with the behavior of P1 when the lack of information might be due to chance (in the LN treatments).[4]

As expected, we find that disclosure of the record of P2s' past choices increases P1s' trust levels as measured by their cooperation choices. As a result, full cooperation by both players is significantly more often observed when such an information-disclosure system is in place. The interplay between information disclosure and partner avoidance can mitigate the problem of dilemma situations. In contrast to the existing literature identifying limits of unraveling, voluntary disclosure is just as effective as mandatory disclosure. When given the choice, P2s reveal their records most of the time, and the probability of information disclosure increases with the quality of P2's record. P1s anticipate this, and they are skeptical: In the voluntary disclosure treatments, they do not trust P2s of whom they do not see the records. As a placebo test, we compare trust by P1s when seeing the record and when not seeing it in the MandLN treatment. In this treatment, the impact of information on P1s' cooperation rates is much smaller than in the VolLN treatment. Importantly, we find that the level of disclosure and skepticism are already high in the first periods of the game, suggesting that learning plays only a limited role. Our results also reveal that information disclosure is somewhat less effective when the information system is not perfectly reliable, ie in treatments with noise. The rather low noise level of the LN treatments is enough to reduce the overall cooperation level significantly.

Our main results are in contrast with the limits of unraveling found in sender–receiver experiments. In particular, we find that Voluntary and Mandatory disclosure are equally effective even very early in the game. This contrasts with Jin et al. (2021) who find that repetition and substantial feedback are necessary to reduce the failure of unraveling. To investigate the robustness of our results, and understand why they do not reflect the failure of unraveling observed in Jin et al. (2021), we run two additional treatments. These treatments closely replicate the MandNN and VolNN treatments, with one change: instead of revealing information about all their past choices, P2s (choose to) disclose only their actions in the previous period. Therefore, the disclosed information is simpler, which might impact disclosure and skepticism (Jin et al. 2022). In addition, short-run reputation allows experimentation early in the game which might affect learning. Nonetheless, we find similar results in the short-run reputation treatments, both qualitatively and quantitatively. Therefore, we can conclude that our results are robust to a change of the complexity of the disclosed information. This further confirms that the unravelling principle might work better in situations in which information about past behavior, and not about the realisation of a random variable, gets disclosed.

The remainder of the paper is organized as follows. Section 2 briefly reviews the related literature. Section 3 details the experimental design, Sect. 4 presents the results from the main treatments, Sect. 5 reports on the effect of short-run reputation, and Sect. 6 discusses our results and concludes.

[4] Note that we do not have a repeated game structure. The players play the game repeatedly but with different partners. Hence, a player cannot punish or reward his/her partner for "bad" or "nice" choices made by his/her partner in the previous periods, since the partner changes from period to period. The spillovers between the periods are purely based on the reputation of P2 and the reactions of the P1's to the P2's reputations.

2 Related literature

We contribute mainly to two streams of literature. First, we contribute to the experimental literature on the role of information about past choices in sustaining cooperation in sequential social dilemmas with stranger matching.[5] Bolton et al. (2004) show in a mini-trust game with stranger matching and finite repetitions that providing history on the second movers' past choices increases trust and trustworthiness significantly compared to a baseline treatment without information. They also show that information disclosure still leads to less efficiency than partner matching. Similar findings are reported in Bohnet and Huck (2004) and Charness et al. (2011).[6] We provide further evidence of the positive effect of information disclosure to establish sequential cooperation and extend it into the sequential prisoner's dilemma. More importantly, our main contribution is to compare the effectiveness of voluntary and mandatory disclosure. This allows us to connect the literature on the effect of information in sequential social dilemmas to the experimental literature on disclosure.

The bulk of the experimental literature on disclosure uses sender–receiver games: a sender privately observes the realization of a random variable, and chooses whether to disclose it to a receiver. The receiver is then asked to report it or, if the sender does not disclose it, to guess it. Payoff functions induce a disclosure: The sender earns more money when the receiver guesses a higher value while the receiver has incentives for accuracy. Following the logic of unraveling, the sender should disclose any realization of the random variable, except for the most unfavorable ones. The main results of this literature are that, while some unraveling is observed, it is often far from complete (see e.g. Jin et al. 2021; Montero and Sheth 2021; Sheth 2021; Hagenbach and Saucet 2022): Receivers are not skeptical enough about undisclosed information and senders often withhold intermediate information. This literature also identified conditions that are more favorable to unraveling. In particular, Jin et al. (2021) show that systematic feedback and experience allow converging to the theoretical predictions. Sheth (2021) shows that competition between senders reduces the extent of the failure of unraveling. In particular, she shows that senders are more likely to disclose intermediate information under competition. On the other hand, receivers' naivety is not impacted by competition.[7]

In our experiment participants disclose their past actions instead of the realization of a random variable. As a consequence, they have to devise at the same time their cooperative and information disclosure strategies. Our results suggest that combining both decisions increases unraveling and reduces the negative effect of the failure of unraveling. This is consistent with the findings of Harrs et al. (2022), who conducted experiments showing that markets where producers can choose their level of social responsibility (represented as a charitable gift) and decide whether to disclose it to

[5] A related literature study the effect of information about choices on cooperation in simultaneous interactions. See for instance Duffy and Ochs (2009), Camera and Casari (2009), Kamei and Putterman (2016).

[6] Duffy et al. (2013) test the effectiveness of information disclosure in an indefinite-horizon game where full cooperation can be supported in equilibrium, with similar results. Cox et al. (2015) also find a positive effect of information about past behavior in a finitely-repeated prisoners dilemma with partner matching.

[7] In a different setup, Penczynski et al. (2022) confirm the effect of competition on senders but surprisingly show that competition can even increase receivers' naivety.

consumers, are equally efficient as markets where social responsibility is chosen but disclosure is mandatory. Importantly, in our settings, very little time and feedback are necessary to converge to unraveling.

To summarize, our main contribution is to connect these two streams of the literature and show that the failure of unraveling generally found in disclosure games does not necessarily lead to voluntary disclosure being less effective than mandatory one when considering social dilemmas. We are aware of only two papers comparing automatic and voluntary disclosure in a social dilemma (Kamei 2017, 2020). Both focus on simultaneous games, while we use a sequential prisoner's dilemma. Sequentiality is common in real-life economic interactions, including online transactions (the buyer must pay before the seller sends the product) or employer-employee relations. In contrast to simultaneous games, in a sequential game all the strategic uncertainty weights on the first mover (as noted by e.g. Ghidoni and Suetens 2022). In SPD, P2 does not have to form a belief about P1's choice, because P1's choice is already known to P2 when P2 makes his own choice. Hence, any cooperation choice of previous P1s should be irrelevant for P2's choice in a particular period. This in turn implies that in our game, P2's record reveals all the information P1 might need, while in the simultaneous game, P1 would also have to know the cooperation choices of P2's prior partner to interpret P2's record correctly. Therefore, in both Kamei (2017, 2020), the information about one's opponent is more complex to interpret. Both papers find that, under random matching, the possibility to freely hide one's identity (Kamei 2017) or one's last action (Kamei 2020) undermines the effect of information disclosure on cooperation. In contrast, our results indicate that voluntary disclosure may be just as effective as mandatory one, in a context in which the interpretation of the disclosed information is relatively straightforward.

3 Design

3.1 The stage game

The stage game is a modified sequential prisoner's dilemma. Compared to a regular sequential prisoner's dilemma the main modification is that in each round P1 can choose whether she enters the game or not. If she decides not to enter, both players get an outside option with a fixed and equal payoff. We introduced the entry option since the possibility to refuse any interaction is a feature found in many real-life situations.[8] For instance, in online markets, customers can avoid suppliers with poor reviews. In the labor market, employers can avoid interviewing candidates based on their application files.[9] The modified sequential prisoner's dilemma is shown in Fig. 1.

P2 plays the strategy method: he takes one decision for the node after P1 decided to enter and defect, and one decision for the node after P1 decided to enter and cooperate.

[8] Relatedly, Feess and Kerzenmacher (2023) in this journal study a situation in which the *second mover* can chose to refuse interactions.

[9] During the pandemic, we collected data online with a slightly modified design in which we dropped the entry decision. The design and the data are described in Appendix D. The overall message is that the entry decision does not impact significantly the behavior in the game.

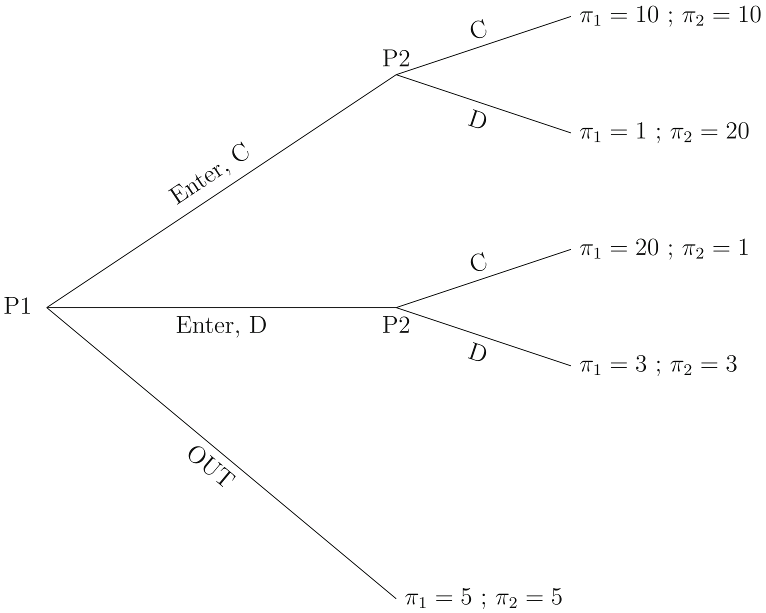

P1
Enter, C
Enter, D
OUT
P2
P2
C
D
C
D
$\pi_1 = 10$; $\pi_2 = 10$
$\pi_1 = 1$; $\pi_2 = 20$
$\pi_1 = 20$; $\pi_2 = 1$
$\pi_1 = 3$; $\pi_2 = 3$
$\pi_1 = 5$; $\pi_2 = 5$

Fig. 1 The stage game

These two choices are made before P2 gets informed about P1's actual choice. The use of the strategy method allows to gather more data and to ensure that the history of play of P2 does not depend on her past partners' behavior. The use of the strategy method is common, and while it may affect behavior (see e.g. Casari and Cason 2009), it generally does not affect treatment effects (Brandts and Charness 2000, 2011).

The parameters were chosen to make cooperation hard to achieve: the temptation to defect for P2 is high, which is expected to deter entry and cooperation of P1 (e.g. Mengel 2017; Gaechter et al. 2022, study the impact of game parameters on cooperation in the prisoner's dilemma). In addition, our parameters make mutual cooperation inefficient, which is expected to further reduce cooperation (see the "temptation treatments" in Clark and Sefton 2001).[10]

3.2 Matching

The stage game is played for 30 periods, which is common knowledge. Roles are fixed: each player is randomly assigned to be P1 or P2, once and for all before the first period. At every period, each P1 is matched with a randomly selected P2 (stranger

[10] Eliciting beliefs of P1s and P2s would have been interesting to explore the mechanism underlying behavior. We chose not to elicit beliefs to keep the experiment simple and because it can affect behavior in strategic games (Gächter and Renner 2010).

Table 1 Summary of the treatments with record disclosure

Record disclosure is:	Mandatory	Voluntary
Certain	MandNN	VolNN
Noisy	MandLN	VolLN

matching).[11] At the end of each period, participants were informed of their earnings for that particular period. Participants are paid for every period.

3.3 Experimental treatments

There are 5 experimental treatments. The **Baseline treatment** is fully described by the stage game and the matching introduced earlier. Besides the **Baseline treatment**, we have 4 treatments in which information about P2 may be revealed to P1 before she decides to enter. We call this information *the record* of P2. The record contains two pieces of information about P2. First, it reports the portion of times that P2 chose "Cooperate" at the Enter Cooperate node. Second, it reports the portion of times that P2 chose "Cooperate" at the Enter Defect node. Remember that P2s' choices are elicited using the strategy method. Therefore, the observed decision of a P2 in a particular period is independent of the actual choice of her P1 in this period and the record of a P2 is independent of the actual choices of the P1s he was matched with in past periods. Figure 9 in Appendix A gives an example of a record as disclosed to P1.

The 4 information treatments differ in the way the record is disclosed to P1. There are two dimensions: First, record disclosure might be mandatory (the "Mand" treatments), or it might be P2's choice whether to disclose his record or not (the "Vol" treatments). Second, we vary whether disclosure is noisy: in the LN treatments, there is a 10% chance that the record is not transmitted to P1 when information transmission is mandatory (in the MandLN treatment), or when P2 chooses to reveal his record (in the VolLN treatment). In the NN treatments, no such noise is introduced. One of our motivations for introducing noise was to test the robustness of our results to small imperfections of the information mechanism. We chose 10% because it is a round number that subjects can understand relatively easily, small enough that one could expect the results to hold, but also large enough that subjects do not completely disregard it when making their decisions. These variations result in 5 treatments: **Baseline, MandNN, MandLN, VolNN, VolLN**. Table 1 summarizes the information treatments.

[11] Given the size of our sessions, participants meet more than once on average. This could in principle affect the evolution of cooperation through contagion. Ghidoni et al. (2019) show that random matching in groups of 6 participants give similar results to perfect stranger's matching. In all our sessions we had more than 6 participants. In addition, all our results hold when we control for the number of participants in the respective session.

Table 2 Summary of the sessions

	Total number of Sessions (in Lille)	Total number of Participants (in Lille)
Baseline	5 (3)	84 (52)
MandNN	5 (2)	74 (38)
VolNN	5 (2)	84 (42)
MandLN	5 (3)	76 (56)
VolLN	5 (1)	68 (20)
Total	25 (11)	386 (208)

Numbers outside of parentheses are totals. Number in parentheses are for the sessions run in Lille

3.4 Procedure

We recruited a total of 386 participants for 25 sessions.[12] 14 sessions were run at the BEEL lab in Brussels in winter 2019/2020 (before the pandemic), while 11 sessions were run at the Anthropo-lab in Lille in fall 2021 (after the pandemic).

Table 2 sums up the distribution of participants and sessions across treatments and cities. Our results are robust to the inclusion of a dummy variable indicating the city in which the session was run.

On average a session lasted about 1 h. The average earnings were €12.87 (S.D.: 2.37). Before starting the experiment, participants had to pass a non-incentivized understanding questionnaire.[13] At the end of the experiment, participants had to fill up a demographic questionnaire.[14] The full set of instructions can be found in Appendix I.

4 Results

Table 3 reports the descriptive statistics by treatment.[15]

The cooperative outcome is relatively rarely observed. This was expected because our parameters make defection tempting. P1s enter in the vast majority of the cases but cooperate much less often. P2s cooperate much more often following cooperation than

[12] A preregistration of the experiment can be found at https://doi.org/10.1257/rct.4937. Note that, due to constraints related to the pandemic, we had to drop two treatments in which a larger noise was introduced. We have a power of 80% to detect an effect of 10 percentage points at the 5% level (more details are in Appendix C). This effect size is less than half the size of the effect of disclosing the second movers' history reported in Charness et al. (2011).

[13] Participants could retake the questionnaire until they passed. In case of mistakes, a prompt asked them to reread the instructions. When necessary, some general explanations were provided by the experimenter. At the end of the experiment, we asked participants to report on a 5-point Likert scale whether it was difficult for them to understand the instructions. Less than 10% reported that it was "difficult" to understand the instructions. More details are given in Table 9 in Appendix B.1.

[14] Table 10 in Appendix B.2 reports the demographic characteristics across cities. Note that the distribution of demographic variables differs across cities, so we control for it in our regressions.

[15] In Table 21 in Appendix G.1, we report the dynamics of the Cooperative outcome for each session separately.

Table 3 Descriptive statistics

Treatment	Coop. outcome	P1 enters	P1 coop.	P2 coop. if:		P2 discloses
				P1 coop.	P1 def.	
Baseline	0.110	0.686	0.289	0.358	0.179	–
MandNN	0.264**	0.801**	0.420**	0.533**	0.159	–
VolNN	0.236**	0.769	0.367	0.530*	0.140	0.619
MandLN	0.157	0.740	0.326	0.365	0.163	–
VolLN	0.213	0.742	0.334	0.473	0.140	0.556

Test Baseline versus respective information treatment based on Logit regressions with standard errors clustered at the session level using bootstrapping

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Table 14 in Appendix E provides details

following defection but still cooperate following defection by P1 in roughly 15% of the cases, which is in line with Miettinen et al. (2020). One may still fear that irrationality plays a big role in both "over-entry" by P1s and cooperation of P2s following defection by P1, but there are alternative explanations. Over-entry makes sense if P1s believe that P2s are concerned about social welfare or if they expect that P2s make errors with a small probability. Cooperation after defection can be explained if we assume that some P2s have social welfare concerns. This is consistent with previous results in the sequential prisoner's dilemma (see e.g. Miettinen et al. 2020). A careful analysis of this question is reported in Appendix H, the result of which is that irrationality is not a big concern.

The dynamics of the cooperative outcome is not impacted by the presence of a reputation system. There is a decline in the likelihood of the cooperative outcome, but it is small and similar with or without record disclosure. See Appendix G.2 for more details.

4.1 Cooperation

Result 1: The existence of a record system increases the likelihood of the cooperative outcome. Voluntary record disclosure is as effective as mandatory disclosure. On the other hand, introducing a small noise decreases the effectiveness of record disclosure slightly.

Support for Result 1: The cooperative outcome occurs in 11% of cases in the baseline, and in 21.7% of cases in the other treatments ($p < 0.05$, see Tables 15 and 16 in Appendix E for details). To disentangle the respective effect of voluntary disclosure and noise, we run Logit regressions explaining the occurrence of the cooperative outcome by dummy variables indicating a voluntary disclosure system and a noisy disclosure system, and the interaction of these two dummies. Standard errors are clustered at the session level using bootstrapping.[16] We control for the presence of a record system

[16] For every bootstrap analysis, we performed 400 bootstrap replications, following the recommendation of Cameron and Trivedi (2010).

Table 4 The effects of voluntary and noisy disclosure on the occurrence of the cooperative outcome

	(1)		(2)		(3)	
	Coop. outcome = 1		Coop. outcome = 1		Coop. outcome = 1	
	coeff.	m.e.	coeff.	m.e.	coeff.	m.e.
Record	1.069***		1.084***		1.040**	
	(0.389)		(0.383)		(0.464)	
Voluntary	−0.151	0.006	−0.153	0.006	−0.131	0.001
	(0.284)	(0.036)	(0.296)	(0.036)	(0.358)	(0.043)
Noisy	−0.655**	−0.061*	−0.662**	−0.061*	−0.605*	−0.063*
	(0.269)	(0.033)	(0.266)	(0.033)	(0.360)	(0.038)
Voluntary × Noisy	0.523		0.528		0.385	
	(0.487)		(0.483)		(0.562)	
Observations	5780		5780		5780	
Session caracteristics	No		No		Yes	
Period FE	No		Yes		Yes	
Sessions	25		25		25	

Standard errors (in parentheses) are clustered at the session level using bootstrapping
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$
The level of observation is one interaction. All treatments are included
Session characteristics: City dummy and size of the session
Individual characteristics: Gender, age, occupational status, experience with experiments

in every model. The outcomes of these regressions are reported in Table 4. For each model, the first column reports the coefficients from the Logit regression, and the second column reports the marginal effects of interest.

Since we used the strategy method, we observe the four pure strategies of P2s: (i) Cooperate when P1 cooperated (conditional cooperation), (ii) Cooperate irrespective of the decision of P1 (unconditional cooperation), (iii) Never cooperate (unconditional defection) and (iv) Cooperate only when P1 defected (mismatch). In what follows, to study cooperation choices of P2, we pool together conditional and unconditional cooperation. Focusing strictly on conditional cooperation does not change our results. More details about the distribution of strategies across treatments are given in Appendix G.3.

Result 2: When there is no noise, the existence of a record system increases cooperation levels of the P1s and P2s irrespective of whether disclosure is mandatory or voluntary. With noise, the record system has no significant effect on cooperation. The effect of a record system on the enter decision is small and weakly significant. Voluntary disclosure and noise do not impact the effect of the presence of a record system on the enter decision.

Support for Result 2: Table 3 provides first evidence in favor of Result 2. To disentangle the respective effects of noise and voluntary disclosure, we run Logit regressions explaining the occurrence of entry, and cooperation of P1 and P2 by dummy variables indicating a voluntary disclosure system and a noisy disclosure system and the

Table 5 The effects of voluntary and noisy disclosure on P1s' and P2s' individual decisions

	(1)		(2)		(3)	
	Enter = 1		P1's Coop = 1		P2's Cond Coop = 1	
	coeff.	m.e.	coeff.	m.e.	coeff.	m.e.
reputation	0.679*		0.643**		0.640*	
	(0.384)		(0.253)		(0.354)	
Voluntary	−0.252	−0.020	−0.249	−0.033	−0.114	0.017
	(0.322)	(0.052)	(0.254)	(0.044)	(0.457)	(0.077)
Noisy	−0.426	−0.051	−0.449**	−0.074*	−0.686**	−0.109*
	(0.384)	(0.051)	(0.213)	(0.040)	(0.310)	(0.057)
Voluntary × Noisy	0.375		0.260		0.506	
	(0.512)		(0.368)		(0.552)	
Observations	5780		5780		5780	
Session caracteristics	Yes		Yes		Yes	
Individual caracteristics	Yes		Yes		Yes	
Period FE	Yes		Yes		Yes	
Sessions	25		25		25	

Standard errors (in parentheses) are clustered at the session level using bootstrapping

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$

The level of observation is the individual decision (of P1s in Column (1) and (2) and of P2s in Column (3)

All treatments are included

Session characteristics: City dummy and size of the session

Individual characteristics: Gender, age, occupational status, experience with experiments

interaction of these two dummies. Standard errors are clustered at the session level using bootstrapping. The outcomes of these regressions are reported in Table 5. For each model, the first column reports the coefficients of the regression and the second column reports the marginal effects of interest. The marginal effect of voluntary disclosure is never significant. The marginal effect of noise on entry is not significant, but it is marginally significant on P1s' and P2s' cooperation levels: noise reduces P1s' cooperation by about 7 percentage points ($p = 0.083$) and P2s' conditional cooperation by about 10 pp ($p = 0.063$).

4.2 Disclosure, record and skepticism

Noise has no effect on the disclosure decisions of P2s. As one can see from Table 3, the disclosure rates are very similar in both voluntary disclosure treatments. When disclosure is voluntary, the disclosure decision depends mainly on P2's record.

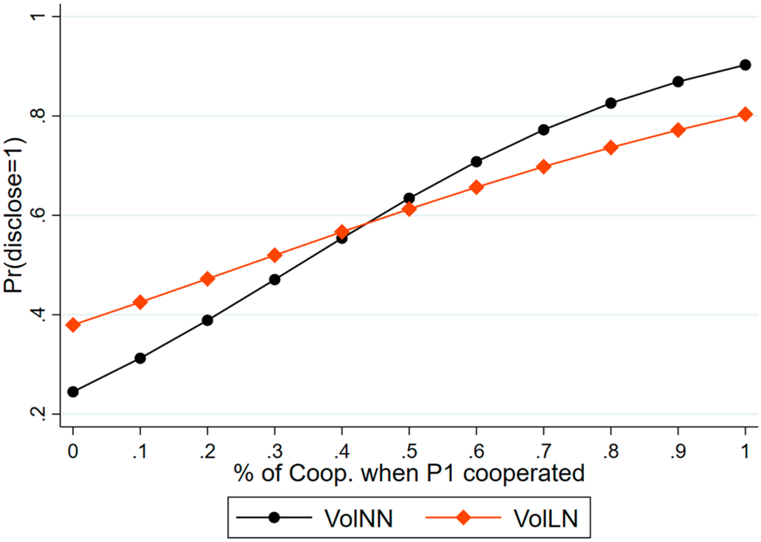


Fig. 2 Predicted probability to disclose depending on one's past record

Result 3: P2s with better records are more likely to disclose it. This is true for both the VolLN and the VolNN treatment. Importantly, there is no time trend in disclosure behavior: P1s are equally likely to disclose in the first periods and in the last periods.

Support for Result 3: We estimated Logit models explaining the decision of a P2 to disclose his record by the rate at which he cooperated following cooperation by P1 and following defection in the past. We use only the data from the Vol treatments. Table 17 in Appendix E.3 reports the outcome of these regressions. We use the results from these regressions to compute predicted probabilities of disclosure depending on one's record, by treatment. These probabilities are represented in Fig. 2. The probability of disclosure increases in the level of cooperation following cooperation of P1 in the past.

Figure 3 shows the disclosure rate of P1s across periods and strongly supports that disclosure does not change over time.

Result 4: P1s enter and cooperate more when they see the record of P2. In addition, a record indicating that P2 cooperated more in the past, irrespective of whether this follows cooperation by P1 or not, increases P1's likelihood to enter. More specifically, a record of more cooperation of P2 following cooperation of P1 increases P1's likelihood to cooperate, while a record of more cooperation after defection leads to less cooperation of P1.

Support for Result 4: Figure 4 shows the entry and cooperation rates of P1s depending on whether P1s saw the record of P2, separated by treatments. Overall, seeing the record increases both entry and cooperation. Pooling all the data from the different treatments, P1s' cooperation rate is about 16 percent higher when P1s saw the record ($p < 0.001$, see Table 18 in Appendix E.4 for details on the tests). The positive effect

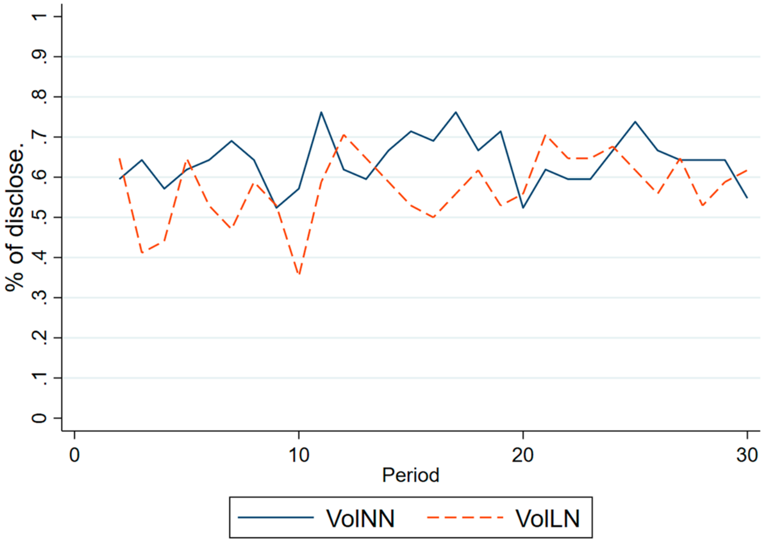


Fig. 3 The dynamic of disclosure (rate averaged at period level)

of seeing the record on cooperation is seen in all three treatments where the disclosure is neither excluded nor guaranteed by design. The effect of seeing the record on entry decisions is also significant overall, albeit somewhat smaller in magnitude.

We run Logit regressions on the subset of the data in which the record of P2 is displayed to P1. We explain the decision of P1 to cooperate depending on the record of P2, i.e. on the portion of times P2 chose cooperation in the past. Standard errors are clustered at the session level using bootstrapping. We present the outcome of several models in Table 6. For all specifications, the coefficient and marginal effects of P2s' previous cooperation are significant at the 1% level. Note that, in contrast with all other parameters, the marginal effect of previous cooperation following defection on cooperation by P1 is negative. This makes intuitive sense: the more likely P2 is to cooperate even after P1s' defection, the more likely P1 is to enter and defect. In the two rightmost columns, we present the marginal effect of our variables of interest over the different period brackets. The effects described above are present in every period bracket but are stronger in the last period bracket than in the first one. This suggests that P1s become more sensitive to the revealed information over time.[17]

Figure 10 in Appendix E.4 reports the predicted probability of cooperation by P1, depending on the record of P2, for each treatment with information disclosure (confidence intervals are omitted for readability). It shows that the better the record, the higher the probability of cooperation by P1. This does not change across treatments.

Result 5: P1s show skepticism: In the Vol treatments P1s cooperate less when they do not see the record of P1 than when they see it. This is also the case in the MandLN

[17] We thank a referee for suggesting this analysis.

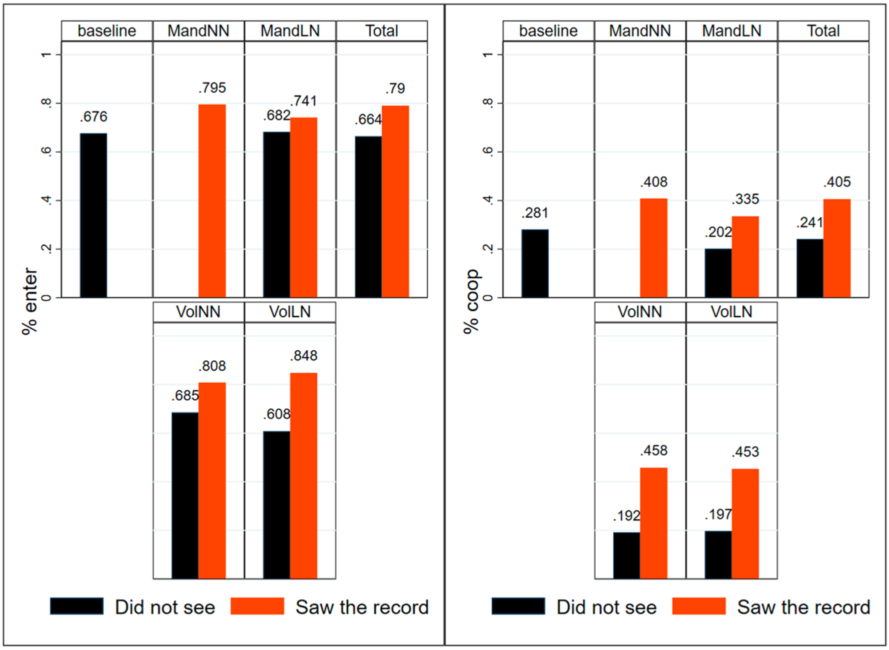


Fig. 4 Entry (left panel) and Cooperation (right panel) rates depending on whether P1 saw the record, separated by treatments

treatment but to a much smaller extent. Skepticism is observed already in the first periods of the game, and limited learning takes place.

Support for Result 5: We already checked that P1s are more likely to cooperate when they see the record of P2. We expect that the impact of non-disclosure on P1s' cooperation is weaker when non-disclosure cannot be attributed to a deliberate decision of P2, since in this case non-disclosure cannot serve as a signal of a poor record.

Non-disclosure reduces P1s' cooperation rates by 10 percentage points (from 37.3 to 27.3%) in the Mand treatments where non-disclosure cannot be voluntary (See Fig. 11 in Appendix E.5). In the Vol treatments, where non-disclosure is mainly due to deliberate decisions of the P2s, the non-disclosure induced a drop in P1s' cooperation rates more than twice as large, namely 26.2 percentage points (from 45.6 to 19.%). Note that a similar pattern is found for entry decisions (left panel of Fig. 11). This is a first indication of the validity of Result 5. We run Logit models explaining P1's decision to cooperate by a dummy variable indicating voluntary disclosure and a dummy variable indicating that P1 saw the record, and interaction of these two dummies. Standard errors are clustered at the session level using bootstrapping. Table 7 shows the outcomes of these regressions. In the second column of Table 7, we report the marginal effect of seeing P2s' records when disclosure is mandatory and when disclosure is voluntary. The marginal effect of seeing the record is significantly greater in the Voluntary treatment (+26.5pp against +11.2pp, $p = 0.004$), which confirms the result suggested by Fig. 11: the effect of (not) seeing the record is much weaker when disclosure is for

Table 6 The effect of the record of P2 on the decision to enter and cooperate by P1

	(1) Enter = 1		(2) Coop. = 1		(3) Enter = 1	(4) Coop. = 1
	coeff.	m.e.	coeff.	m.e.	m.e.	m.e.
% of Coop. of P2 when P1 coop'd	2.883****	0.407****	2.444****	0.506****		
	(0.381)	(0.052)	(0.188)	(0.033)		
% of Coop. of P2 when P1 def'd	1.403****	0.198****	−1.183****	−0.245****		
	(0.298)	(0.040)	(0.244)	(0.047)		
% of Coop. of P2 when P1 coop'd for:						
Period ≤ 10					0.221****	0.342****
					(0.042)	(0.045)
10 < Period ≤ 20					0.419****	0.551****
					(0.069)	(0.057)
Period > 20					0.514****	0.542****
					(0.062)	(0.060)
% of Coop. of P2 when P1 def'd for:						
Period ≤ 10					0.082**	−0.134***
					(0.041)	(0.051)
10 < Period ≤ 20					0.194****	−0.287****
					(0.050)	(0.061)
Period > 20					0.340****	−0.521****
					(0.079)	(0.070)
Observations	3269		3269			
p value diff.		< .001		< .001		
Sessions Char.	Yes		Yes			
Individual Char.	Yes		Yes			
Period F.E.	Yes		Yes			

Standard errors clustered at the session level using bootstrapping. $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$

In Columns (3) and (4), p values are from tests comparing the marginal effect in the first and last period brackets

The level of observation is individual decision of P1s

Data from the baseline and observations in which P1 did not see the record of P2 are excluded

Session characteristics: City dummy and size of the session

Individual characteristics: Gender, age, occupational status, experience with experiments

sure not voluntary. In addition, in the last column, we report the marginal effects of our variables of interest over the different period brackets. The values of the marginal effects suggest that the effect of (not) seeing the record is already present early in the game. Importantly, when disclosure is Voluntary, the effect of (not) seeing the record does not increase significantly across period brackets (none of the pairwise comparisons of the marginal effects for different time brackets in the Voluntary is significant at the 10% level.).

These results show that P1s are skeptical about non-disclosed information. A complementary question is whether skepticism is well-calibrated.[18] We focus on the data from the VolNN treatment, which provides the best setup to test the calibration of skepticism. In this treatment, P1 cooperated 19.1% of the time when P2s did not disclose. P2s who did not disclose cooperated 35% of the time (following cooperation by P1s). P1s who were informed that their respective P2 cooperated between 30% and 40% of the times cooperated 28% of the times. This suggests that, if anything, P1s were too skeptical. Note however that the results from the first periods of the game might not be very reliable/relevant, because the support of the information is mechanically limited. For instance, in period 2, information can take only 2 values (0% or 100%), in period 3, information can take only 3 values (0%, 50% or 100%), etc. If we focus on periods 10 to 30, P1 cooperated 17.7% of the time when P2s did not disclose. P2s who do not disclose cooperated 34% of the time (following cooperation by P1s). P1s who were informed that their respective P2 cooperated between 30% and 40% of the times cooperated 20.9% of the time (57 observations). Our interpretation is that skepticism was overall well-calibrated.

5 The impact of short-run reputation

Our results are surprising, as they run counter to our conjecture based on past experimental results on the failure of unravelling. To assess the robustness of and the reasons for this surprising outcome, we ran two additional treatments, the Short Run reputation treatments (SR).[19] The MandSR is exactly the same as MandNN, except that only P2's strategy choice in the previous period is revealed to P1. Similarly, in the VolSR, P2 can choose to reveal his previous period strategy choice to P1. In the SR treatments, the information disclosed is simpler, and therefore closer to what is disclosed in sender–receiver games [see e.g. Jin et al. (2022) for a study of the role of information complexity on disclosure]. In addition, since information does not carry over across rounds, these treatments allows for experimentation early in the game, which is also the case in sender receiver games. We ran 4 sessions in each treatment in May 2023 in Lille, with a total of 162 participants (82 in the MandSR and 80 in the VolSR).

Table 8 reports the descriptive statistics for the new experiment. We include the data from the baseline, MandNN and VolNN for comparison. The aggregate data are overall very similar in the new treatments compared to those from the main experiment. Note that short-run reputation seems to be slightly less effective than long-run reputation in promoting cooperation, which is consistent with past literature (Keser 2003; Duffy et al. 2013).

[18] We thank a referee for suggesting this analysis.

[19] Note that these treatments were not preregistered.

Table 7 The differential effect of (not) seeing the record, depending on the treatment

	(1) Coop. = 1 coeff.	m.e.	m.e.
Mandatory (base)			
P1 saw the record	0.520^{**}		
	(0.218)		
Voluntary	-0.443^{*}		
	(0.231)		
P1 saw the record × Voluntary	0.762^{***}		
	(0.262)		
Constant	−0.782		
	(0.695)		
Marginal effect of seing the record at:			
Mandatory		0.112^{**}	
		(0.046)	
Voluntary		0.265^{****}	
		(0.034)	
p value diff.		0.004	
Marginal effect of seing the record at:			
Mandatory × Period $\leq$ 10			0.049
			(0.044)
Mandatory ×10 < Period $\leq$ 20			0.171^{****}
			(0.050)
Mandatory × Period > 20			0.107^{*}
			(0.064)
Voluntary × Period $\leq$ 10			0.223^{****}
			(0.046)
Voluntary ×10 < Period $\leq$ 20			0.280^{****}
			(0.035)
Voluntary × Period > 20			0.292^{****}
			(0.055)
Observations	5587		
Sessions Char.	Yes		
Individual Char.	Yes		
Period F.E.	Yes		

Standard errors clustered at the session level using bootstrapping. $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$

The level of observation is individual decision of P1s. All treatments are included

Session characteristics: City dummy and size of the session

Individual characteristics: Gender, age, occupational status, experience with experiments

Table 8 Descriptive statistics of the new experiments

Treatment	Coop. outcome	P1 enters	P1 coop.	P2 coop. if:		P2 discloses
				P1 coop.	P1 def.	
Baseline	0.110	0.686	0.289	0.358	0.179	–
MandNN	0.264****	0.801*	0.420***	0.533**	0.159	–
VolNN	0.236**	0.769*	0.367*	0.530*	0.140	0.619
MandSR	0.232**	0.670	0.406**	0.529**	0.113*	–
VolLNSR	0.190	.655	0.361	0.498*	0.080***	0.565

Test Baseline vs respective information treatment based on Logit regressions with standard errors clustered at the session level using bootstrapping

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Table 19 in Appendix F provides details

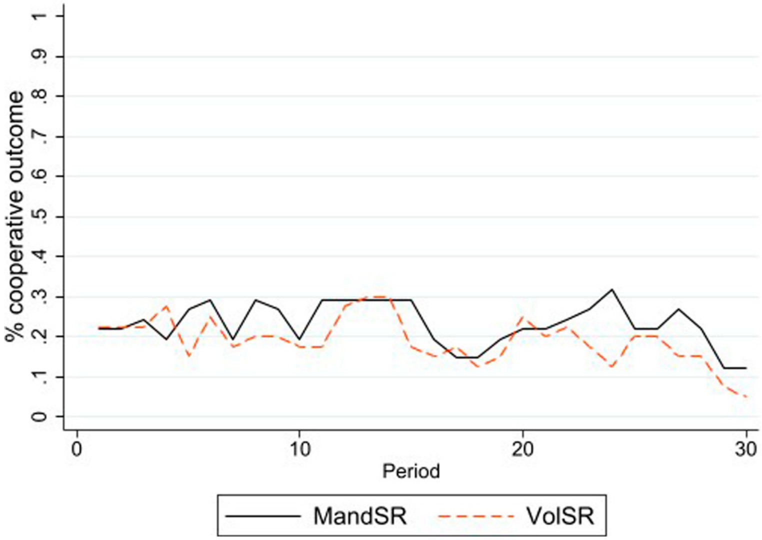


Fig. 5 Cooperative outcomes across periods for the SR treatments

The main result is that voluntary disclosure is as effective as mandatory one even with short-run reputation. In addition, Fig. 5 shows the rate of cooperative outcome in the SR treatments across periods. There is no difference in the time trend between MandSR and VolSR. The possibility to experiment early in the game does not make our results more consistent with the failure of unraveling observed in sender–receiver games. We conclude that our main result is robust to this change in the design.

We now turn to individual decisions. Figure 6 shows the (predicted) probability of cooperation by P1 depending on whether P1 saw the record of P2 in the VolSR treatment, across periods. The regression model used to compute the predicted probabilities is reported in Column (1) of Table 20 in Appendix F. As expected, P1s are much more likely to cooperate when they see the record of P2. Importantly, this is the case already in the very first periods.

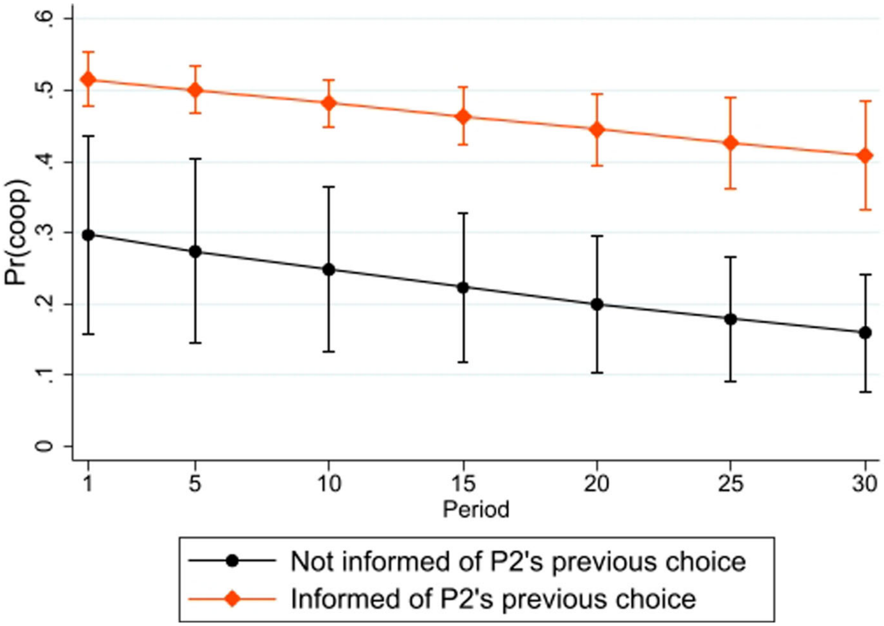


Fig. 6 Probability that P1 cooperates, depending on whether s/he's informed of P2s previous choice, over periods. Vertical lines are 95% confidence intervals

Figure 7 shows the probability that P1 cooperates depending on P2's last choice across periods. We limit the analysis to data in which the information is disclosed to P1 from both SR treatments. The regression model used to compute the predicted probabilities is reported in Column (2) of Table 20 in Appendix F. P1s are much more likely to cooperate when they are informed that their current counterpart cooperated in the previous period. This is the case already in the first periods, and it is consistent with the results from the main experiment.

Figure 8 shows P2s' likelihood to disclose their last choices depending on their choices in the previous period, across periods. We limit the analysis to the VolSR treatment. The regression model used to compute the predicted probabilities is reported in Column (3) of Table 20 in Appendix F. We find that P2s are much more likely to disclose their last choice when they cooperated. This is already true in the very first periods of the game. There is nonetheless some evidence that participants who defected become less likely to disclose through periods, while the opposite holds true for P2s who cooperated in the previous period. This suggests that some learning/experimentation takes place but it is still limited in scope.

Taken together, the data from the SR treatments show the robustness of our main results. Short-run reputation works and generates similar individual behavior. What is interesting is that the data from the SR treatment discards the learning/experimenting explanation for the discrepancy between our results and the failure of unraveling often observed. It suggests that we identified an interesting pattern, namely that a voluntary

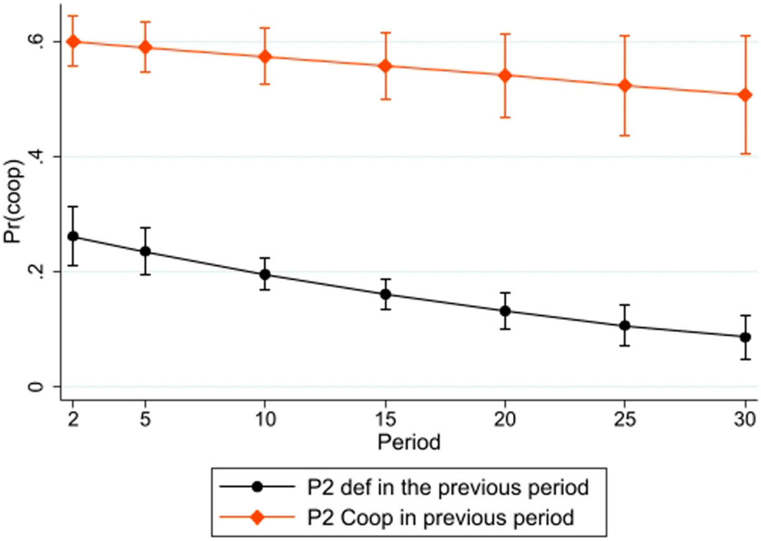


Fig. 7 Probability that P1 cooperates when informed, depending on P2s previous choice, over periods. Vertical lines are 95% confidence intervals

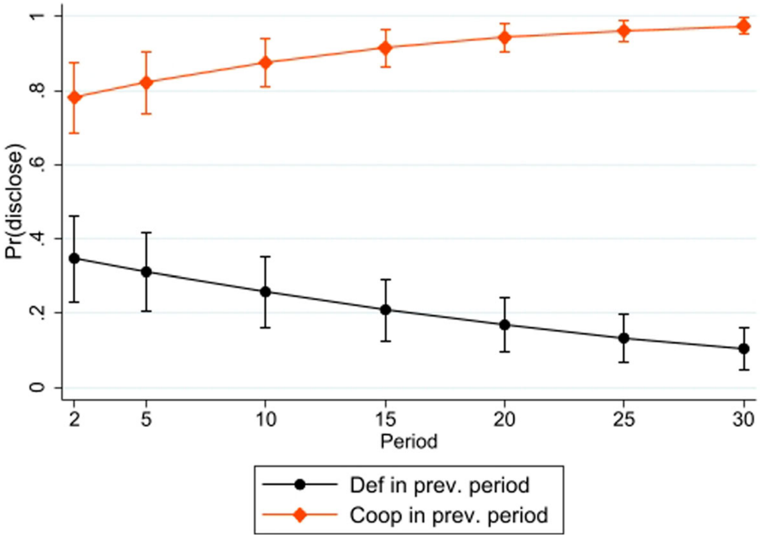


Fig. 8 Likelihood to disclose depending on P2's previous action, over periods. Vertical lines are 95% confidence intervals

disclosure mechanism may be as effective as a mandatory one when individuals reveal their past choices.

6 Discussion and Conclusion

We studied the effect of different information-disclosure mechanisms on cooperation in a sequential social dilemma. We found that such institutions increase the likelihood of reaching the cooperative outcome, as both the first and second movers act more cooperatively. In addition, we found that an institution in which disclosure is voluntary is exactly as effective as one in which it is mandatory. In contrast, we found that a relatively small noise undermines the effect of information disclosure independently of the voluntary aspect.

A striking result is that voluntary disclosure is exactly as effective as mandatory disclosure. This result is consistent with unraveling, but it is in contrast with two previous experiments comparing voluntary and automatic disclosure of information in social dilemma Kamei (2017, 2020). As explained earlier, in our experiment sequentiality, combined with the strategy method makes the interpretation of one's record much more straightforward, which might foster the incentives to behave as a cooperative type.

More importantly, our results also contrast with the literature on the failure of unraveling in sender–receiver games (see e.g. Jin et al. 2021; Montero and Sheth 2021; Sheth 2021). In this literature, senders typically send favorable information, but withhold unfavorable ones, exploiting the naivety of (some of) the receivers. This failure of unraveling is corrected only after some learning has occurred. This suggests that voluntary disclosure should be significantly less effective than mandatory one, if effective at all. In our experiment, when given the choice, a large share of the subjects choose to maintain a good record and disclose it (see Fig. 18 in Appendix G.4.). First movers are skeptical: they generally avoid exposing themselves to being exploited by a second mover who does not disclose his record. This is observed already early in the game, and we observe no significant dynamics in the effectiveness of voluntary disclosure.

What can explain this difference between our results and those of sender–receiver game experiments? In our main experiment, P2's record depends on all his previous choices. Hence P2 cannot hope to get a clean record after defection. In contrast, in the sender–receiver game, the game starts from scratch in every period, which allows for substantial experimentation early in the game, which could in turn lead to the observed dynamic of unraveling. We investigated this hypothesis by running new treatments in which the record is only about P2's last choice, which makes experimentation possible. We found that the "memory" of the records does not explain our main results. Our interpretation is that the extent of the failure of unraveling depends on the *nature* of the information disclosed. In sender–receiver games, participants disclose a random variable that is exogenously determined, while participants in our game disclose their past actions. It's possible that having participants disclose their past actions makes them think differently about disclosure than reporting a random variable. In particular, the disclosed information in our experiment has a signaling value. Building and sustaining a good reputation may have an intrinsic value for participants in a social dilemma game because a good record signals to oneself as well as to others that one is a cooperative, "nice" type (see e.g. Bénabou and Tirole 2006, and the large subsequent literature on identity management). This interpretation is consistent with Harrs et al. (2022), who finds that voluntary disclosure promotes market efficiency as effectively as mandatory one when the disclosed information reveals the social responsibility of producers.

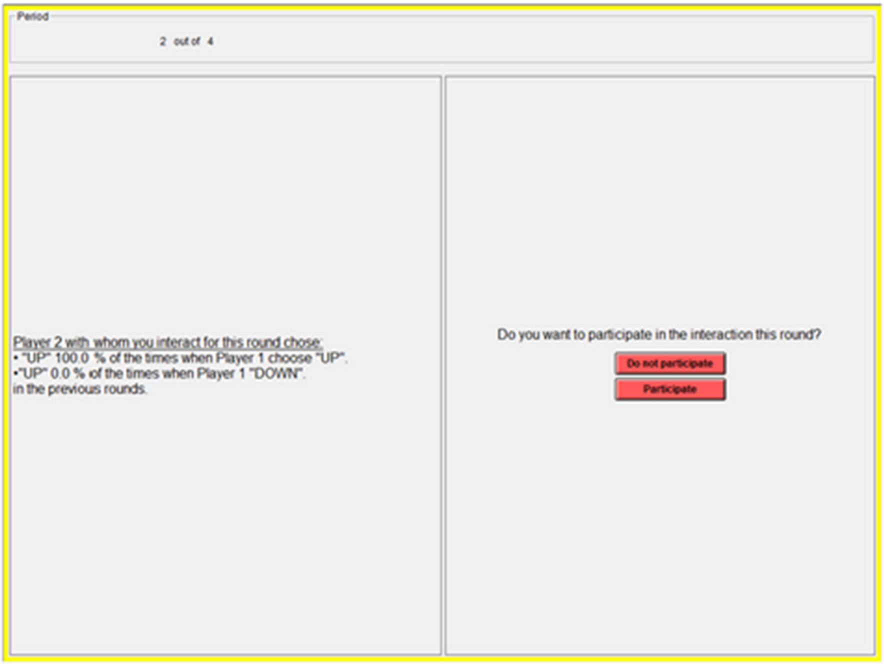
Irrespective of the fundamental reasons, our results suggest that the limits of unraveling identified in past sender–receiver experiments might depend on the nature of the strategic situation and the type of information that is disclosed. Our results should be confirmed experimentally with different parameters, and across different underlying games.

Another noteworthy finding is the impact of noise: the introduction of a relatively small noise diminishes the effectiveness of information disclosure. We designed the treatment with noise with two goals in mind: assessing the robustness of the effect of information disclosure, and providing a placebo test for skepticism (comparing the behavior of P1 when not seeing the record can vs cannot be blamed on luck). The effect of the relatively small noise is consistent with probability weighting: Individuals tend to overweight small probabilities and overestimate the likelihood of unlikely events (Tversky and Kahneman 1992). In our setting, the second movers might overweight the likelihood that their record remains hidden, reducing their incentives to cooperate. This in turn might lead to a lower effectiveness of noisy disclosure in sustaining mutual cooperation.

It would be interesting to see whether agents prefer mandatory disclosure or voluntary disclosure mechanisms in dilemma situations. On the one hand, the experimental results suggest that mandatory disclosure is never worse and in simultaneous games even better than voluntary disclosure. This should induce agents to opt for a mandatory disclosure mechanism for general dilemma situations. On the other hand, it is well known that humans like (the illusion of) control, and this motivation is of course more in line with voluntary disclosure.

A Screenshots

See Fig. 9.



Period

2 out of 4

Player 2 with whom you interact for this round chose:
• "UP" 100.0 % of the times when Player 1 choose "UP".
•"UP" 0.0 % of the times when Player 1 "DOWN".
in the previous rounds.

Do you want to participate in the interaction this round?

Fig. 9 Record of P2, as displayed to P1

B Data from the final questionnaire

B.1 Understanding

Table 9 reports the distribution of answers to the question on understanding (all treatments from the main experiment are pooled).

Table 9 Distribution of answers to the question about understanding (Likert Scale)

Very difficult	Difficult	Neither difficult, nor easy	Easy	Very easy
0%	8.47%	38.25%	38.8%	14.48%

Table 10 Distribution of demographics across cities

	Brussels (%)	Lille (%)
Gender		
Woman	50	74
Man	50	24
Other	0	1
Status		
Student	89	92
Professional	4.4	5.7
Unemployed	5	1
Other	0.6	0.5
Previous participations in experiments		
0	32	67
1	46	10
Between 2 and 5	19	20
More than 5	2	3

B.2 Demographic questionnaire

Table 10 reports the distribution of demographics across cities.

C Power analysis

We followed a simulation approach for our power analysis and sample size determination. The advantage of simulation based power analysis is that we can adapt it to the test we will use on the actual data. We focused on the likelihood of the cooperative outcome as our variable of interest, and proceeded as follows:

- We randomly created samples fixing a number of parameters:
  - The number of observations per session.
  - The number of sessions.
  - The baseline probability of the cooperative outcome.
  - The effect of the existence of an information disclosure mechanism (i.e. an increase in the probability of the cooperative outcome, measured in percentage points).
  - The session effect size (also measured in percentage points).

In each sample, half the observations were allocated to the baseline, and half to a treatment with information disclosure. We generated 100 samples for each vector of parameters, and for each sample, we ran a logit model explaining the probability of the cooperative outcome by the treatment dummy, with errors clustered at the session level. For each vector of parameters, we recorded the percentage of the times the treatment dummy was significant at the 5% level. This gives us the power of our

Table 11 Power for some parameters

Power	Baseline likelihood	Treatment effect	Session effect	N sessions	N participants
.99	.2	.1	.05	10	140
.64	.2	.1	.1	10	140
1	.2	.15	.05	10	140
.96	.2	.15	.1	10	140
.58	.2	.15	.15	10	140

experiment for this vector of parameters. We reproduced it for numerous vectors. In Table 11, we report the power analysis for some parameters. This suggests that, if we expect an effect size of 10 pp with a moderate session effect, and session of, on average, 14 participants, we have appropriate power to detect effect size of 10 or 15 percentage points.

D Data collected online in a modified design

During the pandemic, while the physical labs were closed, we explored the possibility of unrolling our experiment online in a cost-effective way. We adapted the design in 3 ways to make it more suited for online data collection: (i) we reduced the number of periods from 30 to 20 (ii) we fixed the size of the matching groups to 10 participants (iii) most importantly, we dropped the Enter decision. The aim of these adaptations was to reduce the complexity of the experiment, its length, and the potential waiting times to minimize attrition (Arechar et al. 2018). We also chose to focus on the 3 treatments without noise, namely the Baseline, MandNN and VolNN treatments.

The experiment was developed using Lioness-Lab (Giamattei et al. 2020) and participants were recruited on Prolific. Data collection took place in Early 2021. We collected the data from 17 groups in total (170 participants), of which 9 completed the experiment. Table 12 provides information on the distribution of participants and groups across treatments. This important attrition made the data collection costly, so we decided to stop it and wait for the physical lab to reopen to complete our data collection. Nevertheless, the data collected online are very similar to the data from the main lab experiment, as shown in Table 13. This is reinsuring in two ways: (i) the Enter decision probably does not drive our results (ii) our results are robust to change in the experimental design and subject pools.

Table 12 Information on the sample size in the Online experiment

	N groups	N completed groups	N participants	N completed participants	Obs
Baseline	7	4	70	40	479
Mand	2	2	20	20	200
Vol	8	3	80	30	433

Table 13 Occurrence of the cooperative outcome across treatments in the online experiment

	All groups	Completed groups only
Baseline	0.153	0.160
Mand	0.235	0.235
Vol	0.222	0.220

E Support for the results (omitted in the text)

E.1 Test for Table 3

See Table 14.

E.2 Support for Result 1

Table 15 reports descriptive statistics. In the second row, the data from all the treatments with disclosure are pooled.

We regress a dummy variable indicating the cooperative outcome on a dummy variable indicating that there is a disclosure system (all treatments pooled). We report the marginal effects from Logit models. Standard errors are clustered at the session level. Results are in Table 16.

Table 14 Regressions for the significance levels reported in Table 3

	(1)	(2)	(3)	(4)	(5)
	Coop. out. = 1	P1 enters = 1	P1 coops = 1	P2 coops	P2 coops
				when P1 coops = 1	when P1 def. = 1
baseline	–	–	–	–	–
MandNN	0.149** (0.063)	0.125** (0.062)	0.131** (0.061)	0.164** (0.081)	−0.024 (0.062)
VolNN	0.125** (0.063)	0.094 (0.061)	0.074 (0.059)	0.181* (0.104)	−0.039 (0.045)
MandLN	0.051 (0.058)	0.048 (0.076)	0.044 (0.063)	−0.003 (0.080)	−0.018 (0.050)
VolLN	0.088 (0.068)	0.083 (0.069)	0.041 (0.071)	0.098 (0.103)	−0.046 (0.052)
Observations	5780	5780	5780	5780	5780
Session characteristics	Yes	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes	Yes
Sessions	25	25	25	25	25

Marginal effects from Logit models. Standard errors (in parentheses) are clustered at the session level using bootstrapping

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$

The level of observation is an interaction in Column (1) and individual decisions of P1s in Column (2) and (3) and of P2 in Column (4) and (5). All treatments are included

Session characteristics: City dummy and size of the session

Table 15 Descriptive statistics, separating baseline and treatments with reputation

Treatment	Coop. outcome	P1 enters	P1 coop	P2 coop. if:	
				P1 coop	P1 def
Baseline	0.110	0.686	0.289	0.358	0.179
Disclosure[a]	0.217	0.76	0.362	0.476	0.150

[a] Data from MandNN, MandLN, VolNN & VolLN pooled

Table 16 The effect of the existence of a disclosure system on the likelihood of the cooperative outcome

	(1)	(2)	(3)
	Coop. outcome = 1	Coop. outcome = 1	Coop. outcome = 1
disclosure	0.126** (0.056)	0.127** (0.057)	0.122** (0.059)
Observations	5780	5780	5780
Session characteristics	No	No	Yes
Period FE	No	Yes	Yes
Sessions	25	25	25

Standard errors (in parentheses) are clustered at the session level using bootstrapping
Marginal effects reported. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$
The level of observation is an interaction. All treatments are included
Session characteristics: City dummy and size of the session

E.3 Support for Result 3

See Table 17.

Table 17 The effect of one's record on the probability to disclose (Used to compute the predicted probabilities in Fig. 2)

	(1)		(2)	
	Disclose = 1		Disclose = 1	
	Coeff.	m.e.	Coeff.	m.e.
% of Coop. of P2 when P1 coop'd	2.997****	0.551****	3.731****	
	(0.590)	(0.093)	(0.991)	
% of Coop. of P2 when P1 def'd	−0.876	−0.161	−1.037	
	(0.573)	(0.109)	(1.233)	
VolNN			0.000	
			(0.000)	
VolLN			0.631	

Table 17 continued

	(1)		(2)	
	Disclose = 1		Disclose = 1	
	Coeff.	m.e.	Coeff.	m.e.
			(1.198)	
VolNN × % of Coop. of P2 when P1 coop'd			0.000	
			(0.000)	
VolLN × % of Coop. of P2 when P1 coop'd			−1.626	
			(1.240)	
VolNN × % of Coop. of P2 when P1 def'd			0.000	
			(0.000)	
VolLN × % of Coop. of P2 when P1 def'd			0.413	
			(1.331)	
Constant	−0.881		−1.571	
	(2.787)		(3.789)	
% of Coop. of P2 when P1 coop'd				
VolNN				0.648****
				(0.146)
VolLN				0.406****
				(0.118)
% of Coop. of P2 when P1 def'd				
VolNN				−0.180
				(0.222)
VolLN				−0.120*
				(0.073)
Observations	2204		2204	
Session Char	Yes		Yes	
Indiv. Char	Yes		Yes	

Standard errors (in parentheses) are clustered at the session level using bootstrapping
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$. The level of observation is P2s' individual decisions
Only treatments with Voluntary disclosure are included (VolNN VolLN)
Session char.: City dummy and size of the session
Individual char.: Gender, age, occupational status, experience with experiments

E.4 Support for Result 4

See Table 18.

Table 18 The effect of seeing the information about P2's past choices on P1s' choices. Marginal effects from Logit models

	(1)	(2)	(3)	(4)
	Enter = 1	Enter = 1	Coop. = 1	Coop. = 1
Marginal effect of				
Information disclosed	0.135****		0.178****	
at				
VolNN		0.142****		0.277****
		(0.028)		(0.054)
MandLN		0.084**		0.136***
		(0.035)		(0.047)
VolLN		0.258****		0.254****
		(0.039)		(0.059)
Observations	5587	5587	5587	5587
Session Char	Yes	Yes	Yes	Yes
Indiv. Char	Yes	Yes	Yes	Yes

Standard errors (in parentheses) are clustered at the session level using bootstrapping
Marginal effects reported. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$
The level of observation is P1s' individual decisions. All treatments are included
Session characteristics: City dummy and size of the session
Individual characteristics: Gender, age, occupational status, experience with experiments

E.5 Support Result 5

See Figs. 10 and 11.

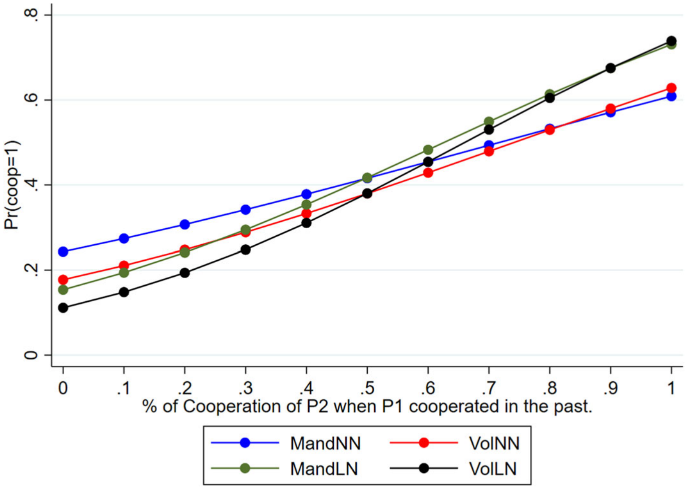


Fig. 10 The predicted probability of cooperation by P1 depending on the record of P2

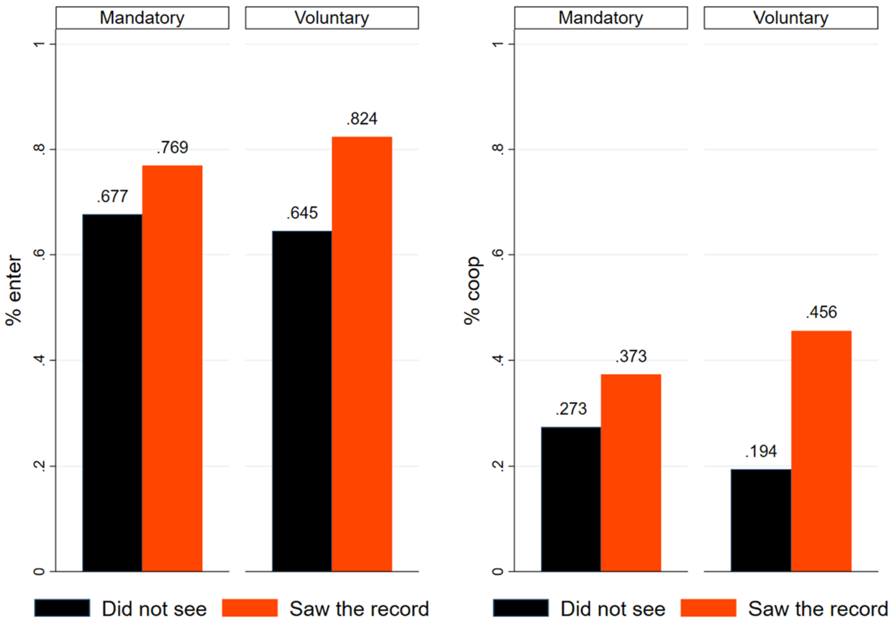


Fig. 11 Entry (left panel) and Cooperation (right panel) depending on whether P1 saw the record, contrasted by voluntary/mandatory disclosure

F Support for the results of the SR treatments

See Tables 19 and 20.

Table 19 Aggregate treatment effects over cooperative outcomes and individual behavior

	(1)	(2)	(3)	(4)	(5)
	Coop. out. = 1	P1 enters = 1	P1 coop. = 1	P2 coop.	P2 coop
				when P1 coop.=1	when P1 def. = 1
baseline	–	–	–	–	–
MandNN	0.155**** (0.047)	0.115* (0.060)	0.132*** (0.051)	0.177** (0.069)	−0.020 (0.059)
VolNN	0.127** (0.055)	0.083* (0.049)	0.079* (0.047)	0.173* (0.099)	−0.039 (0.036)
MandSR	0.124** (0.060)	−0.015 (0.057)	0.118** (0.055)	0.173** (0.078)	−0.065* (0.034)
VolSR	0.082 (0.050)	−0.031 (0.071)	0.074 (0.053)	0.142* (0.073)	−0.098*** (0.035)
Observations	6050	6050	6050	6050	6050
Period FE	Yes	Yes	Yes	Yes	Yes
Sessions	23	23	23	23	23

Marginal effects from Logit models. Standard errors in parentheses are clustered at the session level

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$

The level of observation is an interaction in Column (1) and individual decisions of P1s in Column (2) and (3) and of P2 in Column (4) and (5)

Session characteristics: City dummy and size of the session

Table 20 Support for the results from the SR treatments

	(1)	(2)	(3)
	P1 Coop = 1	P1 Coop = 1	Disclose = 1
Information disclosed	0.960**		
	(0.377)		
Period	−0.029****	−0.049****	−0.055****
	(0.008)	(0.014)	(0.012)
Information disclosed × Period	0.013*		
	(0.007)		
P2 cooperated in previous period		1.433****	1.701****
		(0.194)	(0.509)
P2 cooperated in previous period × Period		0.035*	0.142****
		(0.019)	(0.014)
Constant	−0.903	−0.966	0.307
	(1.874)	(0.973)	(0.640)
Observations	1160	1816	1092
Individual characteristics	Yes	Yes	Yes

Standard errors (in parentheses) are clustered at the session level using bootstrapping

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$

Observation is at the individual level

In Column (1) and (2), P1s' decisions are considered. In Column (1), only the data from VolSR are used

In Column (2), data from both SR treatments in which P1 received the information are used

In Column (3), P2s' decisions are considered. Only data from VolSR are used

Individual characteristics: Gender, age, occupational status, experience with experiments

G Additional results

G.1 Results at the session level

See Fig. 12.

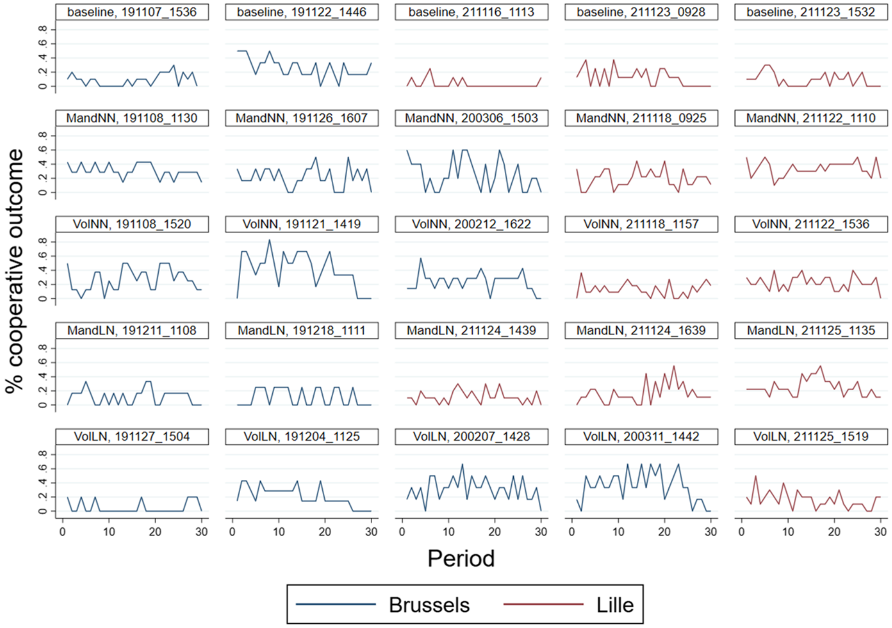


Fig. 12 Cooperative outcome across periods, for each session separately

G.2 The time dynamic of entry, (conditional) cooperation, and disclosure

See Figs. 13, 14, 15 and 16.

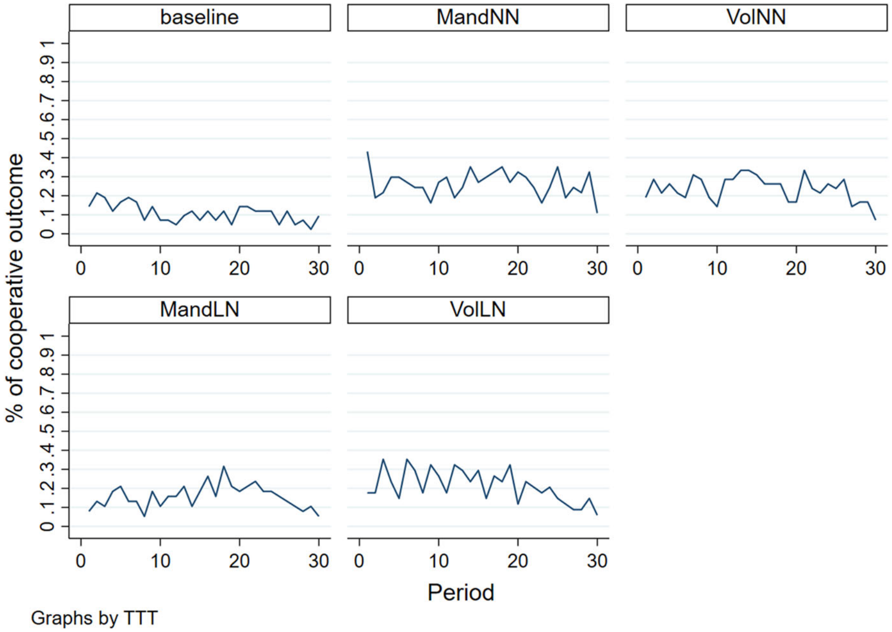


Fig. 13 The dynamic of the cooperative outcome (period averages)

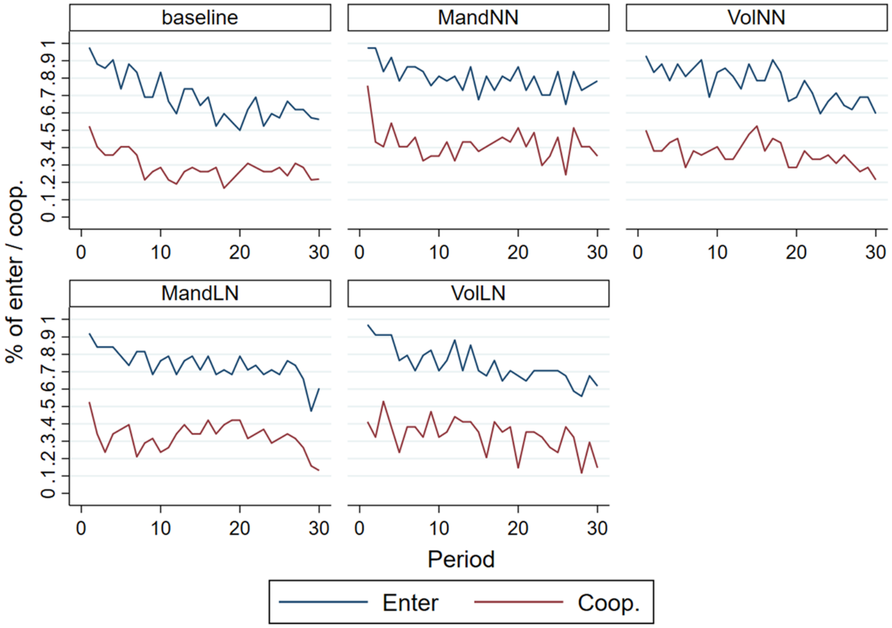


Fig. 14 The dynamic of P1's choices (period averages)

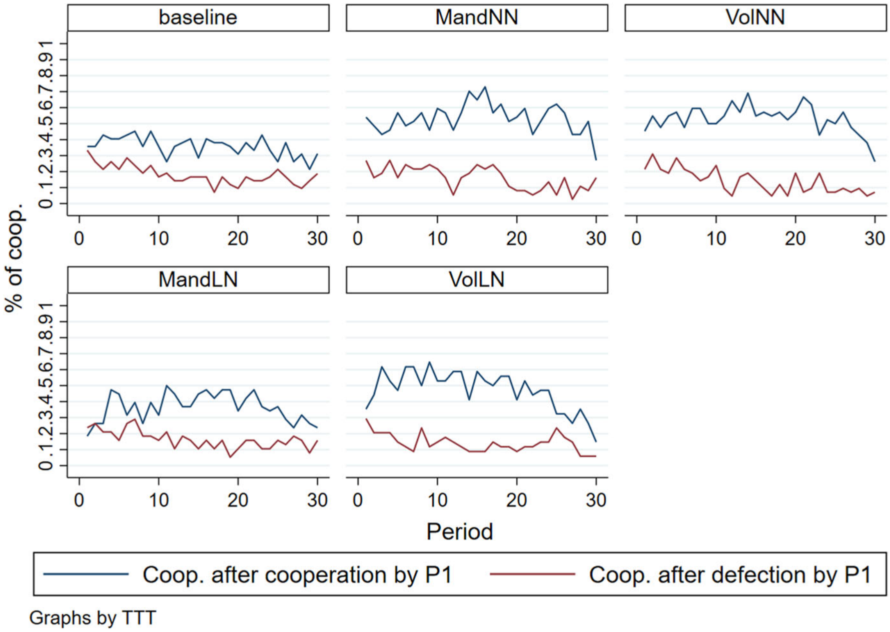


Fig. 15 The dynamic of P2's cooperation choices

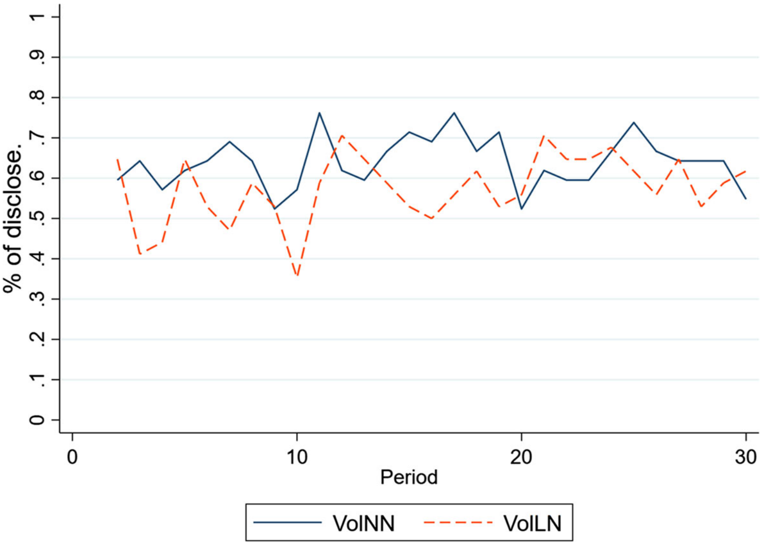


Fig. 16 The dynamic of disclosure (rate averaged at period level)

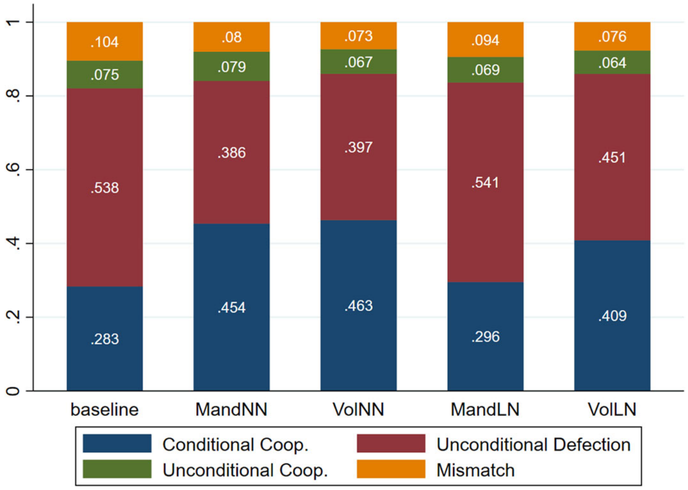


Fig. 17 The distribution of the 4 pure strategies of P2, by Treatment

Table 21 The effects of "Noisy" and "Voluntary" on the choice of pure strategy by P2 (multinomial logit model)

	(1)
Marginal effect of "Voluntary"	
Conditional cooperation	0.038 (0.049)
Unconditional cooperation	−0.005 (0.012)
Unconditional defection	−0.019 (0.051)
Mismatch	−0.014 (0.025)
Marginal effect of "noisy"	
Conditional cooperation	-0.106^{***} (0.039)
Unconditional cooperation	−0.010 (0.013)
Unconditional defection	0.107^{***} (0.040)
Mismatch	0.009 (0.027)
Observations	5780

Standard errors in parentheses are clustered at the session level
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$
The level of observation is one interaction. All treatments are included

G.3 Analysis of the 4 possibles strategies of P2

Figure 17 reports the distribution of pure strategies of P2s across treatments. In Table 21, we report the marginal effect of a multinomial logit regression explaining the strategy choice of P2s by the factorial interaction of "noisy" and "Voluntary". The results suggest that, while "Voluntary" has no effect on strategy choice at all, noise reduces the likelihood of conditional cooperation by 10 pp and increases the likelihood of defection by the same amount.

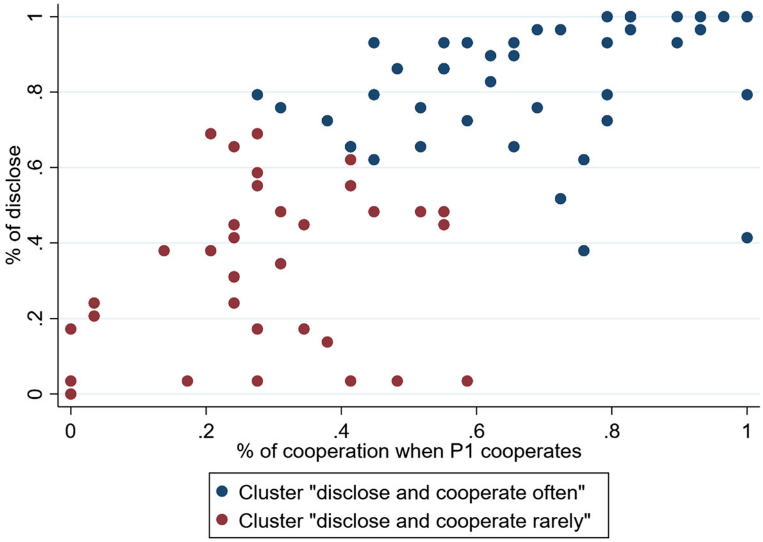


Fig. 18 Cluster of P2s according to disclosure and cooperation

G.4 Individuals and behavioral types

Does it pay to be skeptical? Our results suggest some skepticism among P1s. Here, we check whether being skeptical is beneficial in terms of earnings. To do so, we compute for each individual P1 the extent to which her cooperation rate depends on whether she received the information. This dependence of cooperation on information serves as measure of skepticism. We regress P1s' average stage-game earnings on this measure of skepticism, and we find that more skeptical individuals earn more (OLS with standard errors clustered at the session level: $b = 1.225, \ p = 0.03$).

Does it pay to reveal? P2 chose to reveal in the majority of the cases where he actually had a choice. We saw that disclosure is significantly linked to the quality of the record. Now we want to investigate individuals strategies. To do so, we compute for each P2 in the Endo treatments the percentage of time he cooperates and the percentage of time he discloses. Using the kmeans algorithm, we identify 2 clusters with a natural interpretation: the first cluster is composed of individuals who barely cooperate and disclose (n = 33). The second cluster consists of individuals who cooperate and disclose most of the time (n = 43).[20] Both these strategies make sense: if a P2 believes that P1s are not very skeptical, the first strategy makes sense (Jin et al. (2021) find evidence of such beliefs in a sender–receiver game). Conversely, the second strategy makes sense for a P2 who anticipates that the P1s are skeptical. Figure 18 represent the 2 clusters. We checked which strategy is the most beneficial for a P2, and we found that participants in the second cluster had higher average stage-game profits (7.42 vs 6.48, $p = 0.0017$ in a Mann-Whitney test).

[20] The same message holds if we choose to identify 3 clusters.

H (Seemingly) irrational behavior

Some behavior are hard to rationalize at first sight: In a lot of cases, P1s enter but do not cooperate (overentry) and P2s cooperate even if P1 defected. In this section, we discuss whether such behavior is indeed an indication for irrationality.

Overentry by P1 First, "over" entry by P1 might actually be motivated by social curiosity (one learns the corresponding conditional choice of P2 when entering). In addition, P1s might believe that (there is a small probability that) P2s cooperate in response to defection. Even if this probability is small, this can explain the aforementioned choices: Let's say that P1 believes that P2s cooperate when P1 cooperates with probability q. Let's say that P1 believes that P2s cooperate when P1 defects with probability p. Given our parameters, two conditions must be met so that P1 enters AND defects:

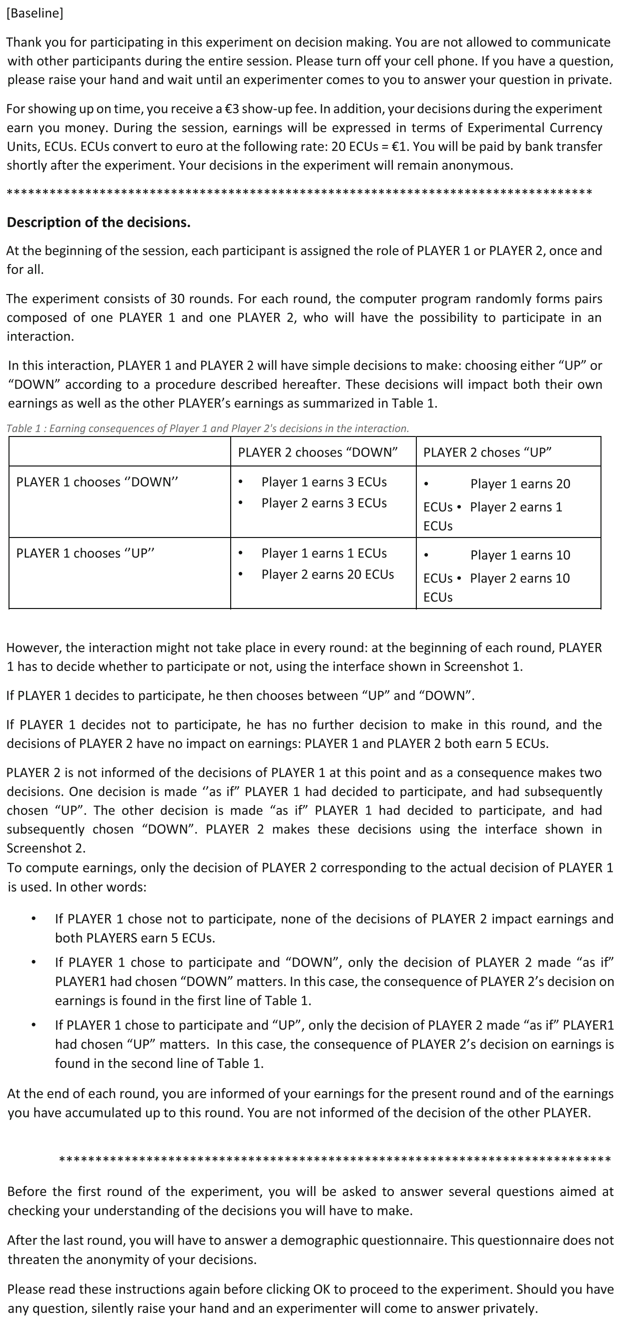
$$(1) : 20p - 3(1-p) \geq 5$$
$$(2) : 20p + 3(1-p) \geq +(1-q)$$

(1) is the condition to "entering conditional on defecting" It imposes $p > \frac{2}{17} \sim 0.11$. (2) is the condition "defecting conditional on entering". It imposes $p > (9q-2)/17$ There is a large set of (p, q) satisfying (1) & (2). Some are à-priori unlikely $(p > q)$, others are more plausible (e.g. $p \geq .15$ & $q \leq .3$ or $p \geq .2$ & $q \leq .5$). Miettinen et al. (2020) for instance elicit the beliefs of participants in a sequential prisoner dilemma and find that first movers on average expect second movers to cooperate 50% of the time when P1 cooperates, and 20% of the time when P1 defects. This might be due to P1s misunderstanding the game, to P1s expecting errors by P2s or to P1s expecting that P2 might be concerned by social welfare.

Cooperation after defection "cooperate when P1 defected" can be due to unconditional cooperation, i.e. P2s who cooperate irrespective of the decision of P1, motivated by e.g. altruism or social welfare. By choosing to cooperate to a defector, one increases the social welfare 3.5 folds (from 6 to 21) or by 15 ECUs (75cents). Such concern for welfare is well documented in the literature (see e.g. Engelmann and Strobel 2004). Miettinen et al. (2020) find that "concern for social welfare" is a good explanation (along with reciprocity, among others) of behaviors in the sequential prisoner dilemma. Interestingly enough, they also find that more or less 15% of P2s "cooperate when P1 defects". This should ease our concerns about irrationality on P2s' behalf.

On the other hand, mismatch, i.e. P2 who choose to cooperate only when P1 defected is more puzzling. In this situation, reciprocity, social welfare concerns or any other theory are of little help. Overall, "mismatch" choices correspond to 8,5% of P2s decisions. This type of behavior is concentrated on a small number of participants: 50% of P2s never mismatch. 75% of P2s mismatch less 5 times. Note that there is a significant negative time trend in mismatching, which is good news (people learn). Our data is overall comparable to the data in Miettinen et al. (2020). Our results are robust to excluding mismatch decisions, or to the exclusion of the 25% of P2 who mismatched 5 times or more.

I Instructions



[Baseline]

Thank you for participating in this experiment on decision making. You are not allowed to communicate with other participants during the entire session. Please turn off your cell phone. If you have a question, please raise your hand and wait until an experimenter comes to you to answer your question in private.

For showing up on time, you receive a €3 show-up fee. In addition, your decisions during the experiment earn you money. During the session, earnings will be expressed in terms of Experimental Currency Units, ECUs. ECUs convert to euro at the following rate: 20 ECUs = €1. You will be paid by bank transfer shortly after the experiment. Your decisions in the experiment will remain anonymous.

****

Description of the decisions.

At the beginning of the session, each participant is assigned the role of PLAYER 1 or PLAYER 2, once and for all.

The experiment consists of 30 rounds. For each round, the computer program randomly forms pairs composed of one PLAYER 1 and one PLAYER 2, who will have the possibility to participate in an interaction.

In this interaction, PLAYER 1 and PLAYER 2 will have simple decisions to make: choosing either “UP” or “DOWN” according to a procedure described hereafter. These decisions will impact both their own earnings as well as the other PLAYER’s earnings as summarized in Table 1.

Table 1 : Earning consequences of Player 1 and Player 2's decisions in the interaction.

	PLAYER 2 chooses “DOWN”	PLAYER 2 choses “UP”
PLAYER 1 chooses ‘’DOWN’’	• Player 1 earns 3 ECUs • Player 2 earns 3 ECUs	• Player 1 earns 20 ECUs • Player 2 earns 1 ECUs
PLAYER 1 chooses ‘’UP’’	• Player 1 earns 1 ECUs • Player 2 earns 20 ECUs	• Player 1 earns 10 ECUs • Player 2 earns 10 ECUs

However, the interaction might not take place in every round: at the beginning of each round, PLAYER 1 has to decide whether to participate or not, using the interface shown in Screenshot 1.

If PLAYER 1 decides to participate, he then chooses between “UP” and “DOWN”.

If PLAYER 1 decides not to participate, he has no further decision to make in this round, and the decisions of PLAYER 2 have no impact on earnings: PLAYER 1 and PLAYER 2 both earn 5 ECUs.

PLAYER 2 is not informed of the decisions of PLAYER 1 at this point and as a consequence makes two decisions. One decision is made ‘’as if” PLAYER 1 had decided to participate, and had subsequently chosen “UP”. The other decision is made “as if” PLAYER 1 had decided to participate, and had subsequently chosen “DOWN”. PLAYER 2 makes these decisions using the interface shown in Screenshot 2.
To compute earnings, only the decision of PLAYER 2 corresponding to the actual decision of PLAYER 1 is used. In other words:

- If PLAYER 1 chose not to participate, none of the decisions of PLAYER 2 impact earnings and both PLAYERS earn 5 ECUs.
- If PLAYER 1 chose to participate and “DOWN”, only the decision of PLAYER 2 made “as if” PLAYER1 had chosen “DOWN” matters. In this case, the consequence of PLAYER 2’s decision on earnings is found in the first line of Table 1.
- If PLAYER 1 chose to participate and “UP”, only the decision of PLAYER 2 made “as if” PLAYER1 had chosen “UP” matters. In this case, the consequence of PLAYER 2’s decision on earnings is found in the second line of Table 1.

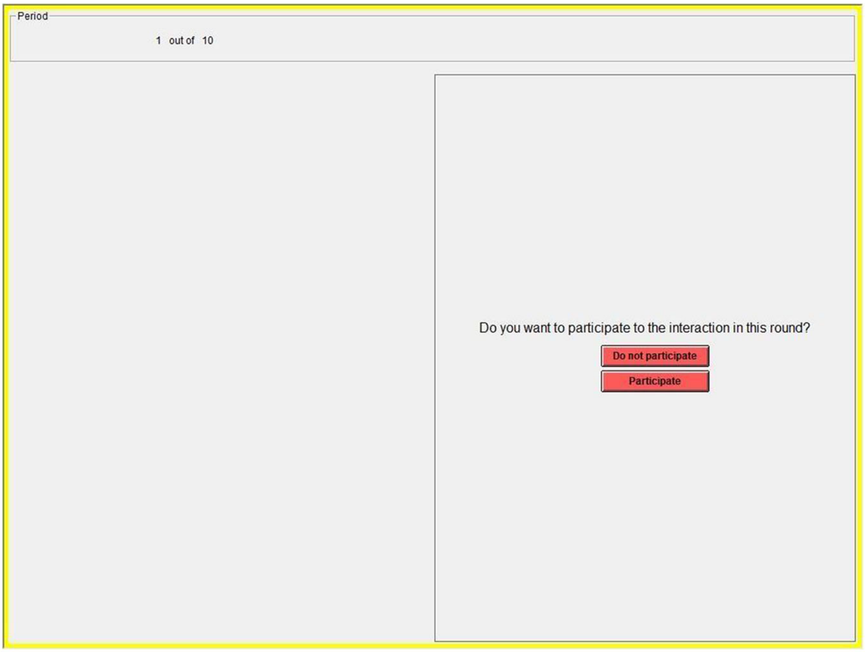
At the end of each round, you are informed of your earnings for the present round and of the earnings you have accumulated up to this round. You are not informed of the decision of the other PLAYER.

****

Before the first round of the experiment, you will be asked to answer several questions aimed at checking your understanding of the decisions you will have to make.

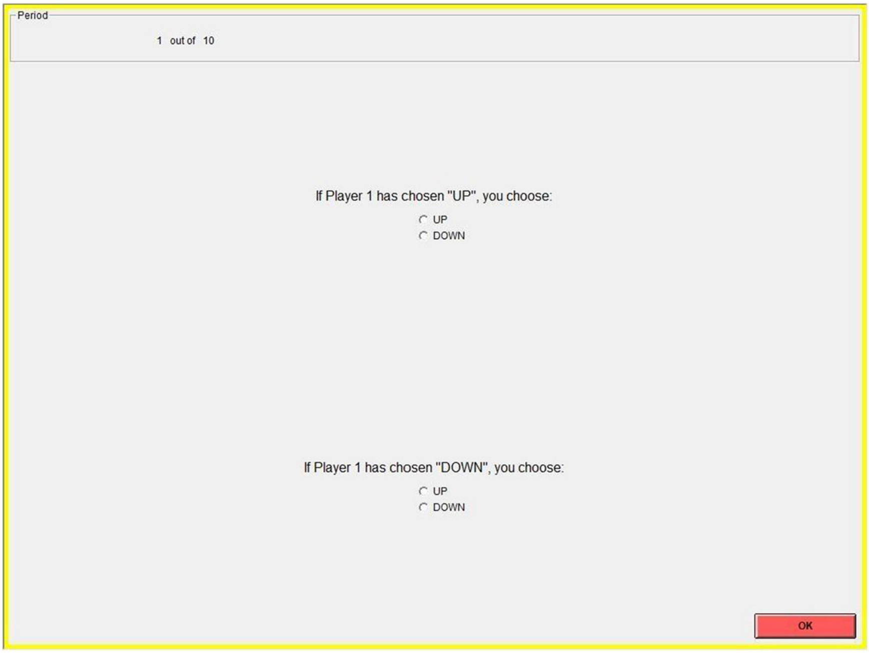
After the last round, you will have to answer a demographic questionnaire. This questionnaire does not threaten the anonymity of your decisions.

Please read these instructions again before clicking OK to proceed to the experiment. Should you have any question, silently raise your hand and an experimenter will come to answer privately.



Do you want to participate to the interaction in this round?

Screenshot 1: Participation decision (PLAYER 1)



If Player 1 has chosen "UP", you choose:

○ UP
○ DOWN

If Player 1 has chosen "DOWN", you choose:

○ UP
○ DOWN

Screenshot 2: UP or DOWN decisions (PLAYER 2)

[Mandatory No Noise]

Thank you for participating in this experiment on decision making. You are not allowed to communicate with other participants during the entire session. Please turn off your cell phone. If you have a question, please raise your hand and wait until an experimenter comes to you to answer your question in private.

For showing up on time, you receive a €3 show-up fee. In addition, your decisions during the experiment earn you money. During the session, earnings will be expressed in terms of Experimental Currency Units, ECU. ECU converts to euro at the following rate: 20 ECUs = €1. You will be paid by bank transfer shortly after the experiment. Your decisions in the experiment will remain anonymous.
****

Description of the decisions.

At the beginning of the session, each participant is assigned to the role of PLAYER 1 or PLAYER 2, once and for all.

The experiment consists of 30 rounds. For each round, the computer program randomly forms pairs composed of one PLAYER 1 and one PLAYER 2, who will have the possibility to participate in an interaction.

In this interaction, PLAYER 1 and PLAYER 2 will have simple decisions to make: choosing either "UP" or "DOWN" according to a procedure described hereafter. These decisions will impact both their own earnings as well as the other PLAYER's earnings as summarized in Table 1.

Table 1: Earning consequences of Player 1 and Player 2's decisions in the interaction.

	PLAYER 2 chooses "DOWN"	PLAYER 2 choses "UP"
PLAYER 1 chooses ''down''	• Player 1 earns 3 ECUs • Player 2 earns 3 ECUs	• Player 1 earns 20 ECUs • Player 2 earns 1 ECUs
PLAYER 1 chooses ''up''	• Player 1 earns 1 ECUs • Player 2 earns 20 ECUs	• Player 1 earns 10 ECUs • Player 2 earns 10 ECUs

However, the interaction might not take place in every round: at the beginning of each round, PLAYER 1 has to decide whether to participate or not, using the interface shown in Screenshot 1.

To help PLAYER 1's participation decision, from round 2 on, the past decisions of PLAYER 2 will be disclosed on the left panel of PLAYER 1's decision screen, as shown in screenshot 2.

If PLAYER 1 decides to participate, he then chooses between "UP" and "DOWN".

If PLAYER 1 decides not to participate, he has no further decision to make in this round, and the decisions of PLAYER 2 have no impact on earnings: PLAYER 1 and PLAYER 2 both earn 5 ECUs.

PLAYER 2 is not informed of the decisions of PLAYER 1 at this point and as a consequence makes two decisions. One decision is made ''as if" PLAYER 1 had decided to participate, and had subsequently chosen "UP". The other decision is made "as if" PLAYER 1 had decided to participate, and had subsequently chosen "DOWN". PLAYER 2 makes these decisions using the interface shown in Screenshot 3.

To compute earnings, only the decision of PLAYER 2 corresponding to the actual decisions of PLAYER 1 is used. In other words:

- If PLAYER 1 chose not to participate, none of the decisions of PLAYER 2 impact earnings and both PLAYERS earn 5 ECUs.
- If PLAYER 1 chose to participate and "DOWN", only the decision of PLAYER 2 taken "as if" PLAYER1 had chosen "DOWN" matters. In this case, the consequence of PLAYER 2's decision on earnings is found in the first line of Table 1.
- If PLAYER 1 chose to participate and "UP", only the decision of PLAYER 2 taken "as if" PLAYER1 had chosen "UP" matters. In this case, the consequence of PLAYER 2's decision on earnings is found in the second line of Table 1.

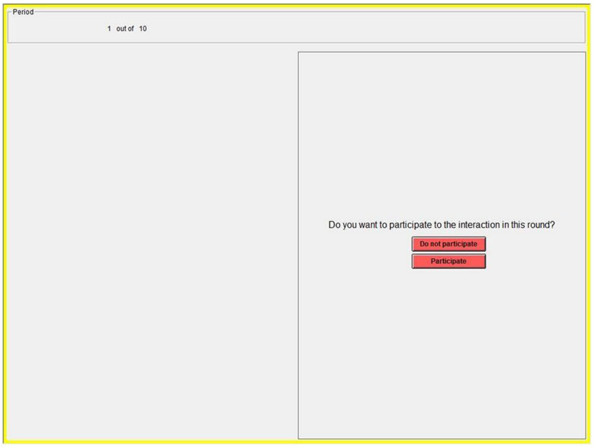
At the end of each round, you are informed of your earnings for the present round and of the earnings you have accumulated up to this round. You are not informed of the decision of the other PLAYER.

****

Before the first round of the experiment, you will be asked to answer several questions aimed at checking your understanding of the decisions you will have to make.

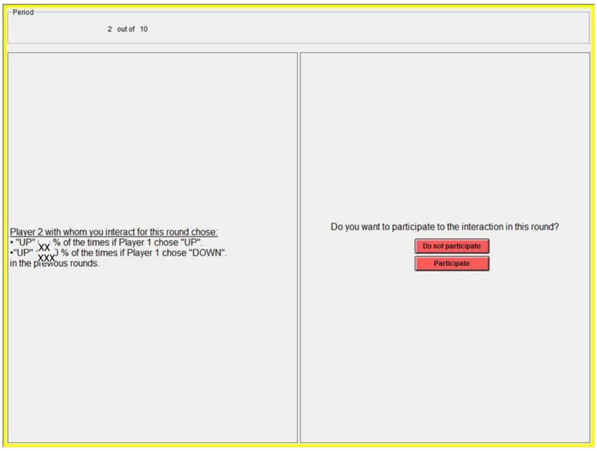
After the last round, you will have to answer a demographic questionnaire. This questionnaire does not threaten the anonymity of your decisions.

Please read these instructions again before clicking OK to proceed to the experiment. Should you have any question, silently raise your hand and an experimenter will come to answer privately.



Do you want to participate to the interaction in this round?

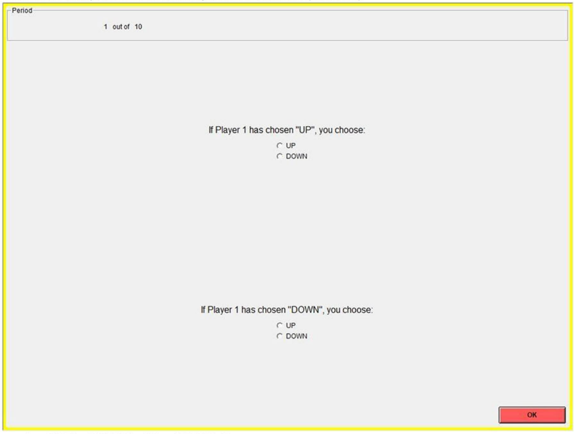
Screenshot 1: Participation decision (PLAYER 1)



Player 2 with whom you interact for this round chose:
• "UP" XX % of the times if Player 1 chose "UP".
•"UP" XXX % of the times if Player 1 chose "DOWN".
in the previous rounds.

Do you want to participate to the interaction in this round?

Screenshot 2: Participation decision with information disclosure (PLAYER 1)



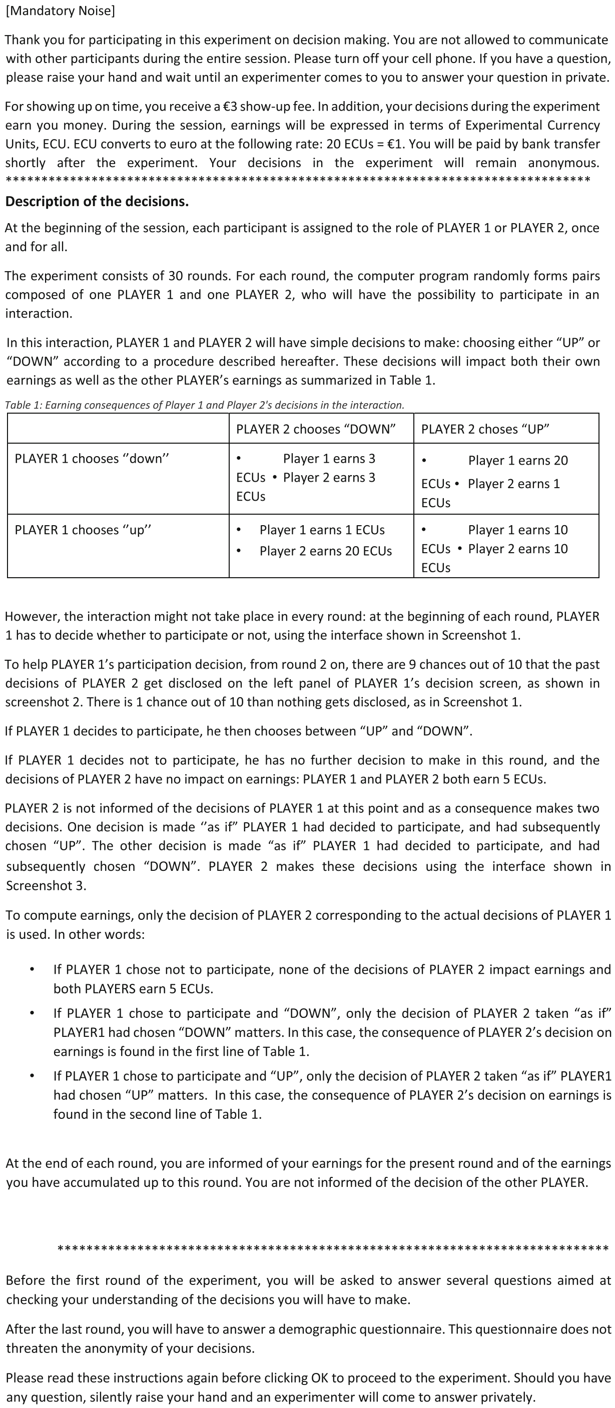
If Player 1 has chosen "UP", you choose:

○ UP
○ DOWN

If Player 1 has chosen "DOWN", you choose:

○ UP
○ DOWN

Screenshot 3: UP or DOWN decisions (PLAYER 2)



[Mandatory Noise]

Thank you for participating in this experiment on decision making. You are not allowed to communicate with other participants during the entire session. Please turn off your cell phone. If you have a question, please raise your hand and wait until an experimenter comes to you to answer your question in private.

For showing up on time, you receive a €3 show-up fee. In addition, your decisions during the experiment earn you money. During the session, earnings will be expressed in terms of Experimental Currency Units, ECU. ECU converts to euro at the following rate: 20 ECUs = €1. You will be paid by bank transfer shortly after the experiment. Your decisions in the experiment will remain anonymous.
****

Description of the decisions.

At the beginning of the session, each participant is assigned to the role of PLAYER 1 or PLAYER 2, once and for all.

The experiment consists of 30 rounds. For each round, the computer program randomly forms pairs composed of one PLAYER 1 and one PLAYER 2, who will have the possibility to participate in an interaction.

In this interaction, PLAYER 1 and PLAYER 2 will have simple decisions to make: choosing either "UP" or "DOWN" according to a procedure described hereafter. These decisions will impact both their own earnings as well as the other PLAYER's earnings as summarized in Table 1.

Table 1: Earning consequences of Player 1 and Player 2's decisions in the interaction.

	PLAYER 2 chooses "DOWN"	PLAYER 2 choses "UP"
PLAYER 1 chooses ''down''	• Player 1 earns 3 ECUs • Player 2 earns 3 ECUs	• Player 1 earns 20 ECUs • Player 2 earns 1 ECUs
PLAYER 1 chooses ''up''	• Player 1 earns 1 ECUs • Player 2 earns 20 ECUs	• Player 1 earns 10 ECUs • Player 2 earns 10 ECUs

However, the interaction might not take place in every round: at the beginning of each round, PLAYER 1 has to decide whether to participate or not, using the interface shown in Screenshot 1.

To help PLAYER 1's participation decision, from round 2 on, there are 9 chances out of 10 that the past decisions of PLAYER 2 get disclosed on the left panel of PLAYER 1's decision screen, as shown in screenshot 2. There is 1 chance out of 10 than nothing gets disclosed, as in Screenshot 1.

If PLAYER 1 decides to participate, he then chooses between "UP" and "DOWN".

If PLAYER 1 decides not to participate, he has no further decision to make in this round, and the decisions of PLAYER 2 have no impact on earnings: PLAYER 1 and PLAYER 2 both earn 5 ECUs.

PLAYER 2 is not informed of the decisions of PLAYER 1 at this point and as a consequence makes two decisions. One decision is made ''as if" PLAYER 1 had decided to participate, and had subsequently chosen "UP". The other decision is made "as if" PLAYER 1 had decided to participate, and had subsequently chosen "DOWN". PLAYER 2 makes these decisions using the interface shown in Screenshot 3.

To compute earnings, only the decision of PLAYER 2 corresponding to the actual decisions of PLAYER 1 is used. In other words:

- If PLAYER 1 chose not to participate, none of the decisions of PLAYER 2 impact earnings and both PLAYERS earn 5 ECUs.
- If PLAYER 1 chose to participate and "DOWN", only the decision of PLAYER 2 taken "as if" PLAYER1 had chosen "DOWN" matters. In this case, the consequence of PLAYER 2's decision on earnings is found in the first line of Table 1.
- If PLAYER 1 chose to participate and "UP", only the decision of PLAYER 2 taken "as if" PLAYER1 had chosen "UP" matters. In this case, the consequence of PLAYER 2's decision on earnings is found in the second line of Table 1.

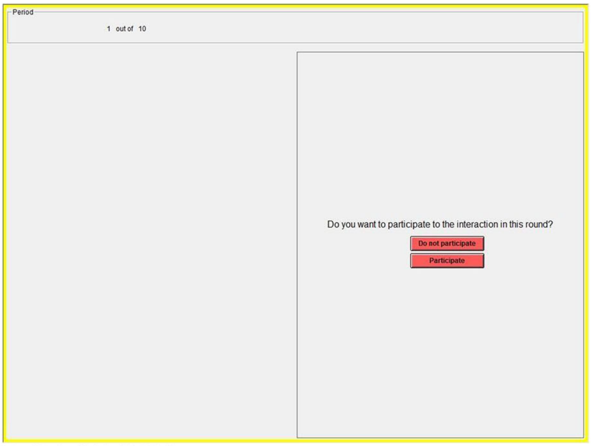
At the end of each round, you are informed of your earnings for the present round and of the earnings you have accumulated up to this round. You are not informed of the decision of the other PLAYER.

****

Before the first round of the experiment, you will be asked to answer several questions aimed at checking your understanding of the decisions you will have to make.

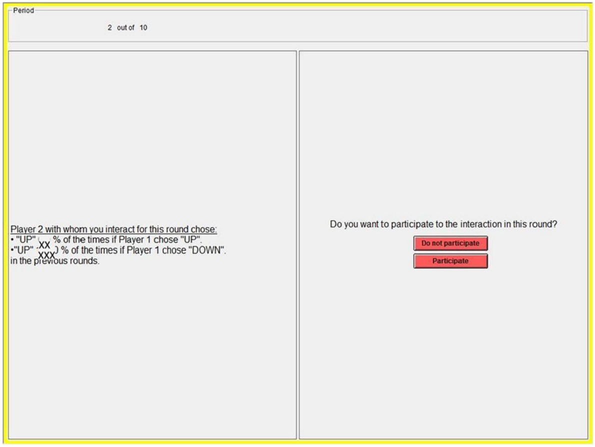
After the last round, you will have to answer a demographic questionnaire. This questionnaire does not threaten the anonymity of your decisions.

Please read these instructions again before clicking OK to proceed to the experiment. Should you have any question, silently raise your hand and an experimenter will come to answer privately.



Do you want to participate to the interaction in this round?

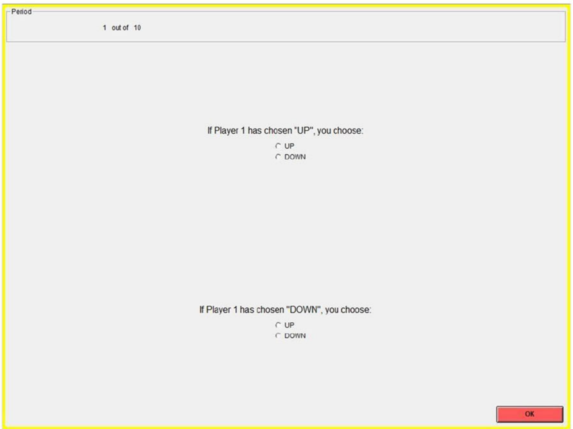
Screenshot 1: Participation decision (PLAYER 1)



<u>Player 2 with whom you interact for this round chose:</u>
• "UP" XX % of the times if Player 1 chose "UP".
•"UP" XXX % of the times if Player 1 chose "DOWN".
in the previous rounds.

Do you want to participate to the interaction in this round?

Screenshot 2: Participation decision with information disclosure (PLAYER 1)



If Player 1 has chosen "UP", you choose:
○ UP
○ DOWN

If Player 1 has chosen "DOWN", you choose:
○ UP
○ DOWN

Screenshot 3: UP or DOWN decisions (PLAYER 2)

[Voluntary No Noise]

Thank you for participating in this experiment on decision making. You are not allowed to communicate with other participants during the entire session. Please turn off your cell phone. If you have a question, please raise your hand and wait until an experimenter comes to you to answer your question in private.

For showing up on time, you receive a €3 show-up fee. In addition, your decisions during the experiment earn you money. During the session, earnings will be expressed in terms of Experimental Currency Units, ECU. ECU converts to euro at the following rate: 20 ECUs = €1. You will be paid by bank transfer shortly after the experiment. Your decisions in the experiment will remain anonymous.
****

Description of the decisions.

At the beginning of the session, each participant is assigned to the role of PLAYER 1 or PLAYER 2, once and for all.

The experiment consists of 30 rounds. For each round, the computer program randomly forms pairs composed of one PLAYER 1 and one PLAYER 2, who will have the possibility to participate in an interaction.

In this interaction, PLAYER 1 and PLAYER 2 will have simple decisions to make: choosing either “UP” or “DOWN” according to a procedure described hereafter. These decisions will impact both their own earnings as well as the other PLAYER’s earnings as summarized in Table 1.

Table 1 : Earning consequences of Player 1 and Player 2's decisions in the interaction.

	PLAYER 2 chooses “DOWN”	PLAYER 2 choses “UP”
PLAYER 1 chooses ‘’down’’	• Player 1 earns 3 ECUs • Player 2 earns 3 ECUs	• Player 1 earns 20 ECUs • Player 2 earns 1 ECUs
PLAYER 1 chooses ‘’up’’	• Player 1 earns 1 ECUs • Player 2 earns 20 ECUs	• Player 1 earns 10 ECUs • Player 2 earns 10 ECUs

However, the interaction might not take place in every round: at the beginning of each round, PLAYER 1 has to decide whether to participate or not, using the interface shown in Screenshot 1.

To help PLAYER 1’s participation decision, from round 2 on PLAYER 2 can choose to disclose his past decisions (see Screenshot 2). If so, the past decisions of PLAYER 2 will be disclosed on the left part of PLAYER 1’s decision screen, as shown in Screenshot 3.

If PLAYER 1 decides to participate, he then chooses between “UP” and “DOWN”.

If PLAYER 1 decides not to participate, he has no further decision to make in this round, and the decisions of PLAYER 2 have no impact on earnings: PLAYER 1 and PLAYER 2 both earn 5 ECUs.

PLAYER 2 is not informed of the decisions of PLAYER 1 at this point and as a consequence makes two decisions. One decision is made ‘’as if” PLAYER 1 had decided to participate, and had subsequently chosen “UP”. The other decision is made “as if” PLAYER 1 had decided to participate, and had subsequently chosen “DOWN”. PLAYER 2 makes these decisions using the interface shown in Screenshot 4.
To compute earnings, only the decision of PLAYER 2 corresponding to the actual decisions of PLAYER 1 is used. In other words:

- If PLAYER 1 chose not to participate, none of the decision of PLAYER 2 impacts earnings and both PLAYERS earn 5 ECUs.
- If PLAYER 1 chose to participate and “DOWN”, only the decision of PLAYER 2 taken “as if” PLAYER1 had chosen “DOWN” matters. In this case, the consequence of PLAYER 2’s decision on earnings is found in the first line of Table 1.
- If PLAYER 1 chose to participate and “UP”, only the decision of PLAYER 2 taken “as if” PLAYER1 had chosen “UP” matters. In this case, the consequence of PLAYER 2’s decision on earnings is found in the second line of Table 1.

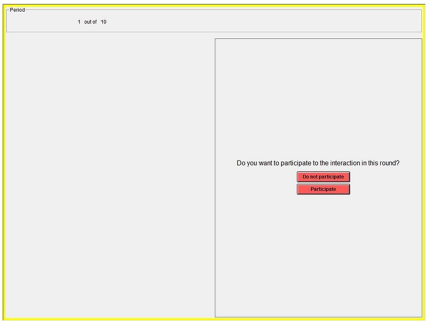
At the end of each round, you are informed of your earnings for the present round and of the earnings you have accumulated up to this round. You are not informed of the decision of the other PLAYER.

****

Before the first round of the experiment, you will be asked to answer several questions aimed at checking your understanding of the decisions you will have to make.

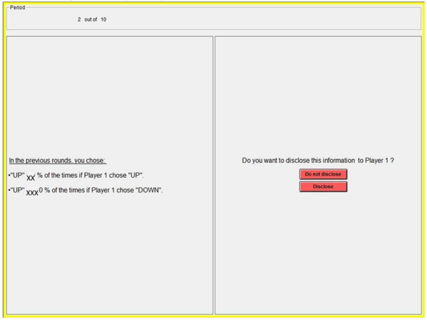
After the last round, you will have to answer a demographic questionnaire. This questionnaire does not threaten the anonymity of your decisions.

Please read these instructions again before clicking OK to proceed to the experiment. Should you have any question, silently raise your hand and an experimenter will come to answer privately.



Do you want to participate to the interaction in this round?

Screenshot 1: Participation decision (PLAYER 1)



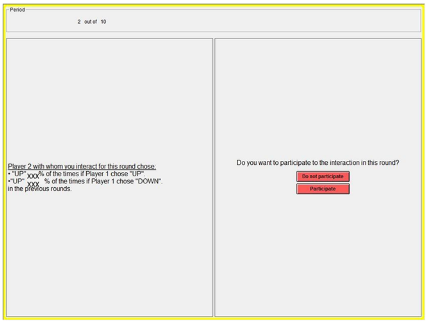
In the previous rounds, you chose:

•"UP" xx % of the times if Player 1 chose "UP".

•"UP" xxx0 % of the times if Player 1 chose "DOWN".

Do you want to disclose this information to Player 1 ?

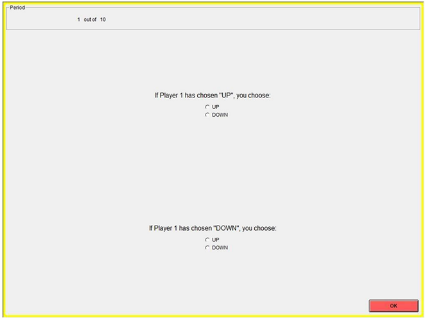
Screenshot 2: Disclosure decision (PLAYER 2)



Player 2 with whom you interact for this round chose:
• "UP" xxx% of the times if Player 1 chose "UP".
•"UP" xxx % of the times if Player 1 chose "DOWN".
in the previous rounds.

Do you want to participate to the interaction in this round?

Screenshot 3: Participation decision with information disclosure (PLAYER 1)



If Player 1 has chosen "UP", you choose:

○ UP
○ DOWN

If Player 1 has chosen "DOWN", you choose:

○ UP
○ DOWN

Screenshot 4: UP or DOWN decisions (PLAYER 2)

[Voluntary Noise]

Thank you for participating in this experiment on decision making. You are not allowed to communicate with other participants during the entire session. Please turn off your cell phone. If you have a question, please raise your hand and wait until an experimenter comes to you to answer your question in private.

For showing up on time, you receive a €3 show-up fee. In addition, your decisions during the experiment earn you money. During the session, earnings will be expressed in terms of Experimental Currency Units, ECU. ECU converts to euro at the following rate: 20 ECUs = €1. You will be paid by bank transfer shortly after the experiment. Your decisions in the experiment will remain anonymous.
****

Description of the decisions.

At the beginning of the session, each participant is assigned to the role of PLAYER 1 or PLAYER 2, once and for all.

The experiment consists of 30 rounds. For each round, the computer program randomly forms pairs composed of one PLAYER 1 and one PLAYER 2, who will have the possibility to participate in an interaction.

In this interaction, PLAYER 1 and PLAYER 2 will have simple decisions to make: choosing either "UP" or "DOWN" according to a procedure described hereafter. These decisions will impact both their own earnings as well as the other PLAYER's earnings as summarized in Table 1.

Table 1 : Earning consequences of Player 1 and Player 2's decisions in the interaction.

	PLAYER 2 chooses "DOWN"	PLAYER 2 choses "UP"
PLAYER 1 chooses ''down''	• Player 1 earns 3 ECUs • Player 2 earns 3 ECUs	• Player 1 earns 20 ECUs • Player 2 earns 1 ECUs
PLAYER 1 chooses ''up''	• Player 1 earns 1 ECUs • Player 2 earns 20 ECUs	• Player 1 earns 10 ECUs • Player 2 earns 10 ECUs

However, the interaction might not take place in every round: at the beginning of each round, PLAYER 1 has to decide whether to participate or not, using the interface shown in Screenshot 1.

To help PLAYER 1's participation decision, from round 2 on PLAYER 2 can choose to disclose his past decisions (see Screenshot 2). If Player 2 chooses to disclose, there is 9 chances out of 10, that the past decisions of PLAYER 2 get disclosed on the left part of PLAYER 1's decision screen, as shown in Screenshot 3, and 1 chance out of 10 that nothing gets disclosed.

If PLAYER 1 decides to participate, he then chooses between "UP" and "DOWN".

If PLAYER 1 decides not to participate, he has no further decision to make in this round, and the decisions of PLAYER 2 have no impact on earnings: PLAYER 1 and PLAYER 2 both earn 5 ECUs.

PLAYER 2 is not informed of the decisions of PLAYER 1 at this point and as a consequence makes two decisions. One decision is made ''as if" PLAYER 1 had decided to participate, and had subsequently chosen "UP". The other decision is made "as if" PLAYER 1 had decided to participate, and had subsequently chosen "DOWN". PLAYER 2 makes these decisions using the interface shown in Screenshot 4.
To compute earnings, only the decision of PLAYER 2 corresponding to the actual decisions of PLAYER 1 is used. In other words:

- If PLAYER 1 chose not to participate, none of the decision of PLAYER 2 impacts earnings and both PLAYERS earn 5 ECUs.
- If PLAYER 1 chose to participate and "DOWN", only the decision of PLAYER 2 taken "as if" PLAYER1 had chosen "DOWN" matters. In this case, the consequence of PLAYER 2's decision on earnings is found in the first line of Table 1.
- If PLAYER 1 chose to participate and "UP", only the decision of PLAYER 2 taken "as if" PLAYER1 had chosen "UP" matters. In this case, the consequence of PLAYER 2's decision on earnings is found in the second line of Table 1.

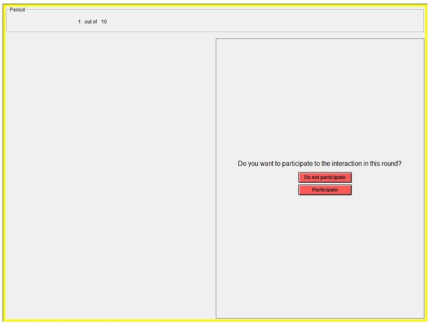
At the end of each round, you are informed of your earnings for the present round and your cumulated earnings, but not of the decision of the other PLAYER.

****

Before the first round of the experiment, you will be asked to answer several questions aimed at checking your understanding of the decisions you will have to make.

After the last round, you will have to answer a demographic questionnaire. This questionnaire does not threaten the anonymity of your decisions.

Please read these instructions again before clicking OK to proceed to the experiment. Should you have any question, silently raise your hand so that an experimenter will come to answer privately.

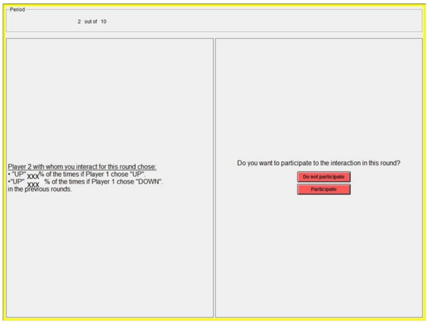


Period

1 out of 10

Do you want to participate to the interaction in this round?

Screenshot 1: Participation decision (PLAYER 1)



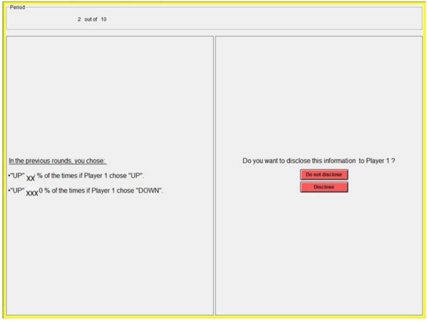
Period

2 out of 10

Player 2 with whom you interact for this round chose:
• "UP" XXX% of the times if Player 1 chose "UP".
•"UP" XXX % of the times if Player 1 chose "DOWN".
in the previous rounds.

Do you want to participate to the interaction in this round?

Screenshot 3: Participation decision with information disclosure (PLAYER 1)



Period

2 out of 10

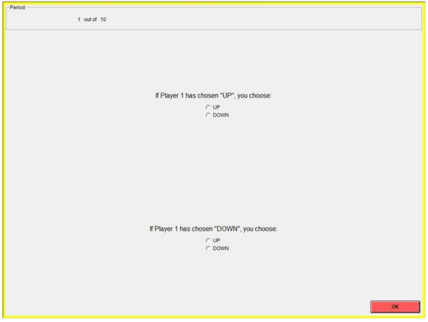
In the previous rounds, you chose:

•"UP" XX % of the times if Player 1 chose "UP".

•"UP" XXX0 % of the times if Player 1 chose "DOWN".

Do you want to disclose this information to Player 1 ?

Screenshot 2: Disclosure decision (PLAYER 2)



Period

1 out of 10

If Player 1 has chosen "UP", you choose:

○ UP
○ DOWN

If Player 1 has chosen "DOWN", you choose:

○ UP
○ DOWN

Screenshot 4: UP or DOWN decisions (PLAYER 2)

References

Acquisti, A., Taylor, C., Wagman, L.: The economics of privacy. J. Econ. Lit. **54**(2), 442–92 (2016)

Arechar, A.A., Gächter, S., Molleman, L.: Conducting interactive experiments online. Exp. Econ. **21**, 99–131 (2018)

Benndorf, V., Kübler, D., Normann, H.-T.: Privacy concerns, voluntary disclosure of information, and unraveling: an experiment. Eur. Econ. Rev. **75**, 43–59 (2015)

Benndorf, V., Normann, H.-T.: The willingness to sell personal data. Scand. J. Econ. **120**(4), 1260–1278 (2018)

Bertomeu, J., Cianciaruso, D.: Verifiable disclosure. Econ. Theory **65**(4), 1011–1044 (2018)

Bohnet, I., Huck, S.: Repetition and reputation: implications for trust and trustworthiness when institutions change. Am. Econ. Rev. **94**(2), 362–366 (2004)

Bolton, G.E., Katok, E., Ockenfels, A.: How effective are electronic reputation mechanisms? An experimental investigation. Manag. Sci. **50**(11), 1587–1602 (2004)

Brandts, J., Charness, G.: Hot vs. cold: sequential responses and preference stability in experimental games. Exp. Econ. **2**(3), 227–238 (2000)

Brandts, J., Charness, G.: The strategy versus the direct-response method: a first survey of experimental comparisons. Exp. Econ. **14**(3), 375–398 (2011)

Brown, A.L., Camerer, C.F., Lovallo, D.: To review or not to review? Limited strategic thinking at the movie box office. Am. Econ. J. Microecon. **4**(2), 1–26 (2012)

Bénabou, R., Tirole, J.: Incentives and prosocial behavior. Am. Econ. Rev. **96**(5), 1652–1678 (2006)

Camera, G., Casari, M.: Cooperation among strangers under the shadow of the future. Am. Econ. Rev. **99**(3), 979–1005 (2009)

Cameron, A.C., Trivedi, P.K., et al.: Microeconometrics using stata, vol. 2. Stata Press, College Station (2010)

Casari, M., Cason, T.N.: The strategy method lowers measured trustworthy behavior. Econ. Lett. **103**(3), 157–159 (2009)

Charness, G., Du, N., Yang, C.-L.: Trust and trustworthiness reputations in an investment game. Games Econ. Behav. **72**(2), 361–375 (2011)

Clark, K., Sefton, M.: The sequential prisoner's dilemma: evidence on reciprocation. Econ. J. **111**(468), 51–68 (2001)

Cox, C.A., Jones, M.T., Pflum, K.E., Healy, P.J.: Revealed reputations in the finitely repeated prisoners' dilemma. Econ. Theory **58**(3), 441–484 (2015)

Dranove, D., Jin, G.Z.: Quality disclosure and certification: theory and practice. J. Econ. Lit. **48**(4), 935–63 (2010)

Duffy, J., Ochs, J.: Cooperative behavior and the frequency of social interaction. Games Econ. Behav. **66**(2), 785–812 (2009)

Duffy, J., Xie, H., Lee, Y.-J.: Social norms, information, and trust among strangers: theory and evidence. Econ. Theory **52**(2), 669–708 (2013)

Engelmann, D., Strobel, M.: Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. Am. Econ. Rev. **94**(4), 857–869 (2004)

Feess, E., Kerzenmacher, F.: Sorting of trustees: the good and the bad stay in the game. Econ. Theory (2023). https://doi.org/10.1007/s00199-023-01491-3

Gächter, S., Renner, E.: The effects of (incentivized) belief elicitation in public goods experiments. Exp. Econ. **13**, 364–377 (2010)

Gaechter, S., Lee, K., Sefton, M., et al.: The variability of conditional cooperation in sequential prisoner's dilemmas. Technical Report (2022)

Ghidoni, R., Cleave, B.L., Suetens, S.: Perfect and imperfect strangers in social dilemmas. Eur. Econ. Rev. **116**, 148–159 (2019)

Ghidoni, R., Suetens, S.: The effect of sequentiality on cooperation in repeated games. Am Econ J Microecon **14**, 58 (2022)

Giamattei, M., Yahosseini, K.S., Gächter, S., Molleman, L.: Lioness lab: a free web-based platform for conducting interactive experiments online. J Econ Sci Assoc **6**, 95–111 (2020)

Hagenbach, J., Saucet, C.: Motivated Skepticism. Working Paper (2022)

Harrs, S., Rockenbach, B., Wenner, L.M.: Revealing good deeds: disclosure of social responsibility in competitive markets. Exp. Econ. **25**, 1–25 (2022)

Jin, G.Z., Luca, M., Martin, D.: Is no news (perceived as) bad news? An experimental investigation of information disclosure. Am. Econ. J. Microecon. **13**(2), 141–73 (2021)

Jin, G.Z., Luca, M., Martin, D.: Complex disclosure. Manag. Sci. **68**(5), 3236–3261 (2022)

Kamei, K.: Endogenous reputation formation under the shadow of the future. J. Econ. Behav. Organ. **142**, 189–204 (2017)

Kamei, K.: Voluntary disclosure of information and cooperation in simultaneous-move economic interactions. J. Econ. Behav. Organ. **171**, 234–246 (2020)

Kamei, K., Putterman, L.: Play it again: partner choice, reputation building and learning from finitely repeated dilemma games. Econ. J. **127**(602), 1069–1095 (2016)

Keser, C.: Experimental games for the design of reputation management systems. IBM Syst. J **42**(3), 498–506 (2003)

Mengel, F.: Risk and temptation: a meta-study on prisoner's dilemma games. Econ. J. **128**(616), 3182–3209 (2017)

Miettinen, T., Kosfeld, M., Fehr, E., Weibull, J.: Revealed preferences in a sequential prisoners' dilemma: a horse-race between six utility functions. J. Econ. Behav. Organ. **173**, 1–25 (2020)

Milgrom, P.R.: Good news and bad news: representation theorems and applications. Bell J. Econ. **12**, 380–391 (1981)

Montero, M., Sheth, J.D.: Naivety about hidden information: an experimental investigation. J. Econ. Behav. Organ. **192**, 92–116 (2021)

Penczynski, S., Koch, C., Zhang S.: Disclosure of verifiable information under competition: an experimental study. Available at SSRN 4130449 (2022)

Schudy, S., Utikal, V.: 'You must not know about me'-on the willingness to share personal data. J. Econ. Behav. Organ. **141**, 1–13 (2017)

Sheth, J.D.: Disclosure of information under competition: an experimental study. Games Econ. Behav. **129**, 158–180 (2021)

Tversky, A., Kahneman, D.: Advances in prospect theory: Cumulative representation of uncertainty. J. Risk Uncertain. **5**(4), 297–323 (1992)

Varian, H.R.: Economic Aspects of Personal Privacy, pp. 101–109. Springer, Boston (2009)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.