Shifting normative views: On why groups behave more antisocially than individuals

Sascha Behnk

Li Hao

Ernesto Reuben

Shifting normative views: On why groups behave more antisocially than individuals

Sascha Behnk
University of Zurich and University of Applied Science Europe, e-mail: sascha.behnk@bf.uzh.ch

Li Hao
Convoy Inc., e-mail: lhao@convoy.com

Ernesto Reuben
New York University Abu Dhabi and the Luxembourg Institute of Socio-Economic Research, e-mail: ereuben@nyu.edu

Abstract

A growing body of research shows that people tend to act more antisocially in groups than alone. However, little is known about why having "partners in crime" has such an effect. We run a laboratory experiment using sender-receiver games in which we elicit subjects' normative views, beliefs, and feelings of guilt to shed light on potential driving factors of this phenomenon. We find that the involvement of an additional sender makes the antisocial actions of senders more normatively acceptable to *all* parties, including receivers. This shift in normative views goes along with senders experiencing less guilt from behaving antisocially. These results apply independently of whether the antisocial action is deceptive or not. Lastly, we identify a necessary condition for this effect: the additional sender has to actively participate in the decision-making.

This version: July 2019

JEL Codes: H41, C92, D63

Keywords: group decision-making, diffusion of responsibility, normative beliefs, social norms, guilt aversion, emotions

1 Introduction

An increasing amount of evidence shows that people are more likely to behave antisocially when acting together than when acting alone. For instance, increased antisocial behavior with "partners in crime" has been observed in several contexts, including altruistic giving (Luhan et al., 2009), reciprocity (Cox, 2002; Kocher and Sutter, 2007), lying (Sutter, 2009; Weisel and Shalvi, 2015; Kocher et al., 2018), whistle-blowing (Choo et al., 2016), and markets of goods with negative externalities (Falk and Szech, 2013; Bartling et al., 2015). Strikingly, Dana et al. (2007) demonstrate that groups behave more antisocially even in one-shot settings where group members cannot interact or communicate in any way. In other words, they find that simply knowing that others are involved in the decision-making is sufficient to reduce prosocial behavior.[1] In this study, we shed light on this phenomenon by investigating whether the increase in antisocial behavior in joint decisions is linked to changes in normative beliefs.

There is a growing body of literature that provides evidence that normative beliefs play a crucial role in decisions involving prosocial and antisocial behaviors.[2] Importantly, recent studies indicate that elicited normative views can predict changes in behavior that are induced by subtle contextual variations (Krupka and Weber, 2013).[3] We extend this literature by investigating in a laboratory experiment whether the mere presence of additional decision-makers makes antisocial actions more normatively acceptable, resulting in more antisocial behavior.

Our experimental setup is designed to evaluate the extent to which an additional decision-maker erodes normative beliefs. To do so, we measure the impact on normative beliefs in multiple ways. First, we elicit the experimental subjects' normative views by asking them to rate the normative acceptability of antisocial actions. To obtain incentive-compatible answers, we also ask subjects to predict the ratings of others and reward them for the accuracy of their prediction. Second, we

[1] This is not to say that there might be additional reasons for increased antisocial behavior in groups. As proposed by Falk and Szech (2013) and Kocher et al. (2018), other explanations include: learning that others are willing to act antisocially might decrease one's willingness to act prosocially; arguments justifying antisocial behavior might be more convincing than those promoting prosocial behavior; specific rules used to aggregate individual preferences (e.g., majority voting) might lead to more antisociality, and materialistic framing embedded in institutions used to make collective decisions might divert attention away from the unacceptability of the antisocial behavior (e.g., bargaining processes in markets).

[2] See, for example, Cialdini et al. (1990), Cialdini (2003), Bicchieri (2006), Bicchieri and Xiao (2009), and Reuben and Riedl (2013).

[3] Normative views have been shown to vary across dictator, ultimatum, trust, and public goods games (Kimbrough and Vostroknutov, 2016), within dictator games due to the type of available actions (Krupka and Weber, 2013) or the presence of peers (Gächter et al., 2017), and in a trust game due to pre-play agreements (Krupka et al., 2017).

elicit the intensity with which subjects experience guilt if they act antisocially. We concentrate on the emotion of guilt because it has been shown empirically to be a crucial determinant of antisocial behavior (e.g., Hopfensitz and Reuben, 2009) and is often modeled as one of the main motivations for complying with social norms (see Elster, 2009; Bicchieri et al., 2018; López-Pérez, 2010). By measuring guilt, we can verify whether the hedonic experience of individuals is consistent with their normative views. Third, we also elicit the subjects' belief of what they think are the expectations of others concerning their own behavior. Given that compliance with normatively-desirable behavior requires both a shared understanding that certain actions are desirable and a shared belief that one is expected to behave in a desirable way (Bicchieri, 2006), measuring the subjects' second-order beliefs allows us to make an important distinction. Namely, does antisocial behavior increase in groups because antisocial actions become more acceptable, or does it increase because acceptable behavior is expected less often?

Finally, we also assess the importance of other group members having a say in the decision-making for there to be a change in the willingness to act more antisocially. In other words, we test whether the mere presence of other group members is sufficient to increase the willingness to choose antisocial actions (as predicted by outcome-based models of social preferences), or whether group members have to be actively involved in determining the group's decision for the erosion of normative beliefs to occur.

To be more specific, we employ sender-receiver games customized to study the abovementioned questions. In the games, a receiver chooses one out of ten options to determine the earnings of all players. The receiver knows the distribution of payoffs among the ten options but does not know what payoffs are associated with particular options. The receiver's only information is a message transmitted by either one or two informed senders. The message is either prosocial (identifies the option that gives everyone an equal payoff) or antisocial (identifies the unequal option that benefits senders at the expense of receivers).[4] We compare a game where the antisocial message is chosen by a single sender, the *1-Sender* game, to a game where the antisocial message needs to be approved by two senders who cannot communicate or bargain with each other, the *2-Sender* game. We use a price list to elicit the minimum monetary compensation required by a sender to transmit the antisocial message, which we call the "antisocial premium." This method allows to attach a monetary value to the willingness of individuals to act prosocially, and therefore measure precisely how much this willingness is eroded once decisions are made jointly with others.

We study two types of antisocial messages: a *truthful* antisocial message that reveals the option

[4]The remaining eight options are Pareto-dominated and pay all players a much smaller amount. Hence, receivers have an incentive to follow the senders' message, even if they know the are receiving the antisocial message.

with unequal payoffs and a *deceptive* antisocial message that points to the option with unequal payoffs but claims it is revealing the option with equal payoffs. Having two different messages in a comparable setting allows us to determine whether changes in behavior and normative beliefs depend on the type of antisocial action that is available. To study whether differences in behavior and normative beliefs are due to the active participation of senders in the decision-making, we also employ a *Passive-Sender* game that is identical to the *2-Sender* game in terms of the number of players and payoffs but, like the *1-Sender* game, has only one player determining which message is sent.

In line with previous literature, we see more antisocial behavior when a choice is made jointly with others. Specifically, antisocial premiums are significantly lower when there are two senders compared to when there is only one. Hence, as Dana et al. (2007), we find that the simple presence of another decision-maker is sufficient to increase the willingness of individuals to act antisocially. In addition, we find that the difference in antisocial premiums is present and of similar magnitude, irrespective of whether the antisocial message is truthful or deceptive.

More importantly, we find evidence that is consistent with a shift in normative views being the reason why joint decisions are more antisocial in the *2-Sender* game. First, we find that subjects in the *2-Sender* game think that sending the antisocial message is significantly more acceptable than subjects in the *1-Sender* game. Importantly, this view is held by both senders and receivers, which demonstrates that the shift in normative views occurs on a general level and is not the result of senders' internal justification of their own choices. Second, we find that senders who end up sending the antisocial message feel less guilt in the *2-Sender* game compared to the *1-Sender* game.

Analyzing the data at the individual level reveals that both normative views and the sender's belief of how much they will disappoint the receiver are essential determinants of antisocial premiums. In particular, and consistent with models of guilt aversion, we find antisocial premiums are affected by the interaction of these two variables. In other words, the senders who act prosocially are those who think antisocial actions are very unacceptable and who believe receivers expect senders to act prosocially.

Finally, we find that antisocial premiums and the unacceptability of sending the antisocial message are higher in the *Passive-Sender* game than in the *2-Sender* game and are very similar to the antisocial premiums and normative vies in the *1-Sender* game. This result shows that the active involvement of the second sender in the decision-making process is crucial for the increase in antisocial behavior in groups to occur.

2 Related Literature

This paper builds on previous studies that investigate how interacting with others affects one's proclivity to act antisocially. Increased antisocial behavior in this regard has been mainly studied in two circumstances: when individuals trade in a market for an antisocial action and when decisions are jointly made in groups.[5] In addition, our work is related to research that assesses whether elicited normative beliefs are associated with behavior.

2.1 Increased antisocial behavior via market interactions

Falk and Szech (2013) show that subjects are more inclined to accept the death of a mouse in return for money when the monetary amount is determined in a bilateral or multilateral double-auction market than when they make their decisions individually. Kirchler et al. (2016) confirm this finding by testing how different interventions concerning morality affect donation decisions differently if they are done with individual choice lists or through double-auction markets. The result that market environments lead to an erosion of morality is further supported by Deckers et al. (2016), who find that personal characteristics, which predict antisocial behavior when decisions are made individually, lose their predictive power in market settings. Finally, Bartling et al. (2015) examine the willingness to maximize own payoffs at the cost of a third player in markets with subjects from countries with low (China) and high (Switzerland) degrees of market-orientation. They find that subjects are less socially responsible in market settings compared to individual decisions and that this pattern is more pronounced among subjects with a lower degree of market-orientation.

Unlike these studies, senders in our games make their decisions independently and simultaneously. Hence, information about the normative views of others is not revealed through market interaction. Moreover, we do not use a substantially different framing between individual and joint decisions (i.e., individual choice vs. market trading) that could fundamentally change how the subjects think about the acceptability of the antisocial outcome. In other words, we concentrate solely on the effect of the inclusion of another decision-maker on the willingness of individuals to behave antisocially.

[5] Increased antisocial behavior due to the involvement of others is also observed in settings where responsibility is transferred through delegation or intermediation. For example, delegation leads to lower payoffs for receivers in dictator games (Hamman et al., 2010; Bartling and Fischbacher, 2012) and intermediation leads to less third-party punishment of unfair allocations, resulting in more selfish decisions (Coffman, 2011; Oexl and Grossman, 2013; Garofalo and Rott, 2017).

2.2 Increased antisocial behavior via group decisions

The second line of literature investigates whether behavior is more antisocial when decisions are made in groups compared to individually. Evidence from multiple economic games provides only mixed support to this assertion.

Various studies have found increased antisocial behavior in groups. In the context of dictator games, Dana et al. (2007) find that two individuals making a joint decision give less than single dictators. Moreover, there is some evidence that groups act more selfishly as proposers in ultimatum games (Bornstein and Yaniv, 1998), as trustees in trust games (Cox, 2002), and as workers in gift-exchange games (Kocher and Sutter, 2007). In the context of deception and lying, Keck (2014) provides evidence that receivers who make joint decisions in an ultimatum game are more likely to deceive the proposer than single receivers. Kocher et al. (2018) find that groups are more willing to lie than individuals in a die-rolling game, even when group members are paid only according to their own decisions.[6] Other variations of the anonymous die-rolling game with joint decisions provide further evidence of comparatively more dishonest acts by groups (Gino et al., 2013; Muehlheusser et al., 2015; Weisel and Shalvi, 2015; Korbel, 2017).[7] Nielsen et al. (2017) find that teams are less likely to keep promises made as trustees in a trust game. In sender-receiver games, Sutter (2009) finds that groups are more likely to use strategic deception by telling the truth when they expect receivers will not follow their message, while Cohen et al. (2009) find that senders making joint decisions deceive receivers more often than individual senders when they are informed of the receiver's decision ahead of time. Finally, Falk and Szech (2016) find that the fraction of unethical decisions (whether to kill a mouse in return for money) is higher when subjects are in groups and therefore cannot know whether their decision was pivotal for the killing.

However, there is also evidence of less antisocial behavior by groups than by individuals. For example, Cason and Mui (1997) find more selfish behavior by dictators when they are individuals as opposed to groups (but see the critique by Luhan et al., 2009). In sender-receiver games, in addition to the results reported above, Sutter (2009) also finds that groups of senders lie less than

[6] Several studies report increased dishonesty when lies benefit others through aligned payoffs, both when decisions are made individually (Wiltermuth, 2011; Gino and Pierce, 2010; Conrads et al., 2013) and jointly (Gino et al., 2013; Weisel and Shalvi, 2015). By contrast, Danilov et al. (2013) find that aligned payoffs result in more dishonesty only when social ties are strong among players.

[7] Barr and Michailidou (2017) use a die-rolling game to study the willingness of individuals to lie to benefit another player, knowing that the other player is making the same lying decision to benefit them. They find more lying when individuals make this lying decision compared to a control group where the beneficiary of the lie is a passive player. Although related, this study is different from ours and others in this literature in that there is no joint antisocial decision to be made.

single senders when they expect receivers will follow their message and Cohen et al. (2009) find that groups deceive less than individuals when the receivers' behavior is unknown at the time decisions whether to lie or not are made.

When comparing individuals to groups, many reasons can explain why groups might act more antisocially. By using the strategy method to elicit the senders' willingness to act antisocially and by eliminating all communication between decision-makers, we can rule out that the senders' choices are driven by strategic considerations (e.g., because of the aggregation rule used in the group decision-making process) or by peer and argumentation effects.

2.3 Normative beliefs and behavior

The third line of research to which our paper contributes is the rapidly growing work on the association between normative beliefs and behavior. Although there has been some discussion about the role of norms in economics for some time (Elster, 1989), studies that empirically test the relationship between directly-elicited normative beliefs and behavior have emerged only recently.

Elicited normative beliefs have been shown to affect behavior in numerous contexts. For instance, they have been found to help explain dictator giving (Bicchieri and Xiao, 2009; Krupka and Weber, 2013; Gächter et al., 2017), reciprocity (Gächter et al., 2013), trust (Krupka et al., 2017), bribery (Banerjee, 2016), punishment (Dimant et al., 2018), and discrimination (Barr et al., 2018). Overall, this literature points to normative beliefs being strong motivators of behavior in social contexts.

To the best of our knowledge, our paper is the first to study whether normative beliefs help explain differences between decisions made jointly and individually. In addition, this is one of the few papers in economics that provides simultaneous evidence on the links between normative beliefs, second-order beliefs, experienced emotions, and behavior (another example is Reuben and van Winden, 2010).

3 Sender-receiver games

We employ a two-by-two experimental design. We vary the number of senders playing a sender-receiver game, either one or two, and the type of antisocial message the senders can use, either truthful or deceptive.

In our *1-Sender* game, subjects are anonymously paired and are then randomly assigned either the role of the sender or the receiver. The receiver's task is to choose one out of ten options to determine both players' earnings. There is one prosocial option that pays €10 to each player,

Table 1. Example of payoff tables in the sender-receiver games (amounts in euros)

A. *1-Sender* game

Option	A	B	C	D	E	F	G	H	I	J
Sender	4	4	10	4	$17 - x$	4	4	4	4	4
Receiver	0	0	10	0	3	0	0	0	0	0

B. *2-Sender* game

Option	A	B	C	D	E	F	G	H	I	J
Sender A	4	4	10	4	$17 - x$	4	4	4	4	4
Sender B	4	4	10	4	$10 + x$	4	4	4	4	4
Receiver	0	0	10	0	3	0	0	0	0	0

one antisocial option that pays the sender €17 minus an amount $x \in [€0, €6.50]$ and €3 to the receiver, and eight Pareto-dominated options that pay €4 to the sender and €0 to the receiver. At the beginning of the game, the computer randomly labels each of the ten options with a single letter ranging from A to J. Although both players know the payoff consequences of a particular option being chosen, only the sender knows how each of the ten options is labeled. Table 1A is an example of a letter assignment and how this information is presented to the sender.

In the *2-Sender* game, subjects are anonymously matched into groups of three and are then randomly assigned either the role of sender A, sender B, or the receiver. The payoff structure for sender A and the receiver are identical to those of the sender and receiver in the *1-Sender* game. The new player, sender B, receives identical payoffs as sender A in all nine options except in the antisocial option where sender B receives €10 plus the amount x, as seen in Table 1B.

In both games, the only information available to the receiver regarding the label assignment of the ten options is due to a message. In the *1-Sender* game, the sender decides which message is sent to the receiver, whereas in the *2-Sender* game, sender A and sender B jointly make this decision. There are two available messages. The first message, Message I, accurately reveals the label of the prosocial option and reads "Option [letter paying the receiver €10] will earn you 10 euros". The second message, Message II, is one of two types, depending on the *context*. In the *Bitter-pill* context, Message II accurately reveals the label of the antisocial option and reads "Option [letter paying the receiver €3] will earn you 3 euros". In the *Deception* context, Message II is deceptive in that it reveals the label of the antisocial option but claims it is the label of the prosocial option: "Option [letter paying the receiver €3] will earn you 10 euros". Like senders, receivers are aware that there are two available messages and that, in *Deception*, a message can be deceptive. Hence, it is common knowledge that a message always reveals the label of either the prosocial or the

antisocial option and never the label of one of the eight Pareto-dominated options.

Our aim with this design is to let receivers make an informed decision and have well-defined beliefs about the senders' behavior (in contrast to papers based on the design of Gneezy, 2005) while maintaining the senders' incentive to reveal their preferences through their choices. In other words, we selected the payoffs and number of Pareto-dominated options to ensure that senders have a powerful incentive to choose the message that corresponds to their preferred outcome. To see that this is the case, denote $U(A)$ as the sender's utility if the antisocial option is implemented, $U(P)$ as her utility if the prosocial option is implemented, and $U(D)$ as her utility if a dominated option is implemented. Furthermore, let p be the sender's expected probability with which the receiver follows her message. In this case, the sender's expected utility of sending Message I is $pU(P) + (1-p)(1/9)U(A) + (1-p)(8/9)U(D)$ and that of sending Message II is $pU(A) + (1-p)(1/9)U(P) + (1-p)(8/9)U(D)$. It is easy to calculate that, as long as $p > 1/9$, senders who prefer the prosocial option (i.e., for whom $U(P) > U(A)$) are better off choosing Message I and senders who prefer the antisocial option (i.e., for whom $U(P) < U(A)$) are better off choosing Message II. We chose payoffs under which it would be highly unlikely for senders to expect that less than 11% of the receivers follow their message. The experimental data supports our guess. We find that 94% of the receivers followed the message they received and 99% of the senders expected at least 11% of the receivers would follow their message.

3.1 The antisocial premium

We use the strategy method to measure the senders' willingness to send an antisocial message. Specifically, senders choose between Message I and Message II for 14 different values of x. The rows of Table 2A correspond to the choices of senders in the *1-Sender* game, those in Table 2B to the choices of Senders A in the *2-Sender* game, and those in Table 2C to the choices of Senders B in the *2-Sender* game. Senders A and B make their decisions simultaneously. After that, the computer randomly selects one row to determine which message is sent.[8] In the *2-Sender* game, Message II is sent only if both senders choose Message II, and otherwise Message I is sent. In other words, the antisocial message is sent only with the consent of both senders. Importantly, this procedure ensures that senders in the *2-Sender* game always have a (weakly) positive incentive to choose their preferred outcome.

While Message I always pays €10, the payoff of Message II depends on the amount x. By systematically varying x, we measure the minimum monetary compensation senders must receive to be willing to send the antisocial Message II instead of the prosocial Message I. Accordingly, we

[8] When receivers see the message, they are not informed which row was selected by the computer.

Table 2. Senders' choice lists

Note: The table displays the choice lists used by senders. Each row displays the value of x, the sender's payoff if Message I is implemented (in euros), and the sender's payoff if Message II is implemented (in euros). Once choices were made, the computer randomly selected one row to determine which message is sent.

A. Sender in *1-Sender*

Row	x	Message	
		I	II
1	0.00	10.00	17.00
2	0.50	10.00	16.50
3	1.00	10.00	16.00
4	1.50	10.00	15.50
5	2.00	10.00	15.00
6	2.50	10.00	14.50
7	3.00	10.00	14.00
8	3.50	10.00	13.50
9	4.00	10.00	13.00
10	4.50	10.00	12.50
11	5.00	10.00	12.50
12	5.50	10.00	11.00
13	6.00	10.00	11.00
14	6.50	10.00	10.50

B. Sender A in *2-Sender*

Row	x	Message	
		I	II
1	0.00	10.00	17.00
2	0.50	10.00	16.50
3	1.00	10.00	16.00
4	1.50	10.00	15.50
5	2.00	10.00	15.00
6	2.50	10.00	14.50
7	3.00	10.00	14.00
8	3.50	10.00	13.50
9	4.00	10.00	13.00
10	4.50	10.00	12.50
11	5.00	10.00	12.50
12	5.50	10.00	11.00
13	6.00	10.00	11.00
14	6.50	10.00	10.50

C. Sender B in *2-Sender*

Row	x	Message	
		I	II
1	0.00	10.00	10.00
2	0.50	10.00	10.50
3	1.00	10.00	11.00
4	1.50	10.00	11.50
5	2.00	10.00	12.00
6	2.50	10.00	12.50
7	3.00	10.00	13.00
8	3.50	10.00	13.50
9	4.00	10.00	14.00
10	4.50	10.00	14.50
11	5.00	10.00	15.50
12	5.50	10.00	15.00
13	6.00	10.00	16.00
14	6.50	10.00	16.50

call this minimum compensation the senders' *antisocial premium*. More specifically, we classify senders who choose Message 2 over Message 1 for a given x as having an antisocial premium in the interval $[€x - 0.5, €x]$.[9]

3.2 Normative views

In all treatments, we elicit the senders' normative views regarding the prosocial and antisocial messages. Specifically, after they made their decisions and delivered the message but before learning the outcome of the game, we ask senders to indicate for each message "How acceptable do you consider it is to deliver Message I [or Message II] to Player 3 [the receiver]?" Answers are recorded with a 5-point Likert scale ranging from "very unacceptable" (1) to "very acceptable" (5).

[9] At the extremes, we classify senders who always choose Message I as having an antisocial premium in the interval $[€7.50, \infty)$ if they played in the *1-Sender* game or as sender A in the *2-Sender* game, and in the interval $[€7.00, \infty)$ if they played as sender B in the *2-Sender* game. The analogous intervals for senders who always choose Message II are $(-\infty, €0.50]$ and $(-\infty, €0.00]$. We did not classify senders who switched more than once or switched in the wrong direction.

In addition to the senders, we also asked receivers to rate the acceptability of sending each message. Receivers rate each message after they made their choice but before they learned their final earnings. The receivers' normative views are important for two reasons. First, they allow us to evaluate whether the senders' normative views are self-serving. Second, they can tell us whether the inclusion of a second sender affects only the senders' normative perceptions or whether it produces a more general perceptual change.

Finally, we also elicited the senders' belief of the receivers' normative views. To do so, after indicating their normative views, we showed senders the question used to measure the normative views of receivers. After that, we asked them to indicate "What do you think was Player 3's [the receiver's] answer to this question?" We incentivized their answer by paying them €0.25 for each correct guess. With this method, we obtain an incentivized measure of the senders' normative views. An alternative method for obtaining an incentivized measure of normative views is to pay subjects if their normative views coincide with those of most other subjects (i.e., have subjects play a coordination game, see Krupka and Weber, 2013). We opted for a different approach because the methodology of Krupka and Weber (2013) implicitly assumes that there is substantial agreement on how acceptable actions are. Therefore, it is not ideally suited for situations where individuals' social perceptions of what is acceptable differ from their normative views,[10] which could easily be the case in our games given the asymmetry between senders and receivers.

3.3 Guilt

Guilt is a powerful emotion that is triggered when an individual acts in a way that violates that individual's perception of normatively acceptable behavior (Baumeister et al., 1994). As such, it provides us with another way of measuring normative views. We measure the senders' experienced guilt to assess whether their hedonic experience is consistent with their previous choices and normative evaluations and to test whether it differs between the *1-Sender* and *2-Sender* games.

We measure guilt the moment senders see the option implemented by the receiver and the earnings of all the players with whom they are matched. At this point, we ask senders to self-report the intensity at which they experienced guilt on a 7-point Likert scale that ranged from "not at

[10]This limitation is not surprising since the methodology of Krupka and Weber (2013) was designed to study the effect of social norms, which they define as *commonly-shared* beliefs of what is acceptable. We elicit separately the subjects' normative views and their perception of the normative views of others to obtain a broader measure of normative beliefs. Distinguishing between one's normative views and one's belief of the normative views of others can be important since they might have different effects on the behavior (see, Schram and Charness, 2015). See Erkut and Reuben (2019) for a more general discussion on the measurement of preferences, including ways of measuring social norms.

all" (1) to "very intensively" (7).[11] We use a self-reported measure because, to the best of our knowledge, there are no precise physiological measures of guilt (Adolphs, 2002). This is not to say that self-reports do not have limitations. In particular, one might worry that subjects do not report their genuine emotions, and instead, they report a fictitious emotional reaction. Reassuringly, considerable research has demonstrated that self-reported emotional experiences are highly correlated with physiological measures like heart rates, facial movements, and brain activation (e.g., Bradley and Lang, 2000; Ben-Shakhar et al., 2007).

3.4 Beliefs

As we will discuss in more detail in Section 4, a potential mechanism for a difference in behavior between the *2-Sender* game and the *1-Sender* game is for the senders' expectations to vary between the two games. Specifically, models of guilt aversion (Battigalli and Dufwenberg, 2007) can predict more antisocial behavior in the *2-Sender* game if senders believe that receivers expect them to send Message II more often in that game compared to the *1-Sender* game. Therefore, we elicit the senders' beliefs about the receivers' expected probability of receiving Message II.[12] To do so, we first elicit the receivers' expected probability of receiving Message II by asking them "How many of the Players 1 [or pairs of Players 1 and 2, which are senders] in the room do you think will deliver Message II?" Thereafter, we show this question to the senders and ask them to indicate "What do you think was Player 3's [the receiver's] answer to this question?" These elicitations are incentivized by paying senders €0.25 and receivers €0.75 per correct answer.[13]

3.5 Procedures

We ran the experiment between February and March 2015 at the Laboratory of Experimental Economics (LEE) at University Jaume I in Castellon, Spain. A total of 197 undergraduate students, including 112 men and 85 women from different faculties were recruited using ORSEE (Greiner, 2015). We conducted eight sessions, each lasting around one and a half hours.

[11] We also measured the senders experienced intensities of shame, anger, happiness, and gratitude. Although guilt is the emotion of interest in our study, and indeed has the most explanatory power, we measured multiple emotions to minimize experimenter demand effects. Evidence that guilt was not excessively salient is that senders typically report higher intensities for other emotions: 90% of senders report experiencing at least one emotion with a strictly higher intensity than guilt.

[12] We also elicited the sender's beliefs about the receivers' behavior by asking them "How many of the Players 3 [receivers] in the room will follow the message they received?"

[13] We paid receivers a higher amount for an accurate answer to partly compensate them for their potentially lower earnings. Subjects were not aware of this difference.

Upon arrival, the subjects were randomly assigned to computer terminals. Thereafter, the instructions for the experiment were read aloud by the experimenter, and subjects were asked to answer a series of control questions (instructions are available in the Appendix). Subjects could ask questions during the whole process. The experiment was run using z-Tree (Fischbacher, 2007).

Once senders made a decision for each of the 14 values of x (see Table 2), the computer randomly selected one of these values and displayed the text of the chosen message on the senders' screen. All senders in the *1-Sender* game and Senders B in the *2-Sender* game were asked to write down the message on a blank sheet of paper located on their desk and then wait for an experimenter to arrive at their desk. The experimenter checked whether the written message coincided with the text on the screen and then guided the sender to their receiver's desk. The sender handed the sheet over to the receiver and then returned to his/her seat. During the delivery process, the experimenter ensured that there would not be any communication between senders and receivers. All subjects were informed about the delivery process in the experimental instructions and knew that communication with other subjects implied not getting paid. Once all senders returned to their desk, receivers were asked to type into the computer screen the message they received and to choose one of the ten options to determine the final earnings of each player. Once the experiment ended, subjects were paid in cash. Average earnings were around €15, including a €5 show-up fee.

4 Hypotheses

In this section, we propose testable hypotheses to guide the data analysis. Our first hypothesis is based on the empirical literature described in the previous section. The literature's main finding is that, more often than not, people making joint decisions end up choosing more antisocial actions than individuals deciding alone. In other words, our first hypothesis simply states that we expect to replicate the most common finding in the literature.

Hypothesis 1 *On average, antisocial premiums are lower in the 2-Sender game than in the 1-Sender game.*

Our subsequent hypotheses are constructed to test the idea that normative beliefs explain the difference in behavior between the *1-Sender* and *2-Sender* games. Note that we do not attempt to compare models of normative beliefs to models that incorporate other motivations, such as models of social preferences (for such an attempt, see Gächter et al., 2013).[14] Instead, our approach is

[14]Broadly speaking, there are three types of models of social preferences that can explain a decrease in antisocial premiums from the *1-Sender* to the *2-Sender* games. The first type includes models in which other-regarding

to measure variables that are explicitly modeled in models of normative beliefs, and then test whether these empirical measures vary between the *1-Sender* and *2-Sender* games in ways that are consistent with normative beliefs being an explanation for the differences in behavior.

Our hypotheses are based on the models of social norms (Krupka and Weber, 2013; Barr et al., 2018, e.g.,). In these models, individual maximize a utility function that includes their monetary payoff and how acceptable they think different actions are. In other words, on their normative views. Cases in which there is broad consensus on the acceptability of actions can be said to be cases where a social norm exists (Bicchieri, 2006). If we assume that sending Message II is less acceptable than sending Message I, then it follows that the senders' antisocial premium depends on how they trade-off the higher monetary payoff of the antisocial outcome with the lower social acceptability of sending Message II. Given that we elicit antisocial premiums for the same monetary payoffs in the *1-Sender* and *2-Sender* games, if there is support for Hypothesis 1, then it must be the case that there is a difference in normative views. This line of thought gives us our second hypothesis.

Hypothesis 2 *Conditional of finding support for Hypothesis 1, on average, senders rate sending the antisocial Message II as more acceptable in the 2-Sender game than in the 1-Sender game.*

Our third hypothesis concerns the senders' experienced feelings of guilt. Although models such as Krupka and Weber (2013) and Barr et al. (2018) do not typically refer to guilt, they assume that choosing an unacceptable action reduces one's utility compared to choosing a more acceptable action. Since an extensive literature in psychology argues that individuals feel guilt if they act in unacceptable ways (Baumeister et al., 1994), it is natural to assume that the difference in utility between acceptable and unacceptable actions is due to differences in the intensity with which individuals feel guilt. Given that there is no reason to feel guilty if one sends the prosocial message in either game, it follows that support for Hypothesis 1 implies a difference between the *1-Sender* and *2-Sender* games in the intensity with which senders feel guilt if they send the antisocial message.

concerns depend on the number of players (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002). In these models, the inclusion of a second sender implies that the weight senders place on receiver's welfare is lowered. The second type includes models that assume that an individual's utility from an outcome depends on how pivotal their choice is in determining the outcome (Engl, 2017; Rothenhäusler et al., 2018). In these models, an increase in the number of decision-makers that makes individual choices less pivotal lead to more antisocial behavior. The third type includes models in which the players' concern for others depends on their beliefs (Geanakoplos et al., 1989). Although these models often complex, it is straightforward to understand that they predict a change in behavior if the inclusion of the second sender alters the senders' expectations. This type of models includes models of guilt aversion (Battigalli and Dufwenberg, 2007), which we think are conceptually related to normative beliefs. As such, we discuss them in more detail in the text.

Hypothesis 3 *Conditional of finding support for Hypothesis 1, on average, senders who send the antisocial message experience lower intensities of guilt in the 2-Sender game than in the 1-Sender game.*

Our fourth hypothesis further elaborates on anticipated guilt as the emotion that deters senders from sending the antisocial message. Recent work in economics argues that beliefs are a determining factor in triggering guilt. In particular, in models of (simple) guilt aversion (Battigalli and Dufwenberg, 2007), the senders' guilt depends on two factors: their sensitivity to guilt and the degree to which they think the antisocial outcome will disappoint the receiver (i.e., senders experience more guilt for sending Message II the more they believe the receiver expects to receive Message I). In other words, these models suggest that antisocial premiums could change from the *2-Sender* to the *1-Sender* game solely due to changes in the senders' second-order beliefs. Specifically, support for Hypothesis 1 is consistent with models of guilt aversion if the senders' second-order beliefs are higher in the *2-Sender* game than in the *1-Sender* game.[15]

Hypothesis 4 *Conditional on finding support for Hypothesis 1, on average, the senders' belief of the receivers' expectation of receiving the antisocial message is higher in the 2-Sender game compared to the 1-Sender game.*

Thus far, we have proposed hypotheses concerning a difference in behavior between the *1-Sender* and *2-Sender* games. Our final hypothesis focuses on individual differences within games. The arguments provided above give us two precise predictions. In each game, the arguments supporting Hypothesis 2 imply a negative relationship between the senders' antisocial premiums and their ratings of the acceptability of sending Message II. In other words, senders who consider sending the antisocial message is more acceptable need to receive less monetary compensation for sending it.[16] The arguments supporting Hypothesis 4 imply a negative relationship between the senders' antisocial premiums and their second-order belief. Specifically, in both games, senders who believe the receiver expects to receive Message II feel less guilt from sending Message II and therefore require less money to send it.

[15] Note that the converse if not necessarily true. Models of guilt aversion can be consistent with support for Hypothesis 1 even if the senders' second-order beliefs are equal across games. This would be the case if including a second sender leaves beliefs unchanged but decreases the senders' guilt sensitivity.

[16] Note that the relationship between antisocial premiums and experienced guilt is not as easy to detect as that between antisocial premiums and normative views. In theory, senders with high antisocial premiums experience more guilt if they send Message II. However, a high antisocial premium also implies that they are also less likely to send Message II. Since we observe guilt only if Message II is sent, the observed relationship between experienced guilt and antisocial premiums will be much smaller than the real relationship between these two variables.

In addition to these straightforward predictions, we can also think about the relationship between normative views and second-order beliefs. Although Battigalli and Dufwenberg (2007) do not specify where guilt sensitivity comes from, we think that the normative views of individuals are a natural interpretation of their guilt sensitivity. In our games, this interpretation states that, for a given set of beliefs, a sender feels more guilt from choosing the antisocial message the more unacceptable they think sending Message II is.[17] If the elicited normative views capture the senders' sensitivity to guilt, then models of guilt aversion predict a positive relationship between antisocial premiums and the *interaction* of normative views and second-order beliefs. In other words, the senders with the highest antisocial premiums are senders who think sending Message II is very unacceptable and believe receivers expect to receive Message I. These predictions constitute our last formal hypothesis.

Hypothesis 5 *Within each game, antisocial premiums are negatively correlated with the senders' acceptability ratings of sending the antisocial message and their belief of the receivers' expectation of receiving the said message. In addition, antisocial premiums are positively correlated with the interaction of the senders' acceptability ratings and their beliefs.*

5 Results

A total of 197 subjects participated in our experiment, 118 as senders and 79 as receivers. Of all the senders, 7 senders switched more than once and 1 sender switched from Message II to Message I as the premium for Message II increased. Since it is not clear what the antisocial premium of these subjects is, we excluded them from the statistical analysis. However, our results are robust to including these subjects and using the number of Message II choices as a measure of antisociality. This exclusion leaves us with 39 senders in the *1-Sender* game (19 in *Bitter-pill* and 20 in *Deception*) and 71 senders in the *2-Sender* game (35 in *Bitter-pill* and 36 in *Deception*).

5.1 The antisocial premium

Figure 1 plots the cumulative distributions of the senders' antisocial premiums in the *1-Sender* and *2-Sender* games, pooling the *Bitter-pill* and *Deception* contexts. Figure 2 depicts the mean antisocial premium depending on the number of senders as well as the context.

[17] This interpretation of models of guilt aversion is also broadly consistent with the argument that social norms are followed only if there is an expectation that sufficiently many others follow them (Bicchieri, 2006). That is to say, even though senders might consider sending Message I is the acceptable action, they will feel guilty from sending Message II only if receivers expect them to choose Message I.

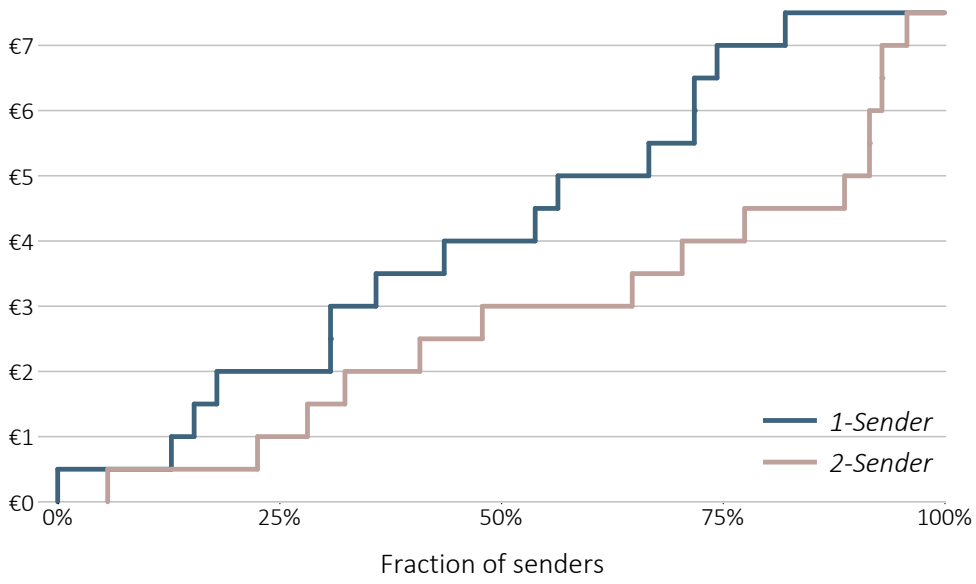


Figure 1. Cumulative distributions of senders' antisocial premiums

Consistent with the literature, the figures reveal that many senders are willing to forego large profits to act prosocially. More interestingly, having a second sender lowers antisocial premiums. On average, senders in the *2-Sender* game require €1.40 less for sending Message II than senders in the *1-Sender* game (€1.49 less in *Bitter-pill* and €1.33 less in *Deception*). This difference is substantial, considering that the overall mean antisocial premium across both games is only €3.28.

To evaluate whether these differences are statistically significant, we use interval regressions with the senders' antisocial premium as the dependent variable. These regressions allow us to account for the fact that if a sender switches from Message I to Message II when the latter pays more than €x, then we know that her antisocial premium lies in the interval [€$x - 0.50$, €x] (Stewart, 1983).[18] All regressions are estimated using robust standard errors and are found in Table A1 of the Appendix. Lastly, since we have a clear directional hypothesis, we report p-values of one-tailed tests.

Consistent with Hypothesis 1, we find that antisocial premiums are significantly lower in the *2-Sender* game compared to the *1-Sender* game ($p = 0.003$ overall; $p = 0.016$ in *Bitter-pill*; $p = 0.022$ in *Deception*).[19] These findings establish our first result.

[18] At the extremes, senders who always choose Message I have an antisocial premium in the interval [€$7.50, \infty$) if they played in the *1-Sender* game or as sender A in the *2-Sender* game. If they played as sender B in the *2-Sender* game, then their antisocial premium is in the interval [€$7.00, \infty$). The analogous intervals for senders who always choose Message II are (∞, €0.50] and (∞, €0.00]. These results are robust if we use linear or ordered probit regressions instead of interval regressions.

[19] A difference-in-differences test reveals that the difference between the *1-Sender* and *2-Sender* games does not differ

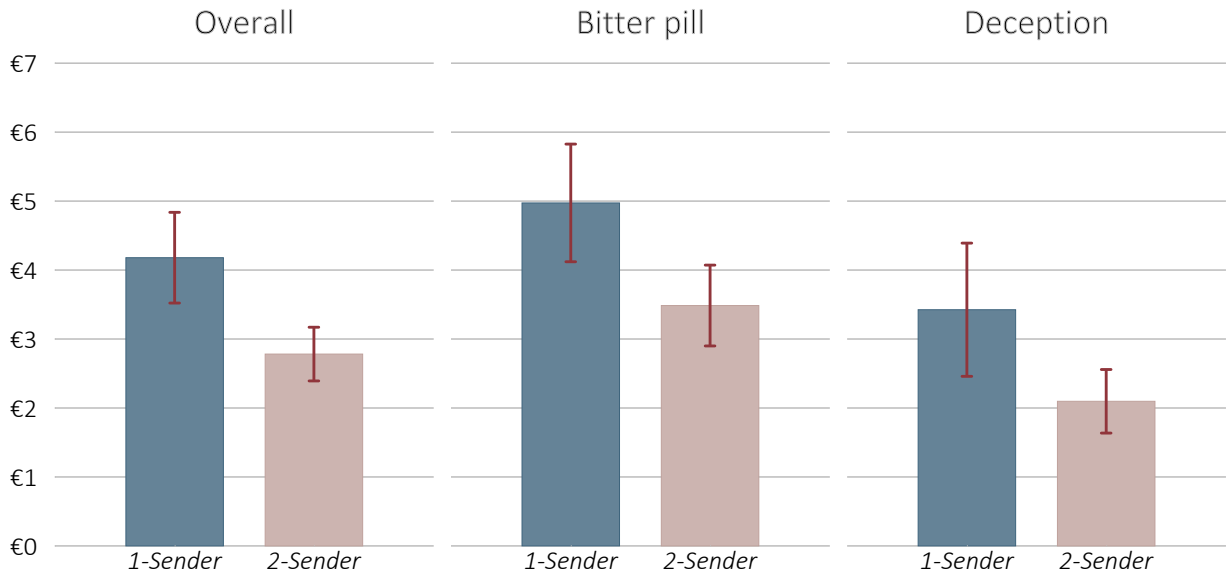


Figure 2. Senders' mean antisocial premium depending on the game and context

Note: Error bars correspond to 90% confidence intervals.

Result 1 *The involvement of a second sender significantly lowers antisocial premiums in both Bitter-pill and Deception contexts.*

5.2 Normative views

Next, we test the hypothesized effects of having a second sender on the subjects' normative views. Figure 3 presents the senders' mean acceptability ratings of sending Message II, the senders' mean belief of the receivers' acceptability ratings, and the receivers' mean acceptability ratings. More detailed summary statistics are reported in Table 3 as well as in Table A2 of the Appendix. Given that normative views are discrete, ranging from very unacceptable (1) to very acceptable (5) (see Section 3.2), we use ordered probit regressions to test whether differences between games are statistically significant. The regression coefficients are provided in Table A3 of the Appendix. As before, since there is a directional hypothesis, we report p-values of one-tailed tests.[20]

In line with Hypothesis 2, we find that senders in *2-Sender* games think it is significantly more acceptable to send Message II than senders in *1-Sender* games (2.97 vs. 2.33, $p = 0.004$).

between *Bitter-pill* and *Deception* ($p = 0.993$). The difference in antisocial premiums between *Bitter-pill* and *Deception* is close to statistical significance in the *1-Sender* game ($p = 0.074$) and is significant in the *2-Sender* game ($p = 0.003$).

[20]Since the senders' normative views were elicited after the message delivery, we cluster standard errors in the *2-Sender* game on the matched pairs. Results are robust to the use of linear regressions instead of ordered probit regressions.

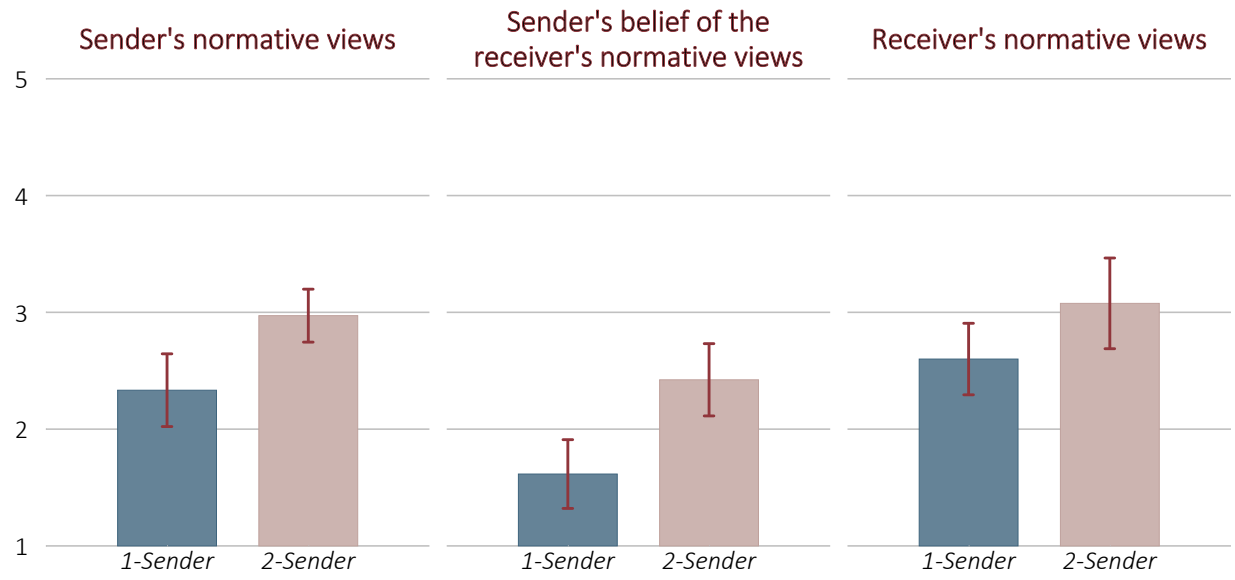


Figure 3. Subjects' mean acceptability of sending Message II (from very acceptable [1] to very unacceptable [5])

Note: Error bars correspond to 90% confidence intervals.

Interestingly, the receivers' acceptability ratings are also higher in the *2-Sender* games (3.08 vs. 2.60, $p = 0.051$), as are the senders' beliefs of the receivers' acceptability ratings (2.42 vs. 1.62, $p = 0.004$).[21] Hence, consistent with Hypothesis 2, all three measures of the subjects' normative views indicate it is more acceptable to send Message II in the *2-Sender* games compared to the *1-Sender* games.

A common concern with self-reported measures, such as the subjects' normative views, is that they could be self-serving and depend on their experience in the game. We argue that this is not the case for our elicited normative views for four reasons. First, antisocial messages are rated as more acceptable in the *2-Sender* games than in the *1-Sender* games not only by senders but also by receivers. This fact suggests that the change in normative views is not due to self-serving reporting by senders.[22] Second, we also observe a difference between games in the senders' beliefs of the receivers' acceptability ratings, which is an incentivized measure and therefore less susceptible to misreporting. Third, there is no evidence that senders in *1-Sender* games rate the acceptability

[21] This pattern persists when we examine each context separately. In *Bitter-pill*, the difference in acceptability ratings between the *1-Sender* and the *2-Sender* games is 0.69 for senders ($p = 0.011$), 0.69 for the senders' belief of the receivers' acceptability ratings ($p = 0.063$), and 0.35 for receivers ($p = 0.190$). In *Deception*, the difference between the *1-Sender* and the *2-Sender* games is 0.59 for senders ($p = 0.050$), 0.92 for the senders' beliefs of the receivers' acceptability ratings ($p = 0.010$) and 0.63 for receivers ($p = 0.058$).

[22] We find that normative views of receivers do not differ significantly from those of senders in either the *1-Sender* ($p = 0.223$) or the *2-Sender* games ($p = 0.727$).

Table 3. Means and standard deviations of subjects' normative views and beliefs

	Overall		***Bitter-pill***		***Deception***	
	1-Sender	*2-Sender*	*1-Sender*	*2-Sender*	*1-Sender*	*2-Sender*
	Normative views of sending Message II					
Senders' normative views	2.33	2.97	2.11	2.80	2.55	3.14
	(1.15)	(1.15)	(1.05)	(1.08)	(1.23)	(1.20)
Senders' belief of the receivers'	1.62	2.42	1.74	2.43	1.50	2.42
normative views	(1.09)	(1.56)	(1.28)	(1.54)	(0.89)	(1.61)
Receivers' normative views	2.60	3.08	2.30	2.65	2.90	3.53
	(1.15)	(1.44)	(0.92)	(1.53)	(1.29)	(1.22)
	Belief of receiving Message II					
Senders' expectation of the	0.57	0.56	0.52	0.49	0.62	0.63
receivers' belief	(0.32)	(0.29)	(0.31)	(0.28)	(0.33)	(0.30)
Receivers' belief	0.56	0.50	0.57	0.34	0.55	0.66
	(0.29)	(0.28)	(0.28)	(0.26)	(0.29)	(0.20)

of Message I differently from senders in *2-Sender* games ($p = 0.290$), which one would expect if normative evaluations were affected by the subjects' experience in the game. Fourth, we also ran the regressions reported above, controlling for subjects' experience up to the point where they reported their normative views. To be precise, we controlled for (i) which one of the two messages was actually delivered, (ii) the choice of the other sender in the *2-Sender* games, and (iii) the senders' earnings if the receiver follows the message. We find, once again, that the senders' normative views, as well as their beliefs about the receivers' normative views, significantly differ between the *1-Sender* and *2-Sender* games ($p = 0.009$ and $p = 0.034$ respectively). Moreover, the control variables are neither jointly significant ($p > 0.770$) nor individually significant ($p > 0.427$) in either regression (see Table A4 of the Appendix).

5.3 Guilt

Next, we analyze the senders' emotional reaction. This analysis can be used to corroborate that the senders' hedonic experience is consistent with their behavior and normative views across the two games. Our analysis focuses on the amount of guilt senders experience when they see the outcome of the game, depending on whether they sent the antisocial or the prosocial message.[23]

[23]In the *1-Sender* game, 22 Message I's and 17 Message II's were delivered; while in the *2-Sender* game, 27 Message I's and 11 Message II's were delivered. We drop the six instances where the outcome was not a direct consequence of the senders' choices because the receiver chose a different option from the one suggested in the message. Our

However, we provide summary statistics for all emotions in Table A6 of the Appendix. Emotions were measured in a 1 to 7 Likert scale.

On average, senders experience significantly more guilt after sending Message II than after sending Message I. The difference is substantial: 4.23 vs. 1.09 in the *1-Sender* game and 2.68 vs. 1.62 in the *2-Sender* game ($p < 0.026$). More importantly and consistent with Hypothesis 3, we also find that senders experience significantly less guilt after sending Message II in the *2-Sender* game compared to the *1-Sender* game (2.68 vs. 4.23, $p < 0.003$).[24] Hence, the effect of a second sender is not only evident in the senders' behavior and normative views but also in their emotional state.[25] In the Appendix, we analyze the association between the senders' guilt and their normative views. We find that senders who deliver Message II experience more guilt the more they consider that sending Message II is normatively unacceptable. Interestingly, this effect is stronger in the *1-Sender* game compared to the *2-Sender* game.[26] We summarize these findings as our second result.

Result 2 *If a second sender is involved, both senders and receivers think it is more normatively acceptable to send Message II. Moreover, if they do send Message II, senders experience less guilt when a second sender is present.*

5.4 Second-order beliefs

Now, we turn to senders' belief of the receiver's expected probability of receiving Message II, to which we refer to as the senders' *second-order belief*. Figure 4 depicts the senders' second-order belief and the receivers' actual expected probability of receiving Message II (see Table 3 for more detailed summary statistics). We use Tobit regressions to test for differences across games as belief responses are censored at 0% and 100%. The regression coefficients are provided in Table A5 of

results remain unchanged if we include these observations.

[24] We obtain these p-values from linear regressions of the senders' experienced guilt (see Table A10 of the Appendix). The p-values for the test of directional Hypothesis 3 are of a one-tailed test.

[25] Once again, a possible concern may be that these differences are due to using a self-reported measure. For instance, one might worry that senders do not report their genuine emotions, and instead, they report the emotional reaction they think the experimenter expects. We believe that this is an unlikely explanation for the difference between games. That is, it is hard to see how subjects could anticipate that the 'expected' emotional reaction is more guilt for Message II in the *1-Sender* than in the *2-Sender* game when subjects took part in only one of these games.

[26] This analysis is found Table A7. Although our interest is on guilt, we also performed a similar analysis for the other elicited emotions. Table A8 contains the regression results of other negative emotions (shame and anger) and Table A9 of positive emotions (happiness and gratitude). By and large, these results are in line with the results for guilt.

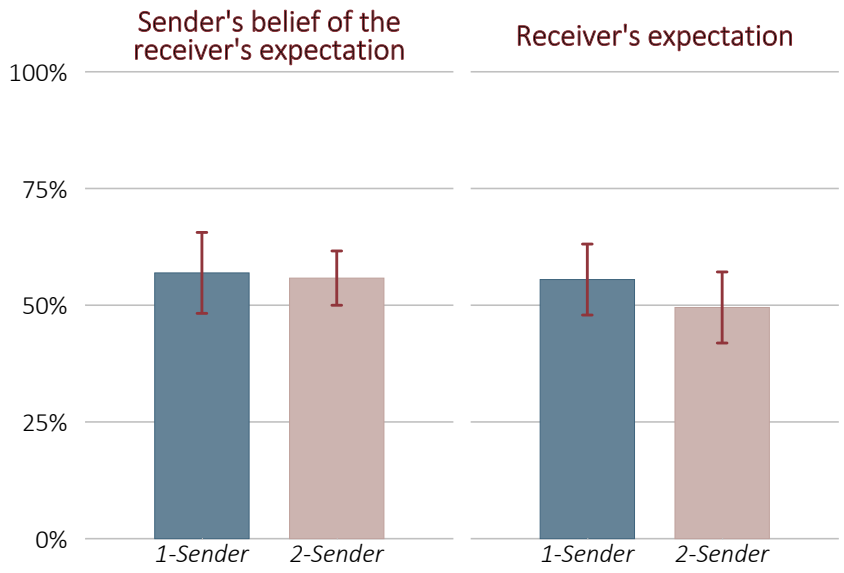


Figure 4. Subjects' expected probability of receiving Message II

Note: Error bars correspond to 90% confidence intervals.

the Appendix.[27]

On average, senders think that receivers expect to receive Message II with probability 0.57 in the *1-Sender* games and 0.56 in the *2-Sender* games ($p = 0.628$). The senders' beliefs are reasonably accurate as receivers expect to receive Message II with probability 0.56 in the *1-Sender* game and 0.50 in the *2-Sender* game ($p = 0.835$). Hence, contrary to Hypothesis 4, we do not find evidence that senders' second-order beliefs are affected by the involvement of a second sender.[28] These findings establish our third result.

Result 3 *The senders' belief of the receivers' expected probability of receiving Message II remain unchanged with the involvement of a second sender.*

5.5 Determinants of the antisocial premium

To further understand the determinants of antisocial premiums, we conduct a series of interval regressions of the senders' antisocial premiums (reported in Table 4). In specification I, we include the senders' normative views, their second-order beliefs, and a dummy variable indicating whether it is the *Deception* or *Bitter-pill* context. This first specification allows us to test whether variation

[27] As before, we cluster standard errors on the matched sender pairs in the *2-Sender* game and report p-values of one-tailed tests. Results are robust to the use of linear regressions instead of Tobit regressions.

[28] If we look at each context separately, we find that, compared to the *1-Sender* game, second-order beliefs in the *2-Sender* game are significantly higher in neither *Bitter-pill* ($p = 0.683$) nor *Deception* ($p = 0.513$).

in normative views and second-order beliefs within games, helps us explain the observed variation in antisocial premiums. In specification II, we add the interaction term of senders' normative views with their second-order beliefs. This specification is inspired by models of guilt aversion, which predict that the senders' antisocial premiums depend on their belief of how much they are disappointing receivers and their guilt sensitivity (Battigalli and Dufwenberg, 2007). Finally, in specification III, we include the following set of control variables: (i) the sender's expected probability that the receiver will implement the option mentioned in the message, (ii) the sender's gender, (iii) age, (iv) age squared, and (v) whether the sender was sender B in the *2-Sender* game.[29] In all regressions, we cluster standard errors on matched sender pairs in the *2-sender* game.

We see a similar pattern across the *1-Sender* and the *2-Sender* games that is broadly consistent with Hypothesis 5. First, normative views have a negative effect on antisocial premiums. That is to say, the premium senders require to send Message II is lower the more acceptable they perceive sending Message II is. Second, the negative coefficients on second-order beliefs indicate that senders are more willing to send the antisocial message if they believe that the receiver expects to receive that message. Third, the interaction between normative views and second-order beliefs is positive, although it is statistically significant only in the *2-Sender* games. In other words, second-order beliefs have a stronger effect on the behavior of senders who think sending Message II is unacceptable compared to senders who think that sending Message II is acceptable. This general pattern is consistent with models of guilt aversion, which predict that senders avoid disappointing the receiver only when their guilt-sensitivity is high.[30]

Despite the similar general pattern, there is a noticeable difference between the *1-Sender* and *2-Sender* games. To visualize this difference in Figure 5, we use the coefficients of specification III to plot the estimated mean antisocial premium depending on the senders' second-order belief at two different values of their normative views (keeping all other variables at their mean). The dark blue line shows the relationship between antisocial premiums and second-order beliefs for senders who think sending Message II is "very unacceptable" (10^{th} percentile of the distribution of normative views), while the light brown line depicts the same relationship for senders who think sending Message II is "very acceptable" (90^{th} percentile of the distribution of normative views).

The key difference between the *1-Sender* and *2-Sender* games lies in the slopes of the light brown

[29] We standardized the control variables so that the constant is comparable across specifications II and III. We present the coefficients and standard errors of these control variables in Table A11 of the Appendix.

[30] We also conducted regressions substituting the senders' normative views with their belief about the receivers' normative views. We present the results in Table A12 of the Appendix. We find that the general pattern is similar to the one reported above. However, the interaction between normative views and second-order beliefs is even weaker in the *1-Sender* game.

Table 4. Determinants of the antisocial premium

Note: Interval regressions of senders' antisocial premiums. 'Normative views' are the senders' appropriateness ratings of sending Message II; 'Second-order beliefs' are the senders' expectation of the receivers' belief of receiving Message II; 'Deception context' is a dummy variable indicating the *Deception* context. Robust standard errors clustered on matched pairs of senders are presented in parentheses. ***, **, and * indicate statistical significance at 0.01, 0.05, and 0.10 using two-tailed tests.

	***1-Sender* game**			***2-Sender* game**		
	I	II	III	I	II	III
Normative views	−0.39	−0.98*	−1.03*	−0.20	−1.06***	−1.00**
	(0.33)	(0.50)	(0.55)	(0.25)	(0.39)	(0.39)
Second-order belief	−6.32***	−9.15***	−9.39***	−3.01***	−7.25***	−7.50***
	(1.48)	(2.68)	(3.02)	(0.91)	(1.80)	(1.83)
Normative views × Second-order belief		1.13	1.20		1.44***	1.31**
		(0.93)	(1.01)		(0.53)	(0.53)
Deception context	−1.00	−0.94	−1.04	−1.08**	−0.89**	−0.68*
	(0.76)	(0.76)	(0.77)	(0.44)	(0.39)	(0.36)
Constant	9.02***	10.48***	10.76***	5.26***	7.66***	10.29***
	(1.08)	(1.47)	(1.77)	(0.97)	(1.38)	(1.62)
Additional controls	No	No	Yes	No	No	Yes
Observations	39	39	39	71	71	71
χ^2	25.49	30.42	40.57	23.61	34.20	64.74

lines. Senders in the *1-Sender* games who think that it is acceptable to send Message II nonetheless increase their antisocial premium to avoid disappointing the receiver. By contrast, senders with these normative views in the *2-Sender* game appear to ignore the receiver's expectations.[31] Hence, even though the introduction of a second sender does not impact mean second-order beliefs (see Result 3), it does change the relationship between the second-order beliefs of some senders and their antisocial premiums. These findings are stated as our fourth result.

Result 4 *Normative views and second-order beliefs are both critical determinants of the senders' antisocial premiums. The relationship between second-order beliefs and antisocial premiums weakens between the 1-Sender and 2-Sender games for senders who view sending Message II as acceptable.*

[31] When testing whether the estimated social premiums differ between the two games, we find that senders who think that it is acceptable to send Message II have significantly higher antisocial premiums in the *1-Sender* game as long as their second-order belief is less than 0.59 (Wald tests, $p < 0.05$).

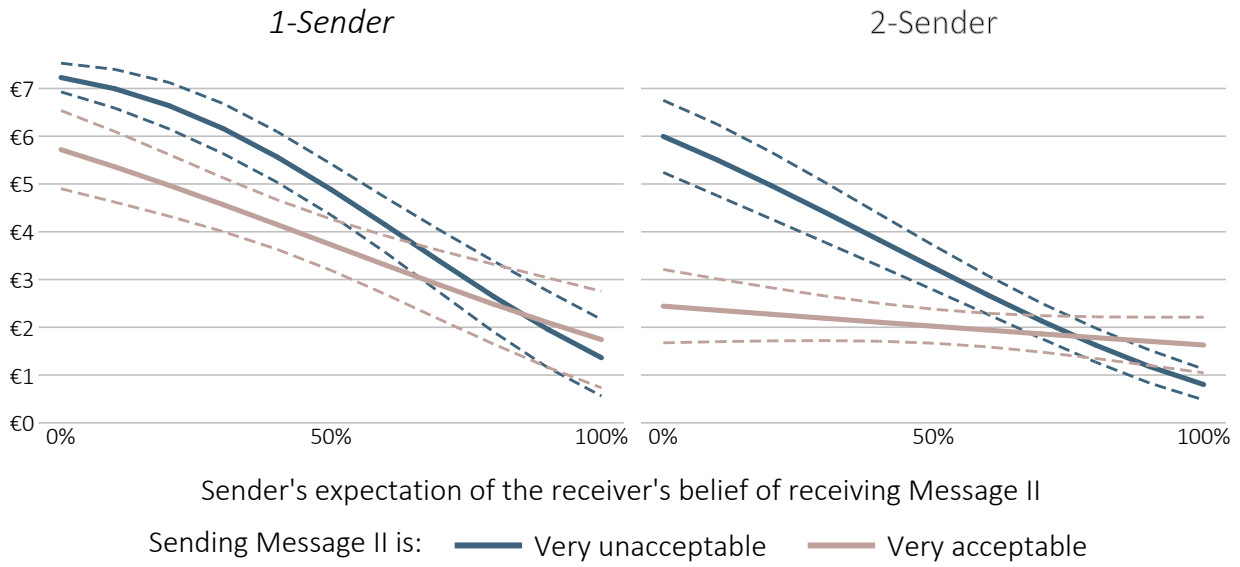


Figure 5. Estimated antisocial premium depending on the sender's normative views and second-order belief

Note: Estimates based on specification III in Table 4. Very unacceptable (acceptable) corresponds to the views of the sender in the 10^{th} (90^{th}) percentile of the normative views distribution. Dotted lines correspond to ± one standard error.

5.6 The importance of active participation

An important question arising from our main findings is: does the increased willingness to act antisocially require sender B's active participation in the decision making or is the presence of a second sender enough to reduce antisocial premiums? To answer this question, we ran additional sessions using a variation of the *2-Sender* game that we refer to as the *Passive-Sender* game.

In the *Passive-Sender* game, sender B is present, delivers the message, and receives the same payoffs as in the *2-Sender* game. The critical difference is that in the *Passive-Sender* game, sender B does not have any say on the content of the message sent to the receiver. The message is picked solely by sender A using the same procedure as in the *1-Sender* game. We ran two sessions of the *Passive-Sender* game in the *Deception* context.[32] A total of 66 subjects participated in these sessions, 22 sender A's, 22 sender B's, and 22 receivers.

As mentioned in Section 4, prominent outcome-based models of social preferences (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002) can predict that adding a second sender lowers antisocial premiums (see footnote 14). This prediction, however, holds irrespective of whether the second sender takes an active part in the decision-making process or

[32]We decided against running sessions of the *Passive-Sender* game in the *Bitter-pill* context because we found that the introduction of a second sender lowered antisocial premiums similarly in both contexts. By concentrating on one context (*Deception*), we increase the power of the comparison between the three games.

not. Hence, with the *Passive-Sender* game, we can determine the extent to which these models explain the observed reduction in antisocial premiums between the *1-Sender* and *2-Sender* games.

Figure 6A depicts the mean antisocial premium in the *Passive-Sender* game as well as in the *Deception* context of the *1-Sender* and *2-Sender* games. Like in section 5.1, we use interval regressions to evaluate whether differences are statistically significant (available in Table A1 of the Appendix). We find that antisocial premiums in the *Passive-Sender* game are close to those in the *1-Sender* game (€3.95 vs. €3.43 on average; $p = 0.558$) and significantly higher than antisocial premiums in the *2-Sender* game (€3.95 vs. €2.10 on average; $p = 0.004$). Therefore, we conclude that active participation is necessary for increased antisocial behavior by groups.

We observe a similar pattern when we compare normative views across the three games. Figure 6B depicts the senders' mean acceptability ratings of sending Message II and their belief of the receivers' acceptability ratings for senders in the *1-Sender*, *2-Sender*, and *Passive-Sender* games (for both senders A and B). By and large, we see that the senders' acceptability ratings of the active and passive senders in the *Passive-Sender* game are similar to those of senders in the *1-Sender* game and below those of senders in the *2-Sender* game.[33] In the Appendix, we further demonstrate that the emotions and second-order beliefs in the *Passive-Sender* game follow the same pattern as in the *1-Sender* game and display similar differences vis-à-vis the *2-Sender* game. Specifically, we find higher levels of experienced guilt if the receiver follows Message II in the *Passive-Sender* game compared to the *2-Sender* game, and we do not see noticeable differences in second-order beliefs (for details see Tables A1, A2, A3, A6, A7, A8 and A12 of the Appendix). We summarize these findings as our last result.

Result 5 *The addition of a passive second sender does not lower antisocial premiums compared to the game with one sender. Consistent with this result, a passive second sender does not make sending Message II more normatively acceptable.*

[33] Like in section 5.2, we use ordered probit regressions to evaluate whether differences are statistically significant (available in Table A3 of the Appendix). There are no statistical differences between senders in the *1-Sender* game and senders A in the *Passive-Sender* game (for acceptability ratings, 2.55 vs. 2.57, $p = 0.979$; for belief of the receivers' acceptability ratings 1.50 vs. 1.67, $p = 0.707$), and senders B in the *Passive-Sender* game (for acceptability ratings, 2.55 vs. 2.23, $p = 0.365$; for belief of the receivers' acceptability ratings 1.50 vs. 1.91, $p = 0.707$). By contrast, senders in the *2-Sender* game tend to view sending Message II to be more acceptable than senders A in the *Passive-Sender* game (for acceptability ratings, 3.14 vs. 2.57, $p = 0.117$; for belief of the receivers' acceptability ratings 2.42 vs. 1.67, $p = 0.055$), and senders B in the *Passive-Sender* game (for acceptability ratings, 3.14 vs. 2.23, $p = 0.009$; for belief of the receivers' acceptability ratings 2.42 vs. 1.91, $p = 0.227$). Similarly, the acceptability ratings of receivers in the *Passive-Sender* game are similar to those of receivers in the *1-Sender* game (2.73 vs. 2.90, $p = 0.337$) but not for senders in the *2-Sender* game (2.73 vs. 3.53, $p = 0.048$).

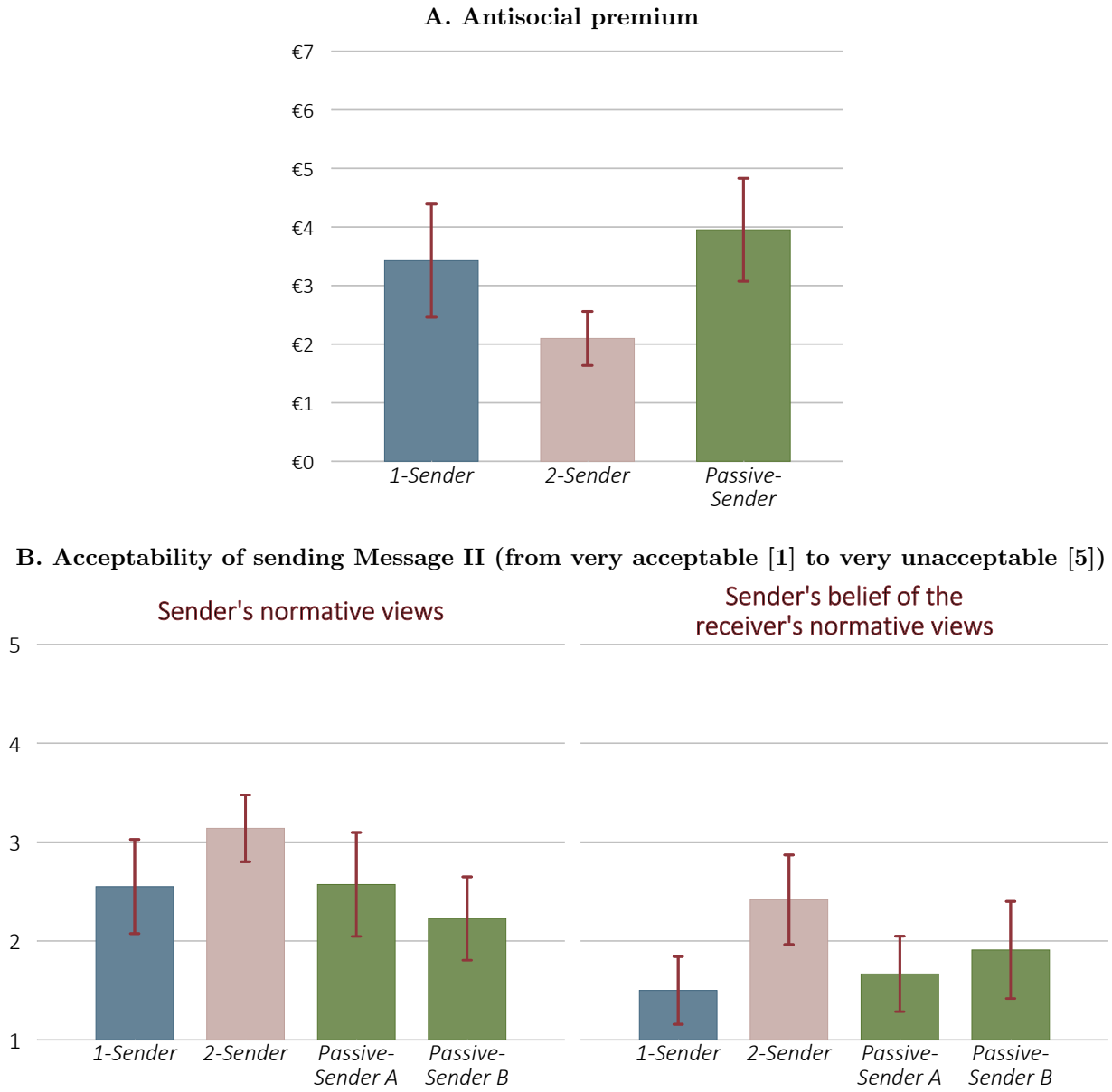


Figure 6. Subjects' mean antisocial premiums and normative views including those of sender A and sender B from the *Passive-Sender* game

Note: The *1-Sender* and *2-Sender* games include subjects from only the *Deception* context. Error bars correspond to 90% confidence intervals.

6 Conclusion

In this study, we first present evidence that individuals tend to behave more antisocially when making a joint decision with a partner than when acting alone. In our experimental design, joint decisions are made without interacting. The absence of interaction enables us to eliminate explanations such as peer effects and information exchange through the deliberation process, while the

lack of a market setting rules out explanations such as social information revealed through the bids and asks of others and the introduction of a more materialistic framing.

We attribute the increased willingness to behave antisocially to a shift in normative beliefs. We observe this shift in multiple ways and different contexts. We find that senders evaluate sending the antisocial message as being more normatively acceptable in the presence of a second sender. We see this difference with both self-reported and incentivized measures of the senders' normative views. Moreover, we see it with two different types of antisocial messages: truthful and deceptive. Finally, we find that the senders' emotional reaction is also consistent with a shift in normative beliefs as they experience less guilt when the antisocial action is a joint decision.

Some of our results give us further insights concerning the senders' shift in normative beliefs. First, we observe a similar shift in the normative views of receivers. This fact implies that the shift is not a self-serving reaction by senders. In other words, senders are not using the presence of the second sender as an "excuse" to misbehave. Second, the results from our *Passive-Sender* game, in which the second sender does not take on an active role, suggests that a necessary condition for the shift in normative views to occur is the active involvement of the partner in the decision-making process. Our results indicate that normative beliefs are clearly important to understand why groups might behave more antisocially than individuals. Although the results of the *Passive-Sender* game allow us to rule out outcome-based models of social preferences as an explanation, we would like to point out that finding evidence of a shift in normative beliefs does not necessarily refute alternative models of this phenomenon. For instance, models of the diffusion of responsibility assume that the motivation of individuals to act prosocially depends on how pivotal their choice is (Engl, 2017; Rothenhäusler et al., 2018). It might very well be that senders think it is more acceptable to send the antisocial message in the *2-Sender* game precisely because their decision counts only if the other sender agrees. In other words, one interpretation of our findings is that they are complementary to existing models and provide insight into the mechanisms through which other-regarding preferences operate.

In this regard, our results also point to interesting ways in which models of guilt aversion can be developed further. In our analysis of the individual determinants of antisocial premiums, we find an interaction between the normative views of senders and their belief of the receivers empirical expectations. This interaction is in line with models of guilt aversion if one interprets individuals' guilt sensitivity as being driven by their normative beliefs. Having a theory of the determinants of guilt sensitivity might allow us to understand why it appears that second-order beliefs seem to matter in some situations but not in others (e.g., see Charness and Dufwenberg, 2006; Vanberg, 2008; Reuben et al., 2009; Ellingsen et al., 2010).

References

Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews*, 1:21–61.

Banerjee, R. (2016). On the interpretation of bribery in a laboratory corruption game: moral frames and social norms. *Experimental Economics*, 19(1):240–267.

Barr, A., Lane, T., and Nosenzo, D. (2018). On the social inappropriateness of discrimination. *Journal of Public Economics*, 164:153–164.

Barr, A. and Michailidou, G. (2017). Complicity without connection or communication. *Journal of Economic Behavior & Organization*, 142:1–10.

Bartling, B. and Fischbacher, U. (2012). Shifting the Blame: On Delegation and Responsibility. *The Review of Economic Studies*, 79(1):67–87.

Bartling, B., Weber, R. A., and Yao, L. (2015). Do Markets Erode Social Responsibility? *The Quarterly Journal of Economics*, 130(1):219–266.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in Games. *American Economic Review*, 97(2):170–176.

Baumeister, R. F., Stillwell, A. M., and Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, 115(2):243–267.

Ben-Shakhar, G., Bornstein, G., Hopfensitz, A., and van Winden, F. (2007). Reciprocity and emotions in bargaining using physiological and self-report measures. *Journal of Economic Psychology*, 28(3):314–323.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, New York.

Bicchieri, C., Muldoon, R., and Sontuoso, A. (2018). Social Norms. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, 90(1):166–193.

Bornstein, G. and Yaniv, I. (1998). Individual and group behavior in the ultimatum game: Are groups more "rational" players? *Experimental Economics*, 1(1):101–108.

Bradley, M. M. and Lang, P. J. (2000). Measuring emotion: behavior, feeling and physiology. In Lang, R. D. and Nadel, L., editors, *Cognitive Neuroscience of Emotions*, pages 242–276. Oxford University Press, Oxford.

Cason, T. N. and Mui, V.-L. (1997). A laboratory study of group polarisation in the team dictator game. *The Economic Journal*, 107(444):1465–1483.

Charness, G. and Dufwenberg, M. (2006). Promises and Partnership. *Econometrica*, 74(6):1579–1601.

Charness, G. and Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Choo, L., Grimm, V., Horvath, G., and Nitta, K. (2016). *Whistleblowing and Diffusion of Responsibility:*

An Experimental Investigation. PhD thesis, German Economic Association Annual Conference 2016, no. 145781.

Cialdini, R. B. (2003). Crafting Normative Messages to Protect the Environment. *Current Directions in Psychological Science*, 12(4):105–109.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015–1026.

Coffman, L. C. (2011). Intermediation Reduces Punishment (and Reward). *American Economic Journal: Microeconomics*, 3(4):77–106.

Cohen, T. R., Gunia, B. C., Kim-Jun, S. Y., and Murnighan, J. K. (2009). Do groups lie more than individuals? Honesty and deception as a function of strategic self-interest. *Journal of Experimental Social Psychology*, 45(6):1321–1324.

Conrads, J., Irlenbusch, B., Rilke, R. M., and Walkowitz, G. (2013). Lying and team incentives. *Journal of Economic Psychology*, 34:1–7.

Cox, J. C. (2002). Trust, Reciprocity, and Other-Regarding Preferences: Groups Vs. Individuals and Males Vs. Females. In Zwick, R. and Rapoport, A., editors, *Experimental Business Research*, pages 331–350. Springer US, Boston, MA.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Danilov, A., Biemann, T., Kring, T., and Sliwka, D. (2013). The dark side of team incentives: Experimental evidence on advice quality from financial service professionals. *Journal of Economic Behavior & Organization*, 93:266–272.

Deckers, T., Falk, A., Kosse, F., and Szech, N. (2016). Homo Moralis: Personal Characteristics, Institutions, and Moral Decision-Making. CESifo Working Paper 5800.

Dimant, E., Bicchieri, C., and Xiao, E. (2018). *Deviant or Wrong? The Effects of Norm Information on the Efficacy of Punishment*. PhD thesis, SSRN Working Paper 3294371.

Ellingsen, T., Johannesson, M., Tjøtta, S., and Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, 68(1):95–107.

Elster, J. (1989). *The Cement of Society - A Study of Social Order*. Cambridge University Press, Cambridge.

Elster, J. (2009). Norms. In Bearman, P. and Hedström, P., editors, *The Oxford Handbook of Analytical Sociology*, chapter 9, pages 195–217. Oxford University Press, Oxford, UK.

Engl, F. (2017). A Theory of Causal Responsibility Attribution. SSRN Working paper 2932769.

Erkut, H. and Reuben, E. (2019). Preference measurement and manipulation in experimental economics. In Schram, A. and Ule, A., editors, *Handbook of Research Methods and Applications in Experimental Economics*. Edward Elgar Publishing, Surrey, UK.

Falk, A. and Szech, N. (2013). Morals and Markets. *Science*, 340(6133):707–711.

Falk, A. and Szech, N. (2016). Diffusion of Being Pivotal and Immoral Outcomes. Human Capital and Economic Opportunity Global Working Group, Working Paper 2016-013.

Fehr, E. and Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Gächter, S., Gerhards, L., and Nosenzo, D. (2017). The importance of peers for compliance with norms of fair sharing. *European Economic Review*, 97:72–86.

Gächter, S., Nosenzo, D., and Sefton, M. (2013). Peer effects in pro-social behavior: Social norms or social preferences? *Journal of the European Economic Association*, 11(3):548–573.

Garofalo, O. and Rott, C. (2017). Shifting Blame? Experimental Evidence of Delegating Communication. *Management Science*.

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

Gino, F., Ayal, S., and Ariely, D. (2013). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of Economic Behavior & Organization*, 93:285–292.

Gino, F. and Pierce, L. (2010). Lying to Level the Playing Field: Why People May Dishonestly Help or Hurt Others to Create Equity. *Journal of Business Ethics*, 95(S1):89–103.

Gneezy, U. (2005). Deception: The Role of Consequences. *American Economic Review*, 95(1):384–394.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.

Hamman, J. R., Loewenstein, G., and Weber, R. A. (2010). Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship. *American Economic Review*, 100(4):1826–1846.

Hopfensitz, A. and Reuben, E. (2009). The Importance of Emotions for the Effectiveness of Social Punishment. *The Economic Journal*, 119(540):1534–1559.

Keck, S. (2014). Group reactions to dishonesty. *Organizational Behavior and Human Decision Processes*, 124(1):1–10.

Kimbrough, E. O. and Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3):608–638.

Kirchler, M., Huber, J., Stefan, M., and Sutter, M. (2016). Market Design and Moral Behavior. *Management Science*, 62(9):2615–2625.

Kocher, M. G., Schudy, S., and Spantig, L. (2018). I Lie? We Lie! Why? Experimental Evidence on a Dishonesty Shift in Groups. *Management Science*, 64(9):3995–4008.

Kocher, M. G. and Sutter, M. (2007). Individual versus group behavior and the role of the decision making procedure in gift-exchange experiments. *Empirica*, 34(1):63–88.

Korbel, V. (2017). Do we lie in groups? An experimental evidence. *Applied Economics Letters*, 24(15):1107–1111.

Krupka, E. L., Leider, S., and Jiang, M. (2017). A Meeting of the Minds: Informal Agreements and Social Norms. *Management Science*, 63(6):1708–1729.

Krupka, E. L. and Weber, R. A. (2013). Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? *Journal of the European Economic Association*, 11(3):495–524.

López-Pérez, R. (2010). Guilt and shame: An axiomatic analysis. *Theory and Decision*, 69(4):569–586.

Luhan, W. J., Kocher, M. G., and Sutter, M. (2009). Group polarization in the team dictator game reconsidered. *Experimental Economics*, 12(1):26–41.

Muehlheusser, G., Roider, A., and Wallmeier, N. (2015). Gender differences in honesty: Groups versus individuals. *Economics Letters*, 128:25–29.

Nielsen, K., Bhattacharya, P., Kagel, J. H., and Sengupta, A. (2017). Teams Promise But Do Not Deliver. SSRN Working Paper 2998300.

Oexl, R. and Grossman, Z. J. (2013). Shifting the blame to a powerless intermediary. *Experimental Economics*, 16(3):306–312.

Reuben, E. and Riedl, A. (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior*, 77(1):122–137.

Reuben, E., Sapienza, P., and Zingales, L. (2009). Is mistrust self-fulfilling? *Economics Letters*, 104(2):89–91.

Reuben, E. and van Winden, F. (2010). Fairness perceptions and prosocial emotions in the power to take. *Journal of Economic Psychology*, 31(6):908–922.

Rothenhäusler, D., Schweizer, N., and Szech, N. (2018). Guilt in voting and public good games. *European Economic Review*, 101:664–681.

Schram, A. and Charness, G. (2015). Inducing Social Norms in Laboratory Allocation Choices. *Management Science*, 61(7):1531–1546.

Stewart, M. B. (1983). On Least Squares Estimation when the Dependent Variable is Grouped. *The Review of Economic Studies*, 50(4):737.

Sutter, M. (2009). Deception Through Telling the Truth?! Experimental Evidence From Individuals and Teams. *The Economic Journal*, 119(534):47–60.

Vanberg, C. (2008). Why Do People Keep Their Promises? An Experimental Test of Two Explanations. *Econometrica*, 76(6):1467–1480.

Weisel, O. and Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences*, 112(34):10651–10656.

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, 115(2):157–168.

Appendix A Supplementary data analysis

Appendix A contains the regressions presented in the paper, including the coefficients of the control variables, and additional statistical analysis reported in the paper but not fully described there due to space constraints. In section A.1, we present the regressions used to test for treatment differences. In section A.2, we present the complete statistical analysis of the subjects' emotional response. Finally, in section A.3, we include the regressions used to evaluate the determinants of the antisocial premium.

A.1 Treatment comparisons

Table A1 shows the regressions used to evaluate whether the treatment differences in the senders' antisocial premiums are statistically significant. The coefficients are estimated using interval regressions. We classified senders who choose Message II over Message I when they earn at least €x as having an antisocial premium in the interval [€$x-0.5$, €x]. For senders who always choose Message I, we classified them as having an antisocial premium in the interval [€$7.50, \infty$) if they played in the *1-Sender* game or as sender A in the *2-Sender* game, or as having an antisocial premium in the interval [€$7.00, \infty$) if they played as sender B the *2-Sender* game. For senders who always choose Message II, we classified them as having an antisocial premium in the interval ($-\infty$, €0.50] if they played in the *1-Sender* game or as sender A in the *2-Sender* game, or in the interval ($-\infty$, €0.00] if they played as sender B in the *2-Sender* game. All regressions are estimated using robust standard errors. The regressions in columns I and II use data from to the *1-Sender* and the *2-Sender* games. Column III further includes the data from the *Passive-Sender* game.

Table A2 shows the means and standard deviations of selected variables used in the analyses. The first three columns contain the respective values for the *1-Sender* and *2-Sender* games with pooled data from the *Bitter-pill* and *Deception* contexts as well as the values for the *Passive-Sender* game. Columns three to six report these values in the *1-Sender* and *2-Sender* games separately for each context. Note that senders' normative views of sending Message I are elicited on a five-point Likert scale ranging from very unacceptable (1) to very acceptable (5). Moreover, the senders' additional earnings if the receiver follows Message II refers to the surplus that senders gain in the selected row compared to the equal payoff distribution (assuming that Message II is sent and that the receiver implements the option mentioned in this message). Specifically, this amount equals $7-x$ for the sender in the *1-Sender* game as well as sender A in the *2-Sender* and *Passive-Sender* games and to x for sender B in the *2-Sender* and *Passive-Sender* games.

Table A1. Treatment differences in antisocial premiums

Note: Interval regressions of the senders' antisocial premium. Robust standard errors in parentheses. ***, **, and * indicate statistical significance at 0.01, 0.05, and 0.10 using two-tailed tests.

	I	II	III
2-Sender	−1.58***		
	(0.57)		
2-Sender × *Bitter-pill*		−1.57***	−1.57***
		(0.74)	(0.73)
2-Sender × *Deception*		−1.57***	−1.59***
		(0.78)	(0.79)
Passive-Sender			0.56
			(0.95)
Deception		−1.63*	−1.63*
		(0.91)	(0.93)
Constant	3.99***	4.82***	4.82***
	(0.49)	(0.60)	(0.61)
Observations	110	110	131
χ^2	7.73	24.32	25.92

Table A2. Means and standard deviations of selected variables by game and treatment

	Overall			***Bitter-pill***		***Deception***	
	1-Sender	*2-Sender*	*Passive -sender*	*1-Sender*	*2-Sender*	*1-Sender*	*2-Sender*
Fraction of Message IIs sent	0.44	0.29	0.46	0.38	0.11	0.50	0.47
	(0.50)	(0.46)	(0.50)	(0.49)	(0.31)	(0.51)	(0.50)
Senders' normative views of	4.51	4.31	4.37	4.58	4.20	4.45	4.42
sending Message I	(0.85)	(1.04)	(1.00)	(0.69)	(1.21)	(1.00)	(0.84)
Senders' additional earnings if	3.31	3.49	3.57	3.66	3.51	2.98	3.47
Message II is followed	(2.07)	(1.92)	(2.05)	(2.17)	(1.85)	(1.97)	(2.01)
Fraction of receivers who	0.89	0.97	1.00	0.77	0.95	1.00	1.00
followed Message II	(0.32)	(0.16)	(0.00)	(0.43)	(0.23)	(0.00)	(0.00)
Senders' expected fraction of	0.84	0.83	0.85	0.84	0.81	0.85	0.85
followed Message IIs	(0.21)	(0.23)	(0.25)	(0.24)	(0.26)	(0.20)	(0.21)
Fraction of women	0.49	0.41	0.46	0.46	0.36	0.53	0.45
	(0.50)	(0.49)	(0.50)	(0.51)	(0.49)	(0.51)	(0.50)
Age	23.63	22.74	21.85	22.87	22.80	24.38	22.67
	(5.61)	(2.46)	(4.52)	(3.99)	(2.44)	(6.80)	(2.51)

Table A3. Treatment differences in normative views

Note: Ordered probit regressions of the subjects' normative views. Robust standard errors clustered on matched pairs (for senders in *2-Sender*) in parentheses. ***, **, and * indicate statistical significance at 0.01, 0.05, and 0.10 using two-tailed tests.

	Senders' normative views			Senders' expected normative views			Receivers' normative views		
	I	II	III	IV	V	VI	VII	VIII	IX
2-Sender	0.62***			0.65***			0.39		
	(0.23)			(0.24)			(0.24)		
2-Sender × *Bitter-pill*		0.70**	0.66**		0.54	0.55		0.30	0.27
		(0.30)	(0.29)		(0.35)	(0.36)		(0.34)	(0.34)
2-Sender × *Deception*		0.56*	0.53*		0.76**	0.77**		0.53	0.54
		(0.34)	(0.32)		(0.33)	(0.33)		(0.34)	(0.34)
Deception		0.44	0.42		−0.25	−0.25		0.50*	0.49
		(0.36)	(0.34)		(0.37)	(0.38)		(0.30)	(0.30)
Sender A in			−0.01			0.20			
Passive-Sender			(0.37)			(0.34)			
Sender B in			−0.31			0.38			
Passive-Sender			(0.34)			(0.36)			
Receiver in									0.12
Passive-Sender									(0.32)
Observations	110	110	153	110	110	153	79	79	101
Clusters	77	77	99	77	77	99	79	79	101
χ^2	6.94	9.08	12.16	7.23	8.17	11.16	2.67	11.76	11.40

Table A3 shows the regressions used to evaluate whether the treatment differences in the subjects' normative views are statistically significant. All normative views range from very unacceptable (1) to very acceptable (5). Therefore, we estimate all coefficients using ordered probit regressions with robust standard errors. Moreover, we cluster standard errors of senders in the *2-Sender* game on their matched pairs. In columns I to III, the dependent variable is the senders' normative views regarding the acceptability of sending Message II. In columns IV to VI, the dependent variable is the senders' belief of the receivers' normative views regarding the acceptability of sending Message II. Lastly, in columns VII to IX, the dependent variable is the receivers' normative views regarding the acceptability of sending Message II. All regressions include data from the *1-Sender* and *2-Sender* games. Regressions in columns III, VI, and IX further include the data from the *Passive-Sender* game.

Table A4 provides robustness checks for the results from the *1-Sender* and *2-Sender* games

Table A4. Treatment differences in normative views, robustness checks

Note: Ordered probit regressions of the subjects' normative views. Robust standard errors clustered on matched pairs (for senders in *2-Sender*) in parentheses. ***, **, and * indicate statistical significance at 0.01, 0.05, and 0.10 using two-tailed tests.

	I	II	III	IV	V	VI
2-Sender	0.63**		0.58*		−0.26	
	(0.27)		(0.32)		(0.24)	
2-Sender × *Bitter-pill*		0.75**		0.48		−0.40
		(0.33)		(0.40)		(0.33)
2-Sender × *Deception*		0.59		0.68*		−0.11
		(0.37)		(0.39)		(0.35)
Deception		0.50		−0.20		−0.10
		(0.37)		(0.38)		(0.39)
Other sender chose Message II	−0.01	−0.05	0.11	0.10		
	(0.28)	(0.29)	(0.30)	(0.30)		
Message II was sent	0.29	0.40	0.24	0.19		
	(0.59)	(0.62)	(0.58)	(0.60)		
Earnings if message is followed	0.04	0.08	0.06	0.06		
	(0.09)	(0.09)	(0.08)	(0.09)		
Observations	110	110	110	110	110	110
Clusters	77	77	77	77	77	77
χ^2	8.18	12.16	9.75	10.69	1.12	1.68

observed in Table A3. We use ordered probit regressions with robust standard errors clustered on their matched pairs in the *2-Sender* game. In columns I and II, the dependent variable is the senders' normative views concerning the acceptability of sending Message II. Unlike in Table A3, we also include the following control variables: a dummy variable that equals one if the other sender in the *2-Sender* game chose Message II, a dummy variable that equals one if the message sent to the receiver was Message II, and the sender's earnings if the receiver follows the message. Note that the control variables are neither jointly significant in column I ($p = 0.968$) nor column II ($p = 0.844$). In columns III and IV, the dependent variable is the senders' belief of the receivers' normative views concerning the acceptability of sending Message II. These regressions also include the abovementioned control variables. Once again, note that the control variables are neither jointly significant in column III ($p = 0.771$) nor column IV ($p = 0.837$). Unlike in Table A3, in columns V and VI, the dependent variable is the senders' normative views concerning the acceptability of sending Message I.

Table A5 shows the regressions used to evaluate whether the treatment differences in the sub-

Table A5. Treatment differences in senders' second-order beliefs and the receivers' own beliefs regarding the probability of receiving Message II

Note: Tobit regressions of the senders' belief of the receivers' expected probability of receiving Message II and the receivers' expected probability of receiving Message II. Robust standard errors clustered on matched pairs (in *2-Sender* and *Passive-Sender*) in parentheses. ***, **, and * indicate statistical significance at 0.01, 0.05, and 0.10 using two-tailed tests.

	I	II	III	IV	V	VI
2-Sender	−0.03			−0.07		
	(0.08)			(0.07)		
2-Sender × *Bitter-pill*		−0.05	−0.05		−0.27***	−0.40***
		(0.10)	(0.10)		(0.10)	(0.10)
2-Sender × *Deception*		0.00	0.00		0.13	0.13
		(0.12)	(0.12)		(0.09)	(0.09)
Deception		0.11	0.11		−0.03	−0.03
		(0.13)	(0.13)		(0.10)	(0.10)
Sender A in *Passive-Sender*			0.00			
			(0.13)			
Sender B in *Passive-Sender*			0.04			
			(0.12)			
Receiver in *Passive-Sender*						0.01
						(0.11)
Constant	0.59***	0.54***	0.54***	0.56***	0.58***	0.58***
	(0.07)	(0.09)	(0.09)	(0.05)	(0.07)	(0.07)
Observations	110	110	153	79	79	101
Clusters	77	77	99	79	79	101
F-statistic	0.11	1.51	1.73	0.96	6.31	4.73

jects' beliefs about the probability that a receiver receives Message II (referred to as the senders' "second-order beliefs") are statistically significant. Since beliefs are censored at 0% and 100%, we estimate all coefficients using Tobit regressions with robust standard errors clustered on matched pairs in the *2-Sender* game. In columns I to III, the dependent variable is the senders' belief of the receivers' expected probability of receiving Message II. In columns IV and VI, the dependent variable is the receivers' expected probability of receiving Message II. All regressions include data from to the *1-Sender* and the *2-Sender* games. Regressions in columns III and VI further include the data from the *Passive-Sender* game.

Table A6. Means and standard deviations of the senders' self-reported emotions

	1-Sender		*2-Sender*		*Passive-Sender*	
	Message I	Message II	Message I	Message II	Message I	Message II
Guilt	1.09	4.23	1.62	2.68	1.13	3.95
	(0.29)	(2.42)	(1.23)	(1.49)	(0.63)	(2.11)
Shame	1.27	3.85	1.46	2.16	1.22	2.30
	(0.77)	(2.76)	(0.93)	(1.80)	(0.52)	(1.72)
Anger	1.14	1.85	2.22	1.32	1.39	1.50
	(0.35)	(1.57)	(1.73)	(0.95)	(1.03)	(0.95)
Happiness	5.82	5.69	4.62	6.11	5.57	5.10
	(1.26)	(1.55)	(1.69)	(0.99)	(1.12)	(1.37)
Gratitude	5.77	5.38	4.60	5.79	5.30	3.90
	(1.48)	(1.71)	(1.80)	(1.08)	(1.89)	(1.83)

A.2 Emotions regressions

Table A6 provides the means and standard deviations of the senders' self-reported emotions dependent on whether Message I or Message II was sent to the receiver. Senders' emotions refer to the moment they learned the outcomes of all players and were elicited on a seven-point Likert scale ranging from 1 to 7 after the game was played.

Table A7 shows the regressions used to evaluate determinants of the senders' guilt. Guilt was elicited on a seven-point Likert scale ranging from 1 to 7 after the game was played. We use linear regressions with robust standard errors clustered on matched pairs, and the senders' experienced guilt as the dependent variable. Regressions I to III use data from the *1-Sender* game. In regression IV, we pooled the data from the *Passive-Sender* and *1-Sender* games (as opposed to the *2-Sender* game) since these contexts display similar behavior. In regressions V to VII, we use data from the *2-Sender* game. 'Delivered Message I' and 'Delivered Message II' are dummy variables indicating the message that was delivered to and followed by the receiver; 'Normative views' are the senders' normative views of sending Message II; 'Second-order belief' is the senders' belief of the receivers' expected probability of receiving Message II. In regressions III, IV, and VII, we add interaction effects of the message sent and, respectively, normative views and second-order beliefs. Regressions II-IV and VI-VII include further controls: Message II being dishonest ('Deception'), the antisocial premium, the interaction of the message sent with the senders' earnings from Message II, gender, and age. In regression IV, we also include an indicator variable for being sender A in *Passive-Sender*. Regressions VI and VII include an indicator variable for being sender B in *2-Sender*.

Table A7. Determinants of experienced guilt

Note: OLS regressions of the senders' guilt. Robust standard errors clustered on matched pairs (for senders in *2-Sender*) in parentheses. ***, **, and * indicate statistical significance at 0.01, 0.05, and 0.10 using two-tailed tests.

	***1-Sender* game**				***2-Sender* game**		
	I	II	III	IV	V	VI	VII
Delivered Message II	3.14***	3.38***	6.30***	5.00***	1.06**	0.89**	2.76*
	(0.67)	(0.75)	(1.86)	(2.24)	(0.29)	(0.38)	(1.50)
Delivered Message I × Normative			0.00	−0.20			0.00
views			(0.15)	(0.14)			(0.19)
Delivered Message II × Normative			−0.91***	−0.49***			−0.38*
views			(0.29)	(0.38)			(0.22)
Delivered Message I × Second-order			0.18	−0.42			1.13
belief			(0.33)	(0.44)			(0.90)
Delivered Message II × Second-order			−0.65	−1.66			−0.25
belief			(2.79)	(2.52)			(1.39)
Deception		−0.04	−0.19	0.42		0.13	0.26
		(0.37)	(0.37)	(0.46)		(0.40)	(0.42)
Antisocial premium		−0.03	−0.06	−0.03		−0.05	0.05
		(0.12)	(0.11)	(0.13)		(0.06)	(0.12)
Delivered Message I × Sender's		−0.22	−0.05	−0.26		−0.27	−0.24
earnings if Message II is applied		(0.17)	(0.16)	(0.20)		(0.17)	(0.16)
Delivered Message II × Sender's		−1.48***	−1.37***	−0.54		0.67	0.80
earnings if Message II is applied		(0.53)	(0.52)	(0.51)		(0.67)	(0.65)
Female		−0.25	−0.07	−0.06		0.06	0.12
		(0.22)	(0.21)	(0.18)		(0.14)	(0.14)
Age		0.69	0.15	0.44		0.03	0.06
		(0.46)	(0.34)	(0.44)		(0.03)	(0.06)
Age^2		−0.17	−0.04	−0.18		−0.08	−0.12
		(0.11)	(0.08)	(0.13)		(0.08)	(0.10)
Sender A in *Passive-Sender*				−0.18			
				(0.13)			
Sender B in *2-Sender*						−0.14	0.10
						(0.32)	(0.43)
Constant	1.09***	1.46*	1.51*	1.82*	1.62***	2.01***	0.78
	(0.06)	(0.74)	(0.89)	(1.03)	(0.18)	(0.60)	(1.60)
Observations	35	35	35	56	69	69	69
Clusters	35	35	35	56	37	37	37
F-statistic	22.14	19.65	17.77	8.01	13.11	3.79	3.09

Table A8. Determinants of other negative emotions

Note: OLS regressions of the senders' negative emotions. Robust standard errors clustered on matched pairs (for senders in *2-Sender*) in parentheses. ***, **, and * indicate statistical significance at 0.01, 0.05, and 0.10 using two-tailed tests.

	I	II	III	IV	V	VI	VII	VIII
Delivered Message II	3.10***	4.72***	0.63	2.41	0.72	2.28	−0.96**	−0.46*
	(0.99)	(2.25)	(0.48)	(1.49)	(0.54)	(1.53)	(0.42)	(1.25)
Delivered Message I ×		0.15		−0.30***		−0.09		0.32
Normative views		(0.23)		(0.11)		(0.13)		(0.25)
Delivered Message II ×		−1.00***		−0.41		0.02		0.19
Normative views		(0.39)		(0.37)		(0.20)		(0.19)
Delivered Message I ×		0.02		0.37		0.19		0.46
Second-order belief		(0.60)		(0.61)		(0.37)		(1.07)
Delivered Message II ×		1.85		−1.93		−2.36		0.37
Second-order belief		(3.08)		(1.54)		(1.88)		(0.91)
Deception	−0.43	−0.63	0.04	0.27	0.20	0.21	−0.91**	−1.02**
	(0.43)	(0.48)	(0.32)	(0.33)	(0.21)	(0.29)	(0.43)	(0.44)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	35	35	69	69	35	35	69	69
Clusters	35	35	37	37	35	35	37	37
F-statistic	7.31	7.35	1.61	9.26	0.56	0.71	3.26	2.73

In Table A8, we report the regressions used to evaluate potential determinants of the senders' experienced anger and shame. As in the case of guilt, these negative emotions refer to the moment senders learned the outcomes of all players and were elicited on a seven-point Likert scale ranging from 1 to 7 after the game was played. Like in Table A7, we use linear regressions with robust standard errors clustered on matched pairs in all models. In regressions I and II (*1-Sender*) as well as regressions III and IV (*2-Sender*), the dependent variable is the senders' experienced anger. In regressions V and VI (*1-Sender*) as well as regressions VII and VIII (*2-Sender*), the dependent variable is the senders' experienced shame. Regressions I, III, V, and VI, include independent variables that indicate which message (Message I or Message II) was sent and whether Message II was dishonest (instead of honest). We added the interaction of normative views and second-order beliefs with the message that was sent to regressions II, IV, VI, and VIII. All regressions include further controls for the antisocial premium, the interaction of the message sent with the senders' earnings from Message II, gender, age and, in case of the *2-Sender* game, for being sender B.

Table A9 shows similar regressions as in Table A8. The only difference is the dependent variables. In regressions I and II (*1-Sender*) as well as in regressions III and IV (*2-Sender*), the

Table A9. Determinants of other positive emotions

Note: OLS regressions of the senders' positive emotions. Robust standard errors clustered on matched pairs (for senders in *2-Sender*) in parentheses. ***, **, and * indicate statistical significance at 0.01, 0.05, and 0.10 using two-tailed tests.

	I	II	III	IV	V	VI	VII	VIII
Delivered Message II	−0.91	−0.94	1.69***	1.14	0.09	−0.17	1.65***	0.79
	(0.96)	(2.28)	(0.42)	(1.50)	(0.89)	(3.48)	(0.52)	(1.82)
Delivered Message I ×		0.20		−0.07		−0.27		−0.12
Normative views		(0.42)		(0.34)		(0.46)		(0.33)
Delivered Message II ×		0.29		−0.34*		0.17		−0.10
Normative views		(0.19)		(0.18)		(0.35)		(0.29)
Delivered Message I ×		0.58		0.75		−0.79		1.34
Second-order belief		(1.06)		(1.11)		(1.24)		(0.95)
Delivered Message II ×		0.11		2.24*		−1.59		1.87
Second-order belief		(1.83)		(1.21)		(3.39)		(1.50)
Deception	0.29	0.18	0.19	0.24	0.85	1.06	−0.27	−0.19
	(0.48)	(0.48)	(0.50)	(0.56)	(0.60)	(0.71)	(0.52)	(0.56)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	35	35	69	69	35	35	69	69
Clusters	35	35	37	37	35	35	37	37
F-statistic	2.29	2.15	3.96	4.06	2.45	1.72	3.73	5.03

dependent variable is the senders' experienced happiness. In regressions V and VI (*1-Sender*) as well as regressions VII and VIII (*2-Sender*), the dependent variable is the senders' experienced gratitude. As with other emotions, these positive emotions refer to the moment senders learned the outcomes of all players and were elicited on a seven-point Likert scale ranging from 1 to 7 after the game was played.

Table A10 shows the linear regressions used to evaluate whether the treatment differences in the senders' experienced guilt after sending an antisocial message that is followed by the receiver, are statistically significant between the *1-Sender* game and the *2-Sender* game in regression I and between the *1-Sender* game and the *Passive-Sender* game in regression II.

A.3 Antisocial premium regressions

Table A11 shows the regressions used to evaluate the effect of senders' normative views and second-order beliefs about the receivers' expected probability of receiving Message II on the antisocial premium, which is the dependent variable in all six regressions. We use interval regressions to

Table A10. Treatment differences in experienced guilt

Note: OLS regressions of the senders' guilt. Robust standard errors clustered on matched pairs (for senders in *2-Sender*) in parentheses. ***, **, and * indicate statistical significance at 0.01, 0.05, and 0.10 using two-tailed tests.

	I	II
2-Sender	0.53***	
	(0.19)	
Passive-Sender		0.40
		(0.19)
1-Sender × Message II	3.14***	3.14***
	(0.66)	(0.66)
2-Sender × Message II	1.06***	
	(0.29)	
Passive-Sender × Message II		2.82***
		(0.49)
Constant	1.09***	1.09***
	(0.606)	(0.06)
Observations	104	78
Clusters	72	57
F-statistic	23.00	18.83

account for the fact that when we observe a sender who switches from Message I to Message II when the latter pays more than €x, her antisocial premium lies in the interval [€$x - 0.50$, €x]. All regressions are estimated using standard errors clustered on matched pairs.

The regressions in columns I to III use data from the *1Sender* game, the regression in column IV also includes data from the *Passive-Sender* game, and the regressions in columns V to VII use data from the *2-Sender* game. Regressions I and V include senders' normative views, second-order beliefs and an indicator variable for the context in which the game is played. In regressions II and VI, we add the interaction of senders' normative views and second-order belief. In regressions III, IV, and VII, we further control for senders' expected probability that the receiver follows the message, gender, and age. In regression IV, we include an indicator variable for being sender A in the *Passive-Sender* game, and in regression VII, we include an indicator variable for being sender B in the *2-Sender* game. Note that we decided to pool the data from the *Passive-Sender* game with the *1-Sender* game (as opposed to the *2-Sender* game) because these contexts display similar behavior.

Table A12 reports the coefficients of the same regressions used to evaluate the effect of senders'

Table A11. Determinants of the antisocial premium

Note: Interval regressions of the senders' antisocial premium. Robust standard errors in parentheses. ***, **, and * indicate statistical significance at 0.01, 0.05, and 0.10 using two-tailed tests.

	I	II	III	IV	V	VI	VII
Normative views	−0.39	−0.98*	−1.03*	−0.73	−0.20	−1.06***	−1.00**
	(0.33)	(0.50)	(0.55)	(0.50)	(0.25)	(0.39)	(0.39)
Second-order belief	−6.32***	−9.15***	−9.39***	−8.44***	−3.01***	−7.25***	−7.50***
	(1.48)	(2.68)	(3.02)	(2.66)	(0.91)	(1.80)	(1.83)
Normative views ×		1.13	1.20	1.08		1.44***	1.31**
Second-order belief		(0.93)	(1.01)	(0.73)		(0.53)	(0.53)
Deception	−1.00	−0.94	−1.04	−1.07	−1.08**	−0.89**	−0.68*
	(0.76)	(0.76)	(0.77)	(0.80)	(0.44)	(0.39)	(0.36)
Expected probability of			−0.15	−0.48			−0.11
receiver following message			(0.32)	(0.31)			(0.20)
Female			0.36	0.12			0.06
			(0.38)	(0.36)			(0.23)
Age			0.38	−0.01			0.29
			(0.67)	(0.55)			(0.31)
Age^2			−0.12	−0.02			−0.18
			(0.18)	(0.17)			(0.11)
Sender A in *Passive-Sender*				0.50			
				(0.86)			
Sender B in *2-Sender*							−1.55***
							(0.51)
Constant	9.02***	10.48***	10.76***	9.59***	5.26***	7.66***	10.29***
	(1.08)	(1.47)	(1.77)	(1.67)	(0.97)	(1.38)	(1.62)
Observations	39	39	39	60	71	71	71
χ^2	25.49	30.42	40.57	39.95	23.61	34.20	64.74

normative views and second-order beliefs on the antisocial premium with two exceptions. First, in all seven regressions, we use the senders' second-order normative views instead of their normative views, that is, the senders' estimation of how the receiver rated the acceptability of sending Message II on a 5-point Likert scale ranging from very unacceptable (1) to very acceptable (5). Second, in regressions II, III, IV, VI, and VII, we include the interaction of senders' second-order beliefs with their second-order normative views instead of with their normative views.

Table A12. Determinants of the antisocial premium, alternative specifications

Note: Interval regressions of the senders' antisocial premium. Robust standard errors in parentheses. ***, **, and * indicate statistical significance at 0.01, 0.05, and 0.10 using two-tailed tests.

	I	II	III	IV	V	VI	VII
Expected normative views	−0.10	0.95	0.91	−0.02	−0.05	−0.95***	−0.88***
	(0.28)	(0.88)	(1.04)	(0.86)	(0.16)	(0.28)	(0.31)
Second-order belief	−6.36***	−3.40	−3.26	−4.86*	−3.11***	−7.29***	−7.55***
	(1.59)	(3.30)	(2.88)	(2.78)	(0.95)	(1.37)	(1.54)
Exp. normative views ×		−2.24	−2.34	−0.45		1.68***	1.53***
Second-order belief		(1.84)	(1.98)	(1.45)		(0.39)	(0.43)
Deception	−1.16	−1.23*	−1.39*	−1.29*	−1.15**	−1.14***	−0.93***
	(0.76)	(0.74)	(0.74)	(0.76)	(0.45)	(0.42)	(0.36)
Expected probability of			−0.22	−0.51			−0.07
receiver following message			(0.45)	(0.34)			(0.18)
Female			0.38	0.12			0.13
			(0.37)	(0.34)			(0.19)
Age			0.27	−0.13			0.24
			(0.74)	(0.57)			(0.25)
Age^2			−0.07	0.02			−0.21*
			(0.22)	(0.18)			(0.11)
Sender A in *Passive-Sender*				0.81			
				(0.85)			
Sender B in *2-Sender*							−1.46***
							(0.46)
Constant	8.37***	6.94***	7.14***	7.80***	4.87***	7.23***	9.72***
	(1.08)	(1.73)	(1.77)	(1.67)	(0.82)	(1.03)	(1.39)
Observations	39	39	39	60	71	71	71
χ^2	23.42	37.59	26.95	41.33	27.15	48.61	63.72

Appendix B Instructions

Appendix B contains a sample of the instructions and screenshots used in the experiment. Specifically, we provide the instructions and some screenshots from the *2-Sender* game in the *Deception* context. The instructions used in the *Bitter-pill* and *Passive-Sender* games are almost identical and are available from the authors upon request.

General instructions

You are participating in a study on economic decision-making. You have already earned *€5* for showing up on time. Please read these instructions carefully as they describe how you can earn *additional* money. You will be paid all your earnings in cash.

Please do not talk or communicate with other participants in any way. If you have questions, raise your hand and one of us will help you.

In the study, all participants are randomly assigned to groups of three. Within each group, the computer randomly assigns participants to the roles of *Player 1*, *Player 2*, and *Player 3*. You will be informed of your role on the computer screen.

Summary of the study

- There are ten options with payments for each player. Player 1 and Player 2 are informed of the payment each player receives in each option. On the other hand, Player 3 does not receive this information.
- *Player 1 chooses one message out of the two available messages* to be sent to Player 3. Each message states that a specific option is the option that gives the highest payment to Player 3.
- Which message will finally be delivered depends on a *private agreement between Players 1 and 2*. The agreement specifies an amount of money that Player 1 transfers to Player 2 for the delivery.
- *Player 2 delivers the message to Player 3 in person.*
- *Player 3 chooses an option* that determines the earnings of all players.

Specific instructions

There are ten options, each one labelled with a unique letter: A, B, C, D, E, F, G, H, I, or J. The computer will randomly assign one option to pay *€10 to Player 1, €10 to Player 2, and €10 to Player 3* and another option to pay *€17 to Player 1, €10 to Player 2, and €3 to Player 3*. The remaining eight options pay *€4 to Player 1, €4 to Player 2, and €0 to Player 3*.

How much each player earns in each option will be shown only to Player 1 and Player 2. The following table is an example of how payments could be assigned to the various options and how this information would be presented to Player 1 and Player 2.

Option	A	B	*C*	D	*E*	F	G	H	I	J
Player 1's payment	4	4	10	4	17	4	4	4	4	4
Player 2's payment	4	4	10	4	10	4	4	4	4	4
Player 3's payment	0	0	10	0	3	0	0	0	0	0

Player 3 will not know which options provide positive earnings for him/her. The table below shows what Player 3 will see.

Option	A	B	C	D	E	F	G	H	I	J
Player 1's payment	?	?	?	?	?	?	?	?	?	?
Player 2's payment	?	?	?	?	?	?	?	?	?	?
Player 3's payment	?	?	?	?	?	?	?	?	?	?

The only information that Player 3 receives regarding the payments of the various options is the message chosen by Player 1 and delivered to Player 3 by Player 2. After receiving the message, Player 3 chooses one of the ten options. The option chosen by Player 3 determines the earnings of all players in the group.

Player 1 chooses a message and reaches an agreement with Player 2

Player 1 chooses *one message* for Player 3. There are *two available messages*. Each message corresponds to one of the two options with positive earnings for all players.

- *Message I* corresponds to the option that pays €10 to Player 3. The message reads "Option <letter of option that pays €10 to Player 3> will earn you 10 euros".
- *Message II* corresponds to the option that pays €3 to Player 3. The message reads "Option <letter of option that pays €3 to Player 3> will earn you 10 euros".

Note that Player 1 cannot choose a message that corresponds to an option that pays €0 to Player 3. Therefore, when Player 3 receives a message, he/she will not know whether the option mentioned in the message pays him/her €10 or €3, but he/she can be certain that the option does not pay him/her €0.

Example

Suppose that the computer randomly assigns payments to options as shown in the table below.

Option	A	B	C	*D*	E	*F*	G	H	I	J
Player 1's payment	4	4	4	17	4	10	4	4	4	4
Player 2's payment	4	4	4	10	4	10	4	4	4	4
Player 3's payment	0	0	4	3	0	10	0	0	0	0

In this case, Player 3 can receive one of the following two messages:

- "Option F will earn you 10 euros"
- "Option D will earn you 10 euros"

Player 1 cannot deliver the message to Player 3. Only Player 2 is able to deliver the message for him/her. If the option mentioned in the message coincides with the option subsequently chosen by Player 3, then Player 1 transfers between €0 and €6.50 to Player 2 for delivery. The screens below will be used to determine which message is delivered and how much is transferred. Each screen displays a list containing 14 rows, each row representing a possible transfer from Player 1 to Player 2. Player 1 and Player 2 must decide between Message I and Message II in each of the 14 rows. Players 1 and 2 make their 14 decisions *simultaneously*, which means that Player 2 will not know Player 1's decisions while he/she is deciding, and vice-versa for Player 1. Specifically, in each row, Player 1 decides between:

- Choosing *Message I* and transferring *€0* to Player 2.
- Choosing *Message II* and transferring *the amount specified in that row* to Player 2.

Similarly, in each row, Player 2 decides between:

- Delivering *Message I* in exchange for a transfer from Player 1 of *€0*.
- Delivering *Message II* in exchange for a transfer from Player 1 equal to *the amount specified in that row*.

After both players have made their decisions, *one of the 14 rows will be randomly selected* by the computer to determine which message will be delivered to Player 3. All rows have the same chance of being selected; therefore, you should make your decision in each row seriously.

Decisions of Player 1

You are Player 1

Option	A	B	C	D	E	F	G	H	I	J
Player 1's payment	10	4	4	4	4	4	4	17	4	4
Player 2's payment	10	4	4	4	4	4	4	10	4	4
Player 3's payment	10	0	0	0	0	0	0	3	0	0

Please decide between Message I and Message II in **each row.**

Message I:
Option A will earn you 10 euros
Player 3 earns **€ 10** if he/she chooses Option A.

Message II:
Option H will earn you 10 euros
Player 3 earns **€ 3** if he/she chooses Option H.

Row	Transfer	Your payment	Earnings Player 2		Transfer	Your payment	Earnings Player 2
1	0	10	10	○ ○	0.0	17.0	10.0
2	0	10	10	○ ○	0.5	16.5	10.5
3	0	10	10	○ ○	1.0	16.0	11.0
4	0	10	10	○ ○	1.5	15.5	11.5
5	0	10	10	○ ○	2.0	15.0	12.0
6	0	10	10	○ ○	2.5	14.5	12.5
7	0	10	10	○ ○	3.0	14.0	13.0
8	0	10	10	○ ○	3.5	13.5	13.5
9	0	10	10	○ ○	4.0	13.0	14.0
10	0	10	10	○ ○	4.5	12.5	14.5
11	0	10	10	○ ○	5.0	12.0	15.0
12	0	10	10	○ ○	5.5	11.5	15.5
13	0	10	10	○ ○	6.0	11.0	16.0
14	0	10	10	○ ○	6.5	10.5	16.5

Decisions of Player 2

You are Player 1

Option	A	B	C	D	E	F	G	H	I	J
Player 1's payment	10	4	4	4	4	4	4	17	4	4
Player 2's payment	10	4	4	4	4	4	4	10	4	4
Player 3's payment	10	0	0	0	0	0	0	3	0	0

Please decide between Message I and Message II in **each row.**

Message I:
Option A will earn you 10 euros
Player 3 earns **€ 10** if he/she chooses Option A.

Message II:
Option H will earn you 10 euros
Player 3 earns **€ 3** if he/she chooses Option H.

Row	Transfer	Your payment	Earnings Player 2		Transfer	Your payment	Earnings Player 2
1	0	10	10	○ ○	0.0	17.0	10.0
2	0	10	10	○ ○	0.5	16.5	10.5
3	0	10	10	○ ○	1.0	16.0	11.0
4	0	10	10	○ ○	1.5	15.5	11.5
5	0	10	10	○ ○	2.0	15.0	12.0
6	0	10	10	○ ○	2.5	14.5	12.5
7	0	10	10	○ ○	3.0	14.0	13.0
8	0	10	10	○ ○	3.5	13.5	13.5
9	0	10	10	○ ○	4.0	13.0	14.0
10	0	10	10	○ ○	4.5	12.5	14.5
11	0	10	10	○ ○	5.0	12.0	15.0
12	0	10	10	○ ○	5.5	11.5	15.5
13	0	10	10	○ ○	6.0	11.0	16.0
14	0	10	10	○ ○	6.5	10.5	16.5

Player 2 will deliver the message determined by the choices in the selected row in the following way:

- In the selected row, if *Player 1 chooses Message I*, then regardless Player 2's choice, *Player 2 delivers Message I*. In this case, if Player 3 chooses the option corresponding to Message I, then Player 1, Player 2, and Player 3 all earn €10.
- In the selected row, if *Player 2 chooses Message I*, then regardless Player 1's choice, *Player 2 delivers Message I*. In this case, if Player 3 chooses the option corresponding to Message I, then Player 1, Player 2, and Player 3 all earn €10.
- In the selected row, if *both Player 1 and Player 2 choose Message II*, then *Player 2 delivers Message II*. In this case, if Player 3 chooses the option corresponding to Message II, then Player 1 earns €17 minus the transferred amount specified in that row, Player 2 earns €10 plus the transferred amount specified in that row, and Player 3 earns €3.

To summarize, Message II is delivered to Player 3 only when both Player 1 and Player 2 choose Message II in the selected row; otherwise Message I is delivered.

Player 3 will *not* be informed which row was selected by the computer.

Example

Suppose that Player 1 and Player 2 make the choices shown below.

Decisions of Player 1

You are Player 1

Option	A	B	C	D	E	F	G	H	I	J
Player 1's payment	10	4	4	4	4	4	4	17	4	4
Player 2's payment	10	4	4	4	4	4	4	10	4	4
Player 3's payment	10	0	0	0	0	0	0	3	0	0

Please decide between Message I and Message II in **each row.**

	Message I: Option A will earn you 10 euros				**Message II: Option H will earn you 10 euros**		
	Player 3 earns **€ 10** if he/she chooses Option A.				Player 3 earns **€ 3** if he/she chooses Option H.		
Row	**Transfer**	**Your payment**	**Earnings Player 2**		**Transfer**	**Your payment**	**Earnings Player 2**
1	0	10	10	○ ●	0.0	17.0	10.0
2	0	10	10	○ ●	0.5	16.5	10.5
3	0	10	10	○ ●	1.0	16.0	11.0
4	0	10	10	○ ●	1.5	15.5	11.5
5	0	10	10	○ ●	2.0	15.0	12.0
6	0	10	10	○ ●	2.5	14.5	12.5
7	0	10	10	○ ●	3.0	14.0	13.0
8	0	10	10	○ ●	3.5	13.5	13.5
9	0	10	10	○ ●	4.0	13.0	14.0
10	0	10	10	○ ●	4.5	12.5	14.5
11	0	10	10	● ○	5.0	12.0	15.0
12	0	10	10	● ○	5.5	11.5	15.5
13	0	10	10	● ○	6.0	11.0	16.0
14	0	10	10	● ○	6.5	10.5	16.5

Decisions of Player 2

You are Player 2

Option	A	B	C	D	E	F	G	H	I	J
Player 1's payment	10	4	4	4	4	4	4	17	4	4
Player 2's payment	10	4	4	4	4	4	4	10	4	4
Player 3's payment	10	0	0	0	0	0	0	3	0	0

Please decide between Message I and Message II in **each row.**

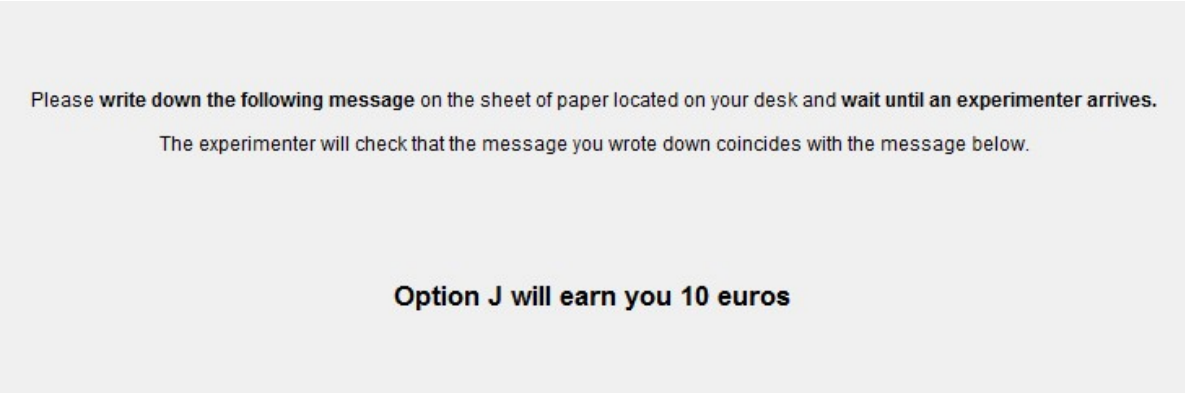
	Message I: Option A will earn you 10 euros				**Message II: Option H will earn you 10 euros**		
	Player 3 earns **€ 10** if he/she chooses Option A.				Player 3 earns **€ 3** if he/she chooses Option H.		
Row	**Transfer**	**Earnings Player 1**	**Your payment**		**Transfer**	**Earnings Player 1**	**Your payment**
1	0	10	10	● ○	0.0	17.0	10.0
2	0	10	10	● ○	0.5	16.5	10.5
3	0	10	10	● ○	1.0	16.0	11.0
4	0	10	10	● ○	1.5	15.5	11.5
5	0	10	10	● ○	2.0	15.0	12.0
6	0	10	10	○ ●	2.5	14.5	12.5
7	0	10	10	○ ●	3.0	14.0	13.0
8	0	10	10	○ ●	3.5	13.5	13.5
9	0	10	10	○ ●	4.0	13.0	14.0
10	0	10	10	○ ●	4.5	12.5	14.5
11	0	10	10	○ ●	5.0	12.0	15.0
12	0	10	10	○ ●	5.5	11.5	15.5
13	0	10	10	○ ●	6.0	11.0	16.0
14	0	10	10	○ ●	6.5	10.5	16.5

In this example, Player 1 is willing to transfer at maximum €4.5 to Player 2 for delivering Message II, while Player 2 demands at least €2.5 for delivering Message II. Given these choices, the following occurs if the computer randomly selects one of the rows below:

- *Row 4:* Since Player 1 chose Message II but Player 2 disagreed in favor of Message I, then Player 2 delivers Message I. Thereafter, if Player 3 chooses the option corresponding to Message I, then Player 1, Player 2, and Player 3 all earn €10.
- *Row 9:* Since Player 1 chose Message II and Player 2 agreed to Message II then Player 2 delivers Message II. Thereafter, if Player 3 chooses the option corresponding to Message II, then Player 1 earns €17 €4 = €13, Player 2 earns €10 + €4 = €14, and Player 3 earns €3.
- *Row 12:* Since Player 1 chose Message I then Player 2 delivers Message I automatically. Thereafter, if Player 3 chooses the option corresponding to Message I, then Player 1, Player 2, and Player 3 all earn €10.

Player 2 delivers the message to Player 3 in person

Once the message is determined, Player 2 will see a screen like the one below.



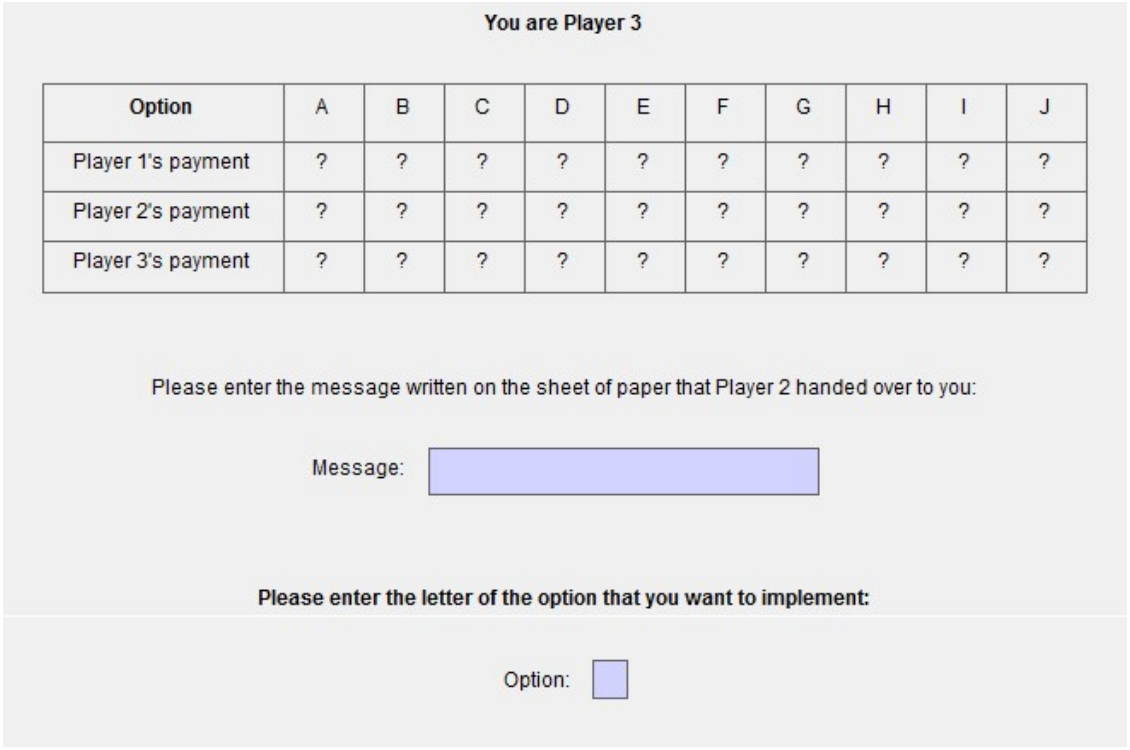
To deliver the message, Player 2 will first *write down the message on the sheet of paper* located on his/her desk. Then, Player 2 will wait until an experimenter arrives. The experimenter will check whether the message written on the sheet of paper is identical to the message shown on the screen. Note that, like Player 3, the experimenter will not know to which payment the option in the message corresponds.

The experimenter will then walk with Player 2 to the desk of the Player 3 of his/her group. At this point, Player 2 will hand the paper with the message to Player 3 and then walk back to his/her desk.

Remember that *any kind of communication between the players is prohibited*, including gestures and facial expressions. In addition, Player 2 is not allowed to write down anything else other than the message on the sheet of paper. Any participant who does not comply with these rules will *not be paid* at the end of the study.

Player 3 chooses an option

Player 3 knows that there are two options with positive payments for him/her, but he/she does not know which two of the ten options contain these payments. *The only information that Player 3 receives is the message delivered to him/her by Player 2.* After receiving the message, Player 3 sees a screen like the one below.



On this screen, Player 3 first confirms the message he/she received by typing it into the text box. Then, he/she chooses one of the ten options. *The option chosen by Player 3 determines the earnings of all players.* Remember that if Player 3 chooses a zero-payment option, the final earnings will be €0 for him/her and €4 for Player 1 and 2.