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

SASCHA BEHNK

University of Zurich and and IU International University of Applied Sciences

LI HAO

Amazon Web Services

ERNESTO REUBEN

New York University Abu Dhabi, Center for Behavioral Institutional Design, and the Luxembourg Institute of Socio-Economic Research

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 an experiment using sender-receiver games in which we elicit subjects' normative and empirical beliefs 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. By contrast, empirical beliefs are unaffected by the additional sender, suggesting that antisocial behavior increases in groups because antisocial actions become more acceptable and not because acceptable behavior is expected less often. We identify a necessary condition for this effect: the additional sender has to actively participate in the decision-making.

This version: March 2022

Note: Reuben gratefully recognizes financial support by Tamkeen under the NYU Abu Dhabi Research Institute Award CG005. Behnk gratefully acknowledges the very helpful support from Aurora García Gallego and Iván Barreda Tarrazona from the Laboratory of Experimental Economics at University Jaume I. This is the authors' version of work that was accepted for publication in the *European Economic Review*. A final version is published in https://doi.org/10.1016/j.euroecorev.2022.104116.

جامعة نيويورك أبوظبي
NYU | ABU DHABI

مركز التصميم السلوكي المؤسساتي
CBID CENTER FOR BEHAVIORAL INSTITUTIONAL DESIGN

1 Introduction

There is increasing evidence showing that people are more likely to behave antisocially when they act together than when they act 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), whistleblowing (Choo et al., 2019), 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. However, it remains unclear why simply knowing that others are involved in the decision-making is sufficient to increase antisocial 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.[2] There is a growing body of literature showing that normative beliefs play a crucial role in decisions involving prosocial and antisocial behaviors.[3] Importantly, recent studies indicate that elicited normative beliefs can predict changes in behavior that are induced by subtle contextual variations (Krupka and Weber, 2013).[4] We extend this literature by investigating whether the mere presence of other decision-makers makes antisocial actions more normatively acceptable, resulting in more antisocial behavior.

Our experimental setup is designed to evaluate the impact of an additional decision-maker on normative beliefs in multiple ways. First, we elicit the normative beliefs of potential offenders and potential victims by asking them to rate the acceptability of the antisocial action in our experiment. This way, we can observe whether the involvement of another decision-maker leads to a general perceptual change in normative beliefs. Second, we ask subjects to predict others' acceptability ratings and reward them for the accuracy of their prediction to obtain an incentivized measure of their normative beliefs. Third, given that compliance with normatively-desirable behavior requires both a shared understanding about the acceptability of actions and a

[1] There are other reasons for increased antisocial behavior in groups. In this paper, we study the effect of the presence of another decision-maker and rule out most other explanations by design.

[2] An often cited explanation for this phenomenon is that there is "diffusion of responsibility" (see, Dana et al., 2007). We think that the study of normative beliefs complements this approach instead of providing a separate explanation. We discuss this further in the context of our results in the conclusions.

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

[4] Normative beliefs 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).

shared belief that others will behave acceptably (Bicchieri, 2006), we also measure the subjects' beliefs about the actions of others. Specifically, we elicit two empirical beliefs: (i) we ask the subjects who decide between a prosocial and an antisocial action about how prevalent the antisocial action is among *other decision-makers*, and (ii) we ask these subjects about the *potential victims'* expected prevalence of the antisocial action. Measuring normative and empirical 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?

To be more specific, we employ sender-receiver games customized to study the above-mentioned questions. In the games, a receiver determines the earnings of all players by choosing one out of ten options. 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 identifies either the option that gives everyone an equal payoff, which we call prosocial, or the unequal option that benefits senders at the expense of receivers, which we call antisocial. The remaining eight options are Pareto-dominated and pay all players a smaller amount. Hence, receivers have an incentive to follow the senders' message, even if they expect to receive the antisocial message.

We compare a treatment where the antisocial message is chosen by a single sender to treatments where sending the antisocial message depends on the choices of two senders who cannot communicate or bargain with each other. To test whether the rule that aggregates the choices of the two senders matters, we run a treatment where unanimity is required to send the antisocial message and another treatment were the antisocial message is sent if any of the two senders chooses it. To study the extent to which differences in behavior and normative beliefs are due to the involvement of a second decision-maker and not just the presence of another person, we also run a treatment in which there are two senders but only one of them determines which message is sent. In other words, a treatment in which one sender is actively involved in the decision while the other sender is passive.[5]

In line with previous literature, we see more antisocial behavior when decisions are made jointly. Moreover, the increase in antisocial behavior occurs for both choice aggregation rules. More importantly, we find evidence consistent with a shift in normative beliefs being the reason why decisions are more antisocial in treatments with two active senders. First, we find that

[5] In the experiment analyzed in Appendices B and C, we include additional design elements. First, we use a price list to elicit the precise monetary value individuals place on acting prosocially. Second, we implement two types of antisocial messages (a deceptive and a truthful message). Third, we elicit the intensity with which subjects experience guilt since it is a crucial emotion for compliance with social norms (see Elster, 2009; Bicchieri et al., 2018; López-Pérez, 2010).

subjects in treatments with two active senders think that sending the antisocial message is significantly more acceptable than subjects in treatment with only one sender. Importantly, this view is held by both senders and receivers, which demonstrates that the shift in normative beliefs occurs on a general level and is not the result of senders' internal justification of their own choices. By contrast, both senders' and receivers' empirical beliefs are statistically indistinguishable across treatments.

Analyzing the data at the individual level reveals that senders' normative and empirical beliefs are significant determinants of antisocial behavior. Consistent with models of social norms (e.g., Bicchieri, 2006; Krupka and Weber, 2013; Barr et al., 2018), we find that senders act prosocially if they think antisocial actions are unacceptable, and they believe other senders will act prosocially as well. Interestingly, in line with models of guilt aversion (Battigalli and Dufwenberg, 2007), we also find that senders act even more prosocially if they also believe that choosing the antisocial action will negatively surprise the receiver.

Finally, we find that the prevalence of antisocial messages and the acceptability of sending the antisocial message are lower in the treatment with an active and a passive sender than in the treatments with two active senders. In fact, antisocial behavior and normative beliefs in this treatment are statistically indistinguishable to those in the treatment with one sender. This result shows that the second sender's active involvement in the decision-making process is crucial for increasing antisocial behavior in groups to occur. Moreover, this result suggests that audience effects and the senders' social preferences regarding other senders do not drive the increase in antisocial behavior in groups.

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.[6] 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 with others in a market than when they make their decisions individually. The result that market environments lead to an erosion

[6] Relatedly, increased antisocial behavior due to the involvement of others has also been observed through delegation (Hamman et al., 2010; Bartling and Fischbacher, 2012) and intermediation (Coffman, 2011; Oexl and Grossman, 2013; Garofalo and Rott, 2018).

of morality is further supported by Kirchler et al. (2016) and Deckers et al. (2016). Bartling et al. (2015) find that subjects are less socially responsible in market settings than in settings with individual decision-making 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 beliefs of others is not revealed through market interactions. Moreover, we do not vary the framing between an individual choice and a market trade that could, on its own, change the subjects' normative beliefs of choosing the antisocial outcome. We concentrate solely on the effect of the inclusion of another decision-maker on individuals' willingness to behave antisocially.

2.2 Increased antisocial behavior via group decisions

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; Nielsen et al., 2019), and as workers in gift-exchange games (Kocher and Sutter, 2007). Falk et al. (2020) find that the fraction of unethical decisions is higher when subjects are in groups and cannot know whether their decision was pivotal. In the context of deception, which is the type of antisocial behavior at the core of our investigation, Keck (2014) provides evidence that receivers who make joint decisions in an ultimatum game are more likely to deceive the proposer than single receivers. In sender-receiver games, Sutter (2009) finds that groups are more likely to use strategic deception by telling the truth when receivers are expected not to follow their message. Cohen et al. (2009) show that senders making joint decisions deceive receivers more often than individual senders when they are informed of the receiver's decision ahead of time. Variations of the die-rolling game (Fischbacher and Föllmi-Heusi, 2013) provide further evidence of comparatively more dishonest acts by groups (Gino et al., 2013; Muehlheusser et al., 2015; Weisel and Shalvi, 2015; Kocher et al., 2018; Korbel, 2017; Barr and Michailidou, 2017). Other studies report increased dishonesty when lies benefit others through aligned payoffs (Wiltermuth, 2011; Gino and Pierce, 2010; Conrads et al., 2013; Gino et al., 2013; Weisel and Shalvi, 2015).[7]

[7] We should also note that there is also some evidence of less antisocial behavior by groups than individuals. Sutter (2009) finds that groups of senders lie less than single senders when they expect receivers will follow their message. Cohen et al. (2009) find that groups deceive less than individuals when the receivers' behavior is unknown. Danilov et al. (2013) find that aligned payoffs for individuals result in more dishonesty only when they share strong social ties. In dictator games, Cason and Mui (1997) find more selfish behavior by individual than groups (but see the critique by Luhan et al., 2009).

When comparing behavior by individuals and groups, there are many reasons why groups might act more antisocially. By eliminating all communication between decision-makers, by design, we can rule out strategic considerations (e.g., because of the aggregation rule used in the group decision-making process), peer influences and signaling, and argumentation effects. We can also isolate audience effects and the effect of social preferences towards other decision-makers using a treatment with one active and one passive sender. Previous studies show that the presence of third-party observers can reduce antisocial behavior (for a review see Dear et al., 2019). Our design is different from these studies in that the passive sender benefits from the antisocial action in the same way as the decision-maker.

2.3 Normative beliefs and behavior

The third line of research to which our paper contributes is research on the association between normative beliefs and behavior. Although the role of social norms has been discussed in economics for some time (e.g. 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 (Reuben and Riedl, 2013; Bicchieri et al., 2021), 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. Also, this is one of the few papers in economics that provides simultaneous evidence on the links between normative beliefs, empirical beliefs (Bicchieri et al., 2020), experienced emotions, and behavior (another example is Reuben and van Winden, 2010). Lastly, it is the first paper to simultaneously explore the role of two types of empirical beliefs: beliefs of other decision-makers' behavior and beliefs about the expectations of the victims of antisocial actions.

3 Experimental design

We ran two experiments to study whether changes in normative beliefs explain the differences in antisocial behavior due to joint decision-making. We ran a lab experiment in the spring of 2015 at the Laboratory of Experimental Economics (LEE) of the University Jaume I in Castellon. We also ran an online experiment in the summer of 2020 and 2021 to replicate findings from the lab experiment using a simpler experimental design. In the main text of the paper, we concentrate on the online experiment. We briefly discuss the design and the findings of the lab

Table 1. Examples of payoff tables in the sender-receiver games (amounts in US dollars)

A. *1-Sender* treatment

Option	A	B	C	D	E	F	G	H	I	J
Sender	2	2	5	2	6.75	2	2	2	2	2
Receiver	0	0	5	0	1.50	0	0	0	0	0

B. *2-Sender-Consensus*, *2-Sender-Unilateral*, & *Passive-Sender* treatments

Option	A	B	C	D	E	F	G	H	I	J
Sender A	2	2	5	2	6.75	2	2	2	2	2
Sender B	2	2	5	2	6.75	2	2	2	2	2
Receiver	0	0	5	0	1.50	0	0	0	0	0

experiment in Section 6.

The online experiment consists of four treatments, each using a different sender-receiver game. In the *1-Sender* treatment, subjects are randomly assigned either the role of sender or 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 $5 to each player; one antisocial option that pays $6.75 to the sender and $1.50 to the receiver, and eight Pareto-dominated options that pay $2 to the sender and $0 to the receiver. At the beginning of the game, the ten options are randomly labeled with letters ranging from A to J. Although both players know the payoff consequences of choosing an option, only the sender knows how the ten options are labeled. Table 1A contains an example of a letter assignment and how we presented this information to the sender.

In the *2-Sender-Consensus*, *2-Sender-Unilateral*, and *Passive-Sender* treatments, subjects are randomly assigned either the role of sender A, sender B, or receiver. The payoffs of sender A and the receiver are identical to those of the sender and receiver in the *1-Sender* treatment. The additional player, sender B, receives identical payoffs as sender A (see Table 1B).

In all treatments, the only information available to the receiver regarding the label assignment of the ten options is due to a message. There are two available messages. The prosocial message (Message I) accurately reveals the prosocial option's label and reads "Option [letter paying the receiver $5] will earn you $5." The antisocial message (Message II) is untruthful in that it reveals the antisocial option's label but claims it is the label of the prosocial option: "Option [letter paying the receiver $1.50] will earn you $5." Like senders, receivers are aware that there are two available messages and that one of them is deceptive. Hence, it is common knowledge that a message always reveals the label of either the prosocial or the antisocial option but never the label of one of the eight Pareto-dominated options.

In *1-Sender* and *Passive-Sender*, one individual chooses the message: the sender in *1-Sender* and sender A in *Passive-Sender*. In *Passive-Sender*, sender B does not make any decisions. By contrast, in *2-Sender-Consensus* and *2-Sender-Unilateral*, sender A and sender B make this decision jointly. Specifically, each sender independently chooses either the prosocial or the antisocial message. In *2-Sender-Consensus*, the antisocial message is sent only if both senders choose the antisocial message while the prosocial message is sent if at least one of the senders chooses the prosocial message. The converse is true in *2-Sender-Unilateral*, where the antisocial message is sent if at least one of the senders chooses the antisocial message while the prosocial message is sent only if both senders choose the prosocial message.

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 the prosocial message is $pU(P) + (1-p)(1/9)U(A) + (1-p)(8/9)U(D)$ and that of sending the antisocial message 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 the prosocial message and senders who prefer the antisocial option (i.e., for whom $U(P) < U(A)$) are better off choosing the antisocial message. 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 80.4% of the receivers followed the message they received, and 98.6% of the senders expected more than 11% of the receivers to follow their message.

3.1 Normative beliefs

In all treatments, we elicit the senders' *normative beliefs* regarding the prosocial and antisocial messages. We do so after senders make their decisions, but before they learn the outcome of the game. Specifically, we ask senders to put themselves in the position of a neutral uninvolved arbitrator and indicate for each message "How socially acceptable is sending Message I [or Message II] to Player [player # of the receiver]?" Answers are recorded with a 5-point Likert scale ranging from "very unacceptable" (1) to "very acceptable" (5).

In addition to the senders, we also ask 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 beliefs are important for two reasons. First, they allow us to evaluate whether the senders' normative beliefs are self-serving. Second, they can tell us whether a second sender's inclusion affects only the senders' normative perceptions or whether it produces a more general perceptual change.

Finally, we also elicit the senders' expectation of the receivers' normative beliefs (Bicchieri and Xiao, 2009; Bicchieri and Chavez, 2010). After indicating their normative beliefs, we show senders the question used to measure the normative beliefs of receivers. Thereafter, we ask them to indicate "What do you think was Player [player # of the receiver]'s answer to this question?" We incentivize their answer by paying them $0.50 for a correct guess. With this method, we obtain an incentivized measure of the senders' normative beliefs.[8] An alternative method for obtaining an incentivized measure of normative beliefs is to pay subjects if their normative beliefs 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 beliefs, which could be the case in our games given the asymmetry between senders and receivers.[9] In vignette studies, Aycinena and Kimbrough (2021) report that these two methods result in highly correlated answers.

There is also some discussion in the literature on the merits of eliciting norms between-subjects (i.e., eliciting norms and behavior from completely separate individuals) or within-subjects. On the one hand, within-subject norm elicitation can investigate questions that cannot be answered by between-subject elicitations. For example, they allow us to test the impact of normative beliefs on behavior at the individual level and identify self-serving biases. On the other hand, one might worry that within-subject elicitation might be affected by motives such as preferences for consistency or self-concept maintenance. Reassuringly, in recent work, D'Adda et al. (2016) conclude that eliciting norms after subjects play a game does not distort norm measurements.

[8] More specifically, since we ask subjects for their own normative judgements, it can be said that we are measuring the senders' and receivers' *personal* normative beliefs, and the senders' expectations about the personal normative beliefs of receivers. There is some discussion in the literature concerning the distinction between personal and social norms (Bicchieri et al., 2018) and whether they have different effects on behavior (Bašić and Verrina, 2021). Since we cannot shed light on this distinction in this study, we use the more general term 'normative beliefs'.

[9] 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 beliefs and their perception of others' normative beliefs because they can have different effects on 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.

3.2 Empirical beliefs

We elicit senders' and receivers' *empirical beliefs*, by which we mean beliefs about the behavior of others. As we will discuss in more detail in Section 4, there are two empirical beliefs that are potential determinants of people's decision to adhere to normatively-prescribed behavior.[10]

- Bicchieri (2006) argues that senders are more likely to adhere to a socially-prescribed action if other senders choose that action as well. Hence, we elicit the senders' expected fraction of other senders choosing the antisocial message. To do so, we ask senders to indicate "Out of 10 Player 1s [or pairs of Player 1s and 2s], how many do you think sent Message II?" We incentivize the belief elicitation by paying senders \$0.50 for a correct guess.

- Theories of guilt aversion (Battigalli and Dufwenberg, 2007) assume that senders will adhere to behavior that satisfies the receivers' expectations. Therefore, we elicit the senders' belief about the receivers' expected fraction of antisocial messages sent. To do so, we first ask receivers to indicate "Out of 10 Player 1s [or pairs of Player 1s and 2s], how many do you think sent Message II?" Thereafter, we show this question to the senders and ask them to indicate "What do you think was Player [player # of the receiver]'s answer?" We also incentivize these elicitations by paying subjects \$0.50 per correct answer.

3.3 Procedures

A total of 1,157 subjects participated in the online experiment, 734 as senders, and 423 as receivers. We recruited subjects using Prolific, an online research subject pool. Subjects were restricted to reside in the United States and be at least 18 years old. Overall, 60.2% of the subjects are male, 72.9% self-identify as 'white,' 66.2% hold at least a bachelor's degree, and 76.1% are employed.

Subjects took part in the experiment asynchronously. Namely, they completed all their decisions before being matched with other subjects. We did the matching the same day and communicated the results by email. To ensure that subjects carefully read the instructions, we included several control questions that subjects had to answer correctly to continue. We warned subjects that an incorrect answer implied automatic removal from the experiment.[11] A sample of the instructions is available in Appendix D. On average, subjects took 13.45 minutes to complete the study and earned \$7.85, including a \$3.30 show-up fee.

[10] We also elicit the sender's beliefs about the receivers' behavior by asking them "Out of 10 Players [player # of receivers], how many will follow the message they received?"

[11] Around 1 out of 3 subjects failed at least of one the control questions.

4 Hypotheses

In this section, we formulate several hypotheses to guide the data analysis. Our first hypothesis is based on the empirical literature described in section 2. 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. Hence, our first hypothesis simply states that we expect to replicate this common finding in the literature.

Hypothesis 1 *More senders choose the antisocial message in the 2-Sender-Consensus and 2-Sender-Unilateral treatments than in the 1-Sender treatment.*

The *1-Sender* treatment and the treatments with two senders differ in two ways. First, in the inclusion of another decision-maker, and second, in the payoff consequences of the different outcomes (both in terms of efficiency and earnings comparisons). The *Passive-Sender* treatment allows us to distinguish these two effects. Differences between the *2-Sender* treatments and the *Passive-Sender* treatment identify the effect of having a second decision-maker, while differences between *1-Sender* and *Passive-Sender* identify the effect of the change in payoff consequences. Note that identifying these two effects allows us to test whether social preferences explain changes in antisocial behavior. In particular, in models such as Fehr and Schmidt (1999), Bolton and Ockenfels (2000), and Charness and Rabin (2002), individuals' other-regarding concerns depend on the number of players. Hence, they predict that senders in the *2-Sender* treatments and the *Passive-Sender* treatment place less weight on the receiver's welfare than senders in the *1-Sender* treatment, which leads to more antisocial behavior in the former treatments.[12]

In addition to these two effects, a couple of models have been proposed recently to explain how increasing the number of decision-makers can lead to more antisocial behavior. These models assume that individuals dislike antisocial actions as long as they are 'responsible' for them. More precisely, they build on a notion of responsibility in which individuals' motivation to act prosocially depends on the probability that their choice is pivotal in determining the antisocial action (Engl, 2017; Rothenhäusler et al., 2018). In our case, senders in *2-Sender-Consensus* are always pivotal in determining the antisocial action (i.e., they have veto power) but this is not the case in *2-Sender-Unilateral*, where a sender choosing the prosocial action might not prevent the antisocial message from being sent. Therefore, these models would predict more antisocial behavior in *2-Sender-Consensus* compared to *2-Sender-Unilateral*.

[12] Models of social preferences that incorporate intentions can predict differences between the *2-Sender* treatments and the *Passive-Sender* treatment. In these models, the players' concern for others depends on their beliefs (Geanakoplos et al., 1989). Although these models are often complex, it is straightforward to see that they predict a change in behavior if the second sender's inclusion alters beliefs. This class 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 later on.

Assuming these three effects are at play, gives us the following hypotheses.

Hypothesis 2A *More senders choose the antisocial message in the 2-Sender-Consensus and 2-Sender-Unilateral treatments than in the Passive-Sender treatment.*

Hypothesis 2B *More senders choose the antisocial message in the Passive-Sender treatment than in the 1-Sender treatment.*

Hypothesis 2C *More senders choose the antisocial message in the 2-Sender-Consensus than in the 2-Sender-Unilateral treatment.*

Our subsequent hypotheses are constructed to test the idea that normative beliefs explain changes in behavior between the *1-Sender* and *2-Sender* treatments. Note that we do not attempt to directly compare models of normative beliefs to models that incorporate other motivations (for such an attempt, see Gächter et al., 2013). Instead, our approach is to measure variables used in models of normative beliefs and then test whether variation in these empirical measures is consistent with differences in behavior between treatments and between individuals within the same treatment.[13]

Our next hypothesis is based on models of social norms (e.g., Krupka and Weber, 2013; Barr et al., 2018). In these models, individuals maximize a utility function that includes their monetary payoff and their belief of the social acceptability of choosing an action (i.e., their normative beliefs). If we assume that sending the antisocial message is less acceptable than sending the prosocial message, then the senders' decision depends on how they trade-off the higher monetary payoff of the antisocial outcome with the lower social acceptability of sending the antisocial message. Given that the antisocial outcome has the same monetary payoffs in all treatments, if there is support for Hypothesis 1, we should also see treatment differences in normative beliefs. This line of thought gives us two related hypotheses.

Hypothesis 3A *Conditional on finding support for Hypothesis 1, on average, senders rate sending the antisocial message as more acceptable in the 2-Sender-Consensus and 2-Sender-Unilateral treatments than in the 1-Sender treatment.*

Hypothesis 3B *Within each treatment, there is a positive association between the senders' acceptability rating of sending the antisocial message and choosing the antisocial message.*

Our next hypothesis touches on the role of empirical expectations in models of normative beliefs. By empirical expectations, we mean expectations about the behavior of others. In

[13] Since most of the empirical literature does not have a control treatment analogous to the *Passive-Sender* treatment, we do not formulate further specific hypotheses concerning this treatment. Nevertheless, we still test whether differences in behavior concerning the *Passive-Sender* treatment are congruent with differences in their normative and empirical beliefs.

her seminal work on social norms, Bicchieri (2006) argues that individuals adhere to socially acceptable behavior only if sufficiently many others do so as well. In a one-shot setting like ours, this conditionality in norm adherence can be studied by looking at the senders' beliefs about the behavior of other senders. Namely, it predicts that, senders who expect other senders to behave antisocially are more likely to send the antisocial message than senders who expect other senders to behave prosocially. In the context of our experiment, these models suggest that the frequency of antisocial messages could change from the *2-Sender* treatments to the *1-Sender* treatment solely due to changes in the senders' empirical beliefs. Specifically, finding support for Hypothesis 1 and finding that senders expect other senders to send the antisocial message more in the *2-Sender* treatments than in the *1-Sender* treatment would be evidence consistent with Bicchieri (2006).[14] This argument gives us our next hypotheses.

Hypothesis 4A *Conditional on finding support for Hypothesis 1, on average, senders expect a higher fraction of other senders to choose the antisocial message in the 2-Sender-Consensus and 2-Sender-Unilateral treatments than in the 1-Sender treatment.*

Hypothesis 4B *Within each treatment, there is a positive association between the senders' expected fraction of other senders choosing the antisocial message and choosing the antisocial message.*

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. We find it natural to interpret this difference in utility between acceptable and unacceptable actions as differences in the intensity with which individuals feel guilt (Baumeister et al., 1994). Thus, we also consider the predictions of prominent models in economics that explain prosocial behavior as a consequence of individuals avoiding feelings of guilt. In models of (simple) guilt aversion (Battigalli and Dufwenberg, 2007), the sender's message choice depends on two factors: the sender's sensitivity to guilt and the degree to which the sender thinks the antisocial outcome will disappoint the receiver. Hence, in these models, empirical beliefs are also a determining factor of antisocial behavior. The difference between these models and those of social norms is that in models of guilt aversion, the relevant empirical belief is the senders' belief of the receivers' expected fraction of senders choosing the antisocial message. Consequently, support for Hypothesis 1 is consistent with guilt aversion models if these beliefs are higher in the *2-Sender* treatments than in the *1-Sender* treatment.[15]

[14] Note that the converse is not necessarily true. Social norms, as modeled by Bicchieri (2006), are consistent with support for Hypothesis 1 even if the senders' empirical beliefs are equal across treatments. This would be the case if including a second sender leaves empirical beliefs unchanged but changes the senders' normative beliefs (as in Hypothesis 3A).

[15] Once again, the converse is not necessarily true. Models of guilt aversion can be consistent with Hypothesis 1

Hypothesis 5A *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-Consensus and 2-Sender-Unilateral treatments compared to the 1-Sender treatment.*

Hypothesis 5B *Within each treatment, there is a positive association between the senders' belief of the receivers' expectation of receiving the antisocial message and choosing the antisocial message.*

Our final two hypotheses concern the relationship between normative beliefs and empirical expectations. According to Bicchieri (2006), individuals behave according to a social norm when both normative and empirical expectations coincide. That is to say, the senders who are most likely to choose the prosocial message are senders who think that sending the antisocial message is normatively unacceptable and expect most other senders will send the prosocial message.

Similarly, even though Battigalli and Dufwenberg (2007) do not specify where guilt sensitivity comes from, we think that the normative beliefs of individuals are a natural interpretation of their guilt sensitivity. If the elicited normative beliefs capture the senders' sensitivity to guilt, then models of guilt aversion predict a positive relationship between choosing the antisocial message and the *interaction* of the senders' normative beliefs and their belief of the receivers' expected fraction of senders choosing the antisocial message. In other words, the senders with the highest likelihood of sending the prosocial message are senders who think that sending the antisocial message is very unacceptable and believe receivers expect to receive the prosocial message. These predictions constitute our next formal hypotheses.

Hypothesis 6A *Within each treatment, choosing the antisocial message is negatively associated with the interaction of the senders' acceptability ratings and their expected fraction of other senders choosing the antisocial message.*

Hypothesis 6B *Within each treatment, choosing the antisocial message is negatively associated with the interaction of the senders' acceptability ratings and their belief of the receivers' expectation of receiving the antisocial message.*

5 Results

As mentioned above, here, we present the analysis of the online experiment. We discuss the results of the lab experiment in Section 6. Throughout this section, we report p-values of two-sided tests. Moreover, when we perform multiple treatment comparisons, we report both an uncorrected p-value (labeled with a p) and a p-value corrected for multiple testing using the

if including a second sender leaves beliefs unchanged but decreases the senders' guilt sensitivity.

Benjamini and Hochberg (1995) method (labeled with a p^c). The data of the online experiment is available at Behnk et al. (2022).

5.1 Antisocial behavior

Figure 1 plots the fraction of senders who send the antisocial message to the receiver in the *1-Sender*, *Passive-Sender*, *2-Sender-Consensus*, and *2-Sender-Unilateral* treatments. In line with the literature, only some senders are willing to profit by acting antisocially. In all treatments, a majority of senders send the prosocial message. Before proceeding to make pairwise comparisons, we use a Fisher's exact test to test the null hypothesis of no treatment differences. We can reject this null hypothesis with a high degree of certainty ($p = 0.006$).

Consistent with Hypothesis 1, having a second sender significantly increases the fraction of senders who send the antisocial message. In *1-Sender*, 18.87% of senders act antisocially compared to 34.55% in *2-Sender-Consensus* (test of proportions; $p = 0.004$, $p^c = 0.022$) and 31.07% in *2-Sender-Unilateral* (test of proportions, $p = 0.024$, $p^c = 0.049$). Interestingly, we also find that the fraction of senders who send the antisocial message in *Passive-Sender*, 21.37%, is close to that in *1-Sender* (test of proportions; $p = 0.642$, $p^c = 0.642$) and significantly lower than in *2-Sender-Consensus* (test of proportions; $p = 0.012$, $p^c = 0.036$) and *2-Sender-Unilateral* (test of proportions; $p = 0.067$, $p^c = 0.101$). Lastly, we do not find that the fraction of senders who send the antisocial message differs between *2-Sender-Consensus* and *2-Sender-Unilateral* (test of proportions; $p = 0.465$, $p^c = 0.558$). Therefore, in line with Hypothesis 2A but contrary to Hypotheses 2B and 2C, increased antisocial behavior by groups appears to be driven solely by the inclusion of additional decision-makers.[16] These findings establish our first result.

Result 1 *A second sender's involvement significantly increases antisocial behavior as long as the second sender is actively involved in making the decision.*

5.2 Normative beliefs

Next, we test the hypothesized effects of having a second sender on the subjects' normative beliefs. Figure 2 depicts the senders' average acceptability rating of sending the antisocial message and their average belief of the receivers' acceptability rating. In *Passive-Sender*, we distinguish between the active sender (A) and the passive sender (B). The figure also depicts the receivers' average acceptability rating. Given that normative beliefs are measured in a discrete scale, ranging from very unacceptable (1) to very acceptable (5), we use ordered probit

[16] We are well-powered to detect significant differences of the magnitude seen in the experiment. With a power of 80% and the number of observations we have in each treatment, the minimal detectable difference for the pairwise comparisons ranges from 13.88% to 16.66%. For reference, the difference between *1-Sender* and *2-Sender-Consensus* is 15.68%.

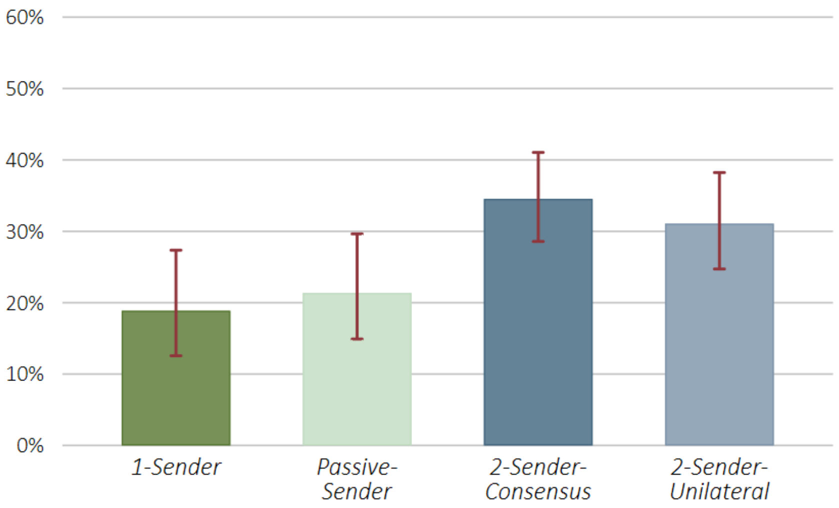


Figure 1. Fraction of senders who send the antisocial message by treatment

Note: Error bars correspond to 95% Wilson confidence intervals.

regressions to test whether the treatment differences are statistically significant. The regression coefficients are provided in Table A1 in Appendix A. Note that in all three regressions, we can reject the null hypothesis of no treatment differences (χ^2 tests; $p < 0.028$).

In line with Hypothesis 3A, we find that senders in *1-Sender* think it is less acceptable to send the antisocial message than senders in *2-Sender-Consensus* ($p = 0.001$, $p^c = 0.004$) and *2-Sender-Unilateral* ($p < 0.001$, $p^c = 0.003$). We see the same pattern if we look at our incentivized measure of normative beliefs (i.e., the senders' belief of the receivers' acceptability ratings). Senders in *1-Sender* think receivers find the antisocial message to be significantly less acceptable than senders in *2-Sender-Consensus* ($p = 0.001$, $p^c = 0.007$) and *2-Sender-Unilateral* ($p < 0.001$, $p^c = 0.001$).[17] Importantly, these treatment differences are not unique to senders. Receivers in *1-Sender* also think it is less acceptable to send the antisocial message than receivers in *2-Sender-Consensus* ($p = 0.015$, $p^c = 0.090$) and *2-Sender-Unilateral* ($p = 0.077$, $p^c = 0.154$).[18]

Consistent with the behavioral differences between treatments. We do not find differences between the *1-Sender* treatment and the *Passive-Sender* treatment in the senders' acceptability

[17] Interestingly, in all treatments, senders incorrectly expect a self-serving bias. Namely, they think that receivers believe sending the antisocial message is less acceptable than their own belief (Wilcoxon signed-rank tests, $p < 0.001$, $p^c < 0.001$). However, the receivers actual normative beliefs are not significantly different from those of senders (Mann-Whitney U tests, $p > 0.314$, $p > 0.833$).

[18] We see the same pattern if we compare the *Passive-Sender* treatment to the *2-Sender-Consensus* and *2-Sender-Unilateral* treatments. Specifically, compared to the senders in the two *2-Sender* treatments, both active and passive senders in *Passive-Sender* think that sending the antisocial message is less acceptable ($p < 0.004$, $p^c < 0.009$) and expect receivers to think it is less acceptable ($p < 0.050$, $p^c < 0.086$). Similarly, the receivers' acceptability ratings in *Passive-Sender* are lower than those in *2-Sender-Consensus* ($p = 0.017$, $p^c = 0.052$) and *2-Sender-Unilateral* ($p = 0.080$, $p^c = 0.121$).

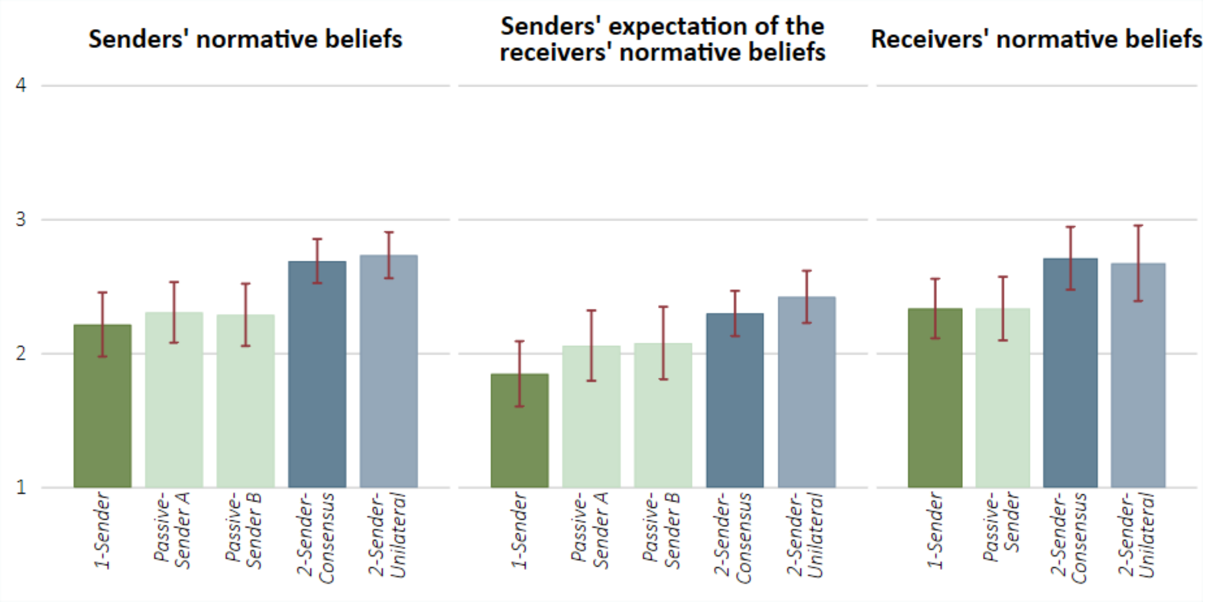


Figure 2. Normative beliefs and expected normative beliefs by treatment and role

Note: Average acceptability of sending the antisocial message (from very unacceptable [1] to very acceptable [5]). For the *Passive-Sender* treatment, we distinguish between active senders (A) and passive senders (B). Error bars correspond to 95% confidence intervals.

ratings ($p > 0.559$, $p^c > 0.799$), the senders' beliefs of the receivers' acceptability ratings ($p > 0.324$, $p^c > 0.364$), and the receivers' acceptability ratings ($p = 0.975$, $p^c = 0.975$). Similarly, we do not find differences between the *2-Sender-Consensus* and *2-Sender-Unilateral* treatments in our three measures of normative beliefs ($p > 0.320$, $p^c > 0.458$).[19] 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 the antisocial message as long as the second sender is actively involved in making the decision.*

A common concern with self-reported measures, such as the senders' normative beliefs, is that they could be noisy or biased due to senders self-justifying their behavior. We argue that this is not the case for our elicited normative beliefs for three reasons. First, we observe treatment differences in the senders' belief of the receivers' acceptability ratings. The elicitation of these beliefs is incentivized and, therefore, less susceptible to misreporting. Second, there is no evidence that senders in either *2-Sender* treatment rate the acceptability of the prosocial message differently from senders in the *1-Sender* ($p > 0.283$) or *Passive-Sender* ($p > 0.506$) treatments, which one would expect if normative evaluations were *post hoc* justifications for behavior. Third, antisocial messages are rated as more acceptable in the *2-Sender* treatments

[19] With a power of 80% and the number of observations we have in each treatment, the minimal detectable difference in normative beliefs for the pairwise comparisons ranges from 0.35 to 0.47 (assuming the observed standard deviations). For reference, the difference between *1-Sender* and *2-Sender-Consensus* is 0.47.

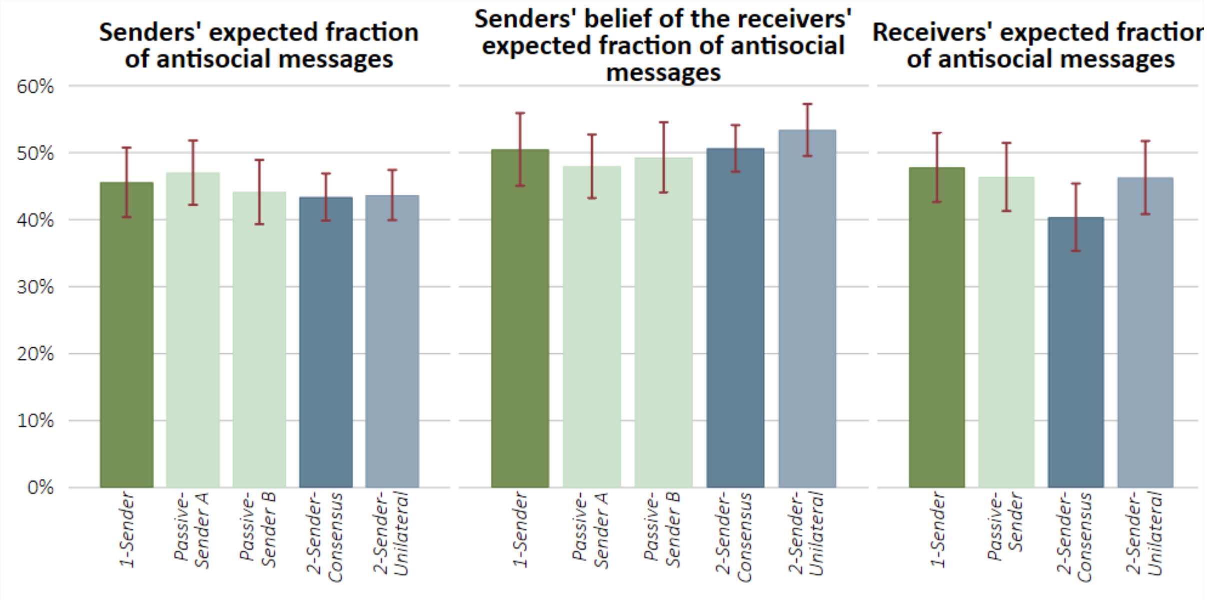


Figure 3. Expected fraction of antisocial messages by treatment and role

Note: Average expected fraction of other senders choosing the antisocial message. For the *Passive-Sender* treatment, we distinguish between the beliefs of active senders (A) and passive senders (B). Error bars correspond to 95% confidence intervals.

not only by senders but also by receivers. This fact demonstrates that the change in normative beliefs is not due to self-serving reporting by senders.

5.3 Empirical beliefs

Now, we turn to subjects' beliefs about the actions of others. We elicited two beliefs for the senders: (i) their expected fraction of other senders choosing the antisocial message, and (ii) their belief of the receiver's expected fraction of senders choosing the antisocial message. For receivers, we elicited their expected fraction of senders choosing the antisocial message. Figure 3 depicts these beliefs for each treatment, separating the beliefs of senders in *Passive-Sender* depending on whether the sender was active (A) or passive (B). We use Tobit regressions to test for treatment differences as beliefs are censored at 0% and 100%. The regression coefficients are provided in Table A2 in Appendix A. In this case, we cannot reject the null hypothesis of no treatment differences in any of the three regressions (F tests; $p > 0.257$).

Overall, we do not find support for Hypothesis 4A as the expected fraction of senders choosing the antisocial message does not vary systematically across treatments. On average, senders think that 45.6% of senders choose the antisocial message in *1-Sender*, 45.5% in *Passive-Sender*, 43.4% in *2-Sender-Consensus*, and 43.7% in *2-Sender-Unilateral*. These fractions are not significantly different from each other (pairwise tests; $p > 0.221$, $p^c > 0.955$). We also do not find support for Hypothesis 5A. The senders' belief of the receivers' expected fraction of senders choosing the antisocial message is very similar and shows no significant differences across treatments (pairwise tests; $p > 0.110$, $p^c > 0.858$). The same can be said about the receivers' beliefs, where the expected fraction of senders choosing the antisocial message is even slightly higher in *1-Sender* compared to *2-Sender-Consensus* ($p = 0.061$, $p^c = 0.364$) and *2-Sender-Unilateral* ($p = 0.613$, $p^c = 0.919$).[20] These findings establish our third result.

Result 3 *Both the senders' and receivers' belief about the fraction of senders choosing the antisocial message remain unchanged with the involvement of a second sender.*

5.4 Determinants of antisocial behavior

To test the hypotheses about behavior within treatments, we conduct a series of probit regressions of the senders' message choice. In all regressions, the dependent variable equals one if a sender chooses the antisocial message and zero if the sender chooses the prosocial message. We report the regressions' marginal effects in Table 2. Given the similarity in behavior between the *1-Sender* and *Passive-Sender* treatments, we pool the data from these treatments to analyze the case where there is one active decision-maker (Panel A). Similarly, given the similarity in behavior between the *2-Sender-Consensus* and *2-Sender-Unilateral* treatments, we pool these treatments to analyze the case of two decision-makers (Panel B).

In specification I, as independent variables, we include the senders' acceptability rating of choosing the antisocial message ('normative beliefs') and their expected fraction of other senders choosing the antisocial message ('expected antisocial messages'). This first specification allows us to test Hypotheses 3B and 4B. In specification II, we add the interaction term between these two variables to test Hypothesis 6A. This specification is inspired by Bicchieri (2006), who argues that a social norm exists when both normative and empirical expectations coincide. In specification III, we evaluate Hypothesis 5B. Namely, the effect of the senders' belief of the receivers' expected fraction of senders choosing the antisocial message ('belief of receivers' expected antisocial messages'). These beliefs are crucial in models of guilt aversion (Battigalli and Dufwenberg, 2007). In specification IV, we include the interaction between these beliefs and the senders' normative beliefs to test Hypothesis 6B. Finally, in specification V, we analyze the effect of both types of empirical beliefs when they are included simultaneously. As one would expect, these beliefs are highly correlated (Pearson's $r = 0.656$, $p < 0.001$). Hence, we introduce them by adding to specification IV the difference between the senders' expected fraction of antisocial messages and their belief of the receivers' expected fraction of antisocial messages ('difference in empirical beliefs').[21]

[20] With a power of 80% and the number of observations we have in each treatment, the minimal detectable difference in empirical beliefs for the pairwise comparisons ranges from 7.48% to 10.24% (assuming the observed standard deviations). For reference, the difference between *1-Sender* and *2-Sender-Consensus* is 2.20%.

[21] We also conducted regressions substituting the senders' normative beliefs with their expectation of the receivers' normative beliefs. We present the results in Table A3 in Appendix A. We find similar results to the ones reported below.

Table 2. Determinants of choosing the antisocial message

Note: The table presents marginal effects of probit regressions in which the dependent variable equals one if the sender chooses the antisocial message and zero otherwise. ‘Normative beliefs’ are the senders’ acceptability rating of choosing the antisocial message; ‘expected antisocial messages’ are the senders’ expected fraction of other senders choosing the antisocial message; ‘belief of receivers’ expected antisocial messages’ are the senders’ belief of the receivers’ expected fraction of senders choosing the antisocial message; difference in empirical beliefs’ is the difference between the senders’ expected fraction of antisocial messages and their belief of the receivers’ expected fraction of antisocial messages. Robust standard errors are presented in parentheses. ** and * indicate statistical significance at 0.01 and 0.05.

A. *1-Sender* & *Passive-Sender* treatments ($n = 223$)

	I	II	III	IV	V
Normative beliefs	0.01	0.03	0.03	0.09	0.05
	(0.02)	(0.05)	(0.02)	(0.05)	(0.06)
Expected antisocial messages	0.46**	0.45**			
	(0.10)	(0.10)			
Expected antisocial messages		−0.03			
× normative beliefs		(0.08)			
Belief of receivers’ expected antisocial messages			0.31**	0.26**	0.45**
			(0.10)	(0.10)	(0.12)
Belief of receivers’ expected antisocial messages				−0.10	−0.07
× normative beliefs				(0.08)	(0.09)
Difference in empirical beliefs					0.43**
					(0.12)
Difference in empirical beliefs					0.03
× normative beliefs					(0.08)

B. *2-Sender-Consensus* & *2-Sender-Unilateral* treatments ($n = 397$)

	I	II	III	IV	V
Normative beliefs	0.02	0.09*	0.04	0.18**	0.17**
	(0.02)	(0.04)	(0.02)	(0.04)	(0.04)
Expected antisocial messages	0.44**	0.44**			
	(0.09)	(0.09)			
Expected antisocial messages		−0.14*			
× normative beliefs		(0.07)			
Belief of receivers’ expected antisocial messages			0.37**	0.33**	0.46**
			(0.09)	(0.09)	(0.09)
Belief of receivers’ expected antisocial messages				−0.26**	−0.25**
× normative beliefs				(0.07)	(0.07)
Difference in empirical beliefs					0.35**
					(0.12)
Difference in empirical beliefs					0.06
× normative beliefs					(0.11)

In specifications I and II, we find strong support for Hypothesis 4B. In other words, we see a significantly positive association between choosing the antisocial message and the senders' expected fraction of other senders choosing the antisocial message. In specification II, we see an interesting difference between treatments. Unlike in treatments with one active sender, in treatments with two senders, we find a significantly positive association between choosing the antisocial message and the senders' normative beliefs (supporting Hypothesis 3B). Moreover, consistent with Hypothesis 6A, there is a significantly positive association between sending the antisocial message and the interaction between normative beliefs and the senders' expected fraction of other senders choosing the antisocial message. This pattern is consistent with models of social norms, which posit that senders will be motivated to abide with a social norm only when their normative and empirical expectations coincide.

Specifications III and IV provide evidence in favor of Hypothesis 5B. That is to say, the senders' belief of the receivers' expected fraction of antisocial messages is positively associated with choosing the antisocial message in all treatments. In specification IV, we see once again a significantly positive association between choosing the antisocial message and normative beliefs in treatments with two senders but not in treatments with one active sender. In addition, in treatments with two senders, we find support for Hypothesis 6B. Namely, we find a significantly positive association between choosing the antisocial message and the interaction between normative beliefs and the senders' belief of the receivers' expected fraction of antisocial messages. This pattern is consistent with our interpretation of models of guilt aversion, which predict that senders avoid disappointing the receiver only when their guilt-sensitivity is high.

Interestingly, in specification V, we see that both empirical beliefs predict the senders' choice. In other words, the senders' expected fraction of antisocial messages has an additional effect on top of the effect of the senders' belief of the receivers' expected fraction of antisocial messages (simultaneously supporting Hypotheses 4B and 5B). This specification also confirms what appears to be an important difference between the treatments with two senders and those with one active sender. Namely, normative beliefs and their interaction with empirical beliefs are significant determinants of choosing the antisocial message (i.e., Hypotheses 3B and 6B) only when this is a joint decision. Hence, even though the introduction of a joint decision does not impact mean empirical beliefs (Result 3), it does change the relationship between empirical and normative beliefs in determining antisocial behavior. These findings are stated as our last result.

Result 4 *The senders' beliefs of what other senders are doing and their beliefs of the receivers' expectations are critical determinants of the senders' antisocial behavior. The senders' normative beliefs and their interaction with empirical beliefs are also important determinants when senders make decisions jointly.*

6 Lab experiment

In this section, we briefly mention the design and results of the lab experiment. The detailed description of the experimental design is available in Appendix B, the complete analysis of the results in Appendix C, and a sample of the instructions in Appendix D.

The lab and online experiments share many features. For instance, in the lab experiment, we ran sessions for the *1-Sender*, *Passive-Sender*, and *2-Sender-Consensus* treatments using similarly-structured sender-receiver games. The main differences between the two experiments are:

- First, in the lab experiment, we vary the senders' earnings from the antisocial message and use the strategy method to determine the amount of money a sender is willing to forgo to act prosocially. We call this the antisocial premium. This method allows us to have a more precise measure of the impact of making joint decisions on the senders' willingness to act antisocially.

- Second, in the lab experiment, we run treatments with different antisocial messages. Like in the online experiment, in some sessions, the antisocial message contains a lie by claiming to reveal the label of the prosocial option. However, in the lab experiment, we also run sessions where the antisocial message is truthful in that it correctly states that the revealed label corresponds to the antisocial option. These sessions allow us to test whether the results reported here are robust to another context.

- Third, in the lab experiment, we elicit the senders' emotional reaction to sending the antisocial message. Specifically, we ask senders to self-report their experienced guilt when they see the option implemented by the receiver and the earnings of all the players with whom they are matched. As mentioned in Section 4, guilt is an essential emotion to ensure compliance with norms that prescribe prosocial behavior (Baumeister et al., 1994; Hopfensitz and Reuben, 2009). Hence, by analyzing experienced guilt, we have another measure to corroborate that the senders' behavior is explained by their normative beliefs.

- Finally, in the lab experiment, we ran sessions using the *2-Sender-Consensus* aggregation rule but not with the *2-Sender-Unilateral* rule.

The results of the lab experiment are consistent with those of the online experiment. As in the online experiment, we find significantly more antisocial behavior in *2-Sender-Consensus* than in *1-Sender* and *Passive-Sender* (Result 1). On average, senders in *2-Sender-Consensus* require €1.40 less for sending the antisocial message than senders in *1-Sender*. This difference is substantial, considering that the overall mean antisocial premium across treatment is only €3.28. We also find that senders consider choosing the antisocial message more acceptable in *2-Sender-Consensus* than in the other treatments (Result 2), and no significant treatment differences in empirical beliefs (Result 3).

Interestingly, we find the same patterns irrespective of whether the antisocial message is deceitful or truthful. For instance, the difference in antisocial premiums between *2-Sender-Consensus* and *1-Sender* equals €1.33 when antisocial messages are deceitful and €1.49 when they are truthful. Hence, our results also apply to antisocial behaviors that do not include lying.

Finally, the analysis of the senders' experienced guilt shows that senders feel significantly guiltier after sending the antisocial message in *1-Sender* compared to *2-Sender-Consensus*, and this difference is bigger among senders who think that sending the antisocial message is normatively unacceptable. Hence, the effect of a second sender is not only evident in the senders' behavior and normative beliefs but also their emotional reaction.

7 Conclusions

In this study, we present evidence that individuals tend to behave more antisocially when making a joint decision with a partner than when acting alone. We attribute the increased willingness to behave antisocially to a shift in normative beliefs since 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 beliefs.

Our results provide further insights concerning the shift in the senders' normative beliefs. First, we observe a similar shift in the normative beliefs of receivers. This result implies that the shift is not a self-serving reaction by senders. In other words, senders are not using the second sender's presence as an "excuse" to misbehave. Second, the results from our *Passive-Sender* treatment suggest that a necessary condition for the shift in normative beliefs to occur is the active involvement of a partner in the decision-making process. Third, the subjects' empirical beliefs about the fraction of senders choosing the antisocial message are similar across treatments. These results add to the growing evidence that normative and empirical beliefs are independent concepts with distinct effects on behavior even though they are often highly correlated (for more on diverging normative and empirical beliefs see Bicchieri and Xiao, 2009; Bicchieri and Chavez, 2013; Bicchieri et al., 2022). In our experiment, as in most of the literature, normative and empirical beliefs are significantly correlated.[22] However, only normative beliefs are affected by including an actively involved second sender.[23] Taken together, our find-

[22] For example, in all four treatments, the senders' expected fraction of antisocial messages is positively correlated with their normative beliefs (Spearman's rank correlations, $p < 0.001$) and their expectation of the receivers' normative beliefs (Spearman's rank correlations, $p < 0.039$).

[23] We should point out that our experiment is not designed to specifically test the influence of normative beliefs

ings suggest that antisocial behavior increases in groups because antisocial actions become more acceptable and not because acceptable behavior is expected less often.

The analysis of the senders' individual behavior reveals that both normative and empirical beliefs determine antisocial behavior. In all treatments, we find a strong association between the senders' behavior, their beliefs of other senders' antisocial behavior, and their beliefs of the receivers' expected antisocial behavior. In addition, we find that antisocial behavior depends on the interaction of these two empirical beliefs with the senders' normative beliefs. We think that these findings point to two exciting paths for further research. First, the interaction between normative beliefs and beliefs of the receivers' expected antisocial behavior is in line with models of guilt aversion if one interprets individuals' guilt sensitivity as being determined by their normative beliefs. Extending models of guilt aversion to incorporate normative beliefs might allow us to understand why 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). Second, the fact that two types of empirical beliefs predict adherence to normative beliefs suggests that more research is needed to fully understand the relative importance of first-order and second-order empirical beliefs in settings with multiple roles.

It is worth noting that, while we can observe that a shift in normative beliefs occurs, we do not know precisely *why* subjects change their normative beliefs. This is a general weakness of models of social norms, which are typically silent concerning the reasons behind normative evaluations. A common explanation for increased antisocial behavior in groups is that joint decisions diffuse the individuals' responsibility for choosing the antisocial action. However, there is no consensus yet in the literature concerning the definition of "responsibility". Recently, responsibility has been linked to pivotality in determining prosocial options (Falk et al., 2020). However, the lack of difference between the *2-Sender-Consensus* and *2-Sender-Unilateral* treatments suggests that pivotality is not the sole determinant of diffusion of responsibility. Another interpretation is that individuals feel less responsible for acting antisocially when a decision is made in a group simply because the decision-making process includes other individuals, and therefore the decision can be attributed to the group and not to one person. This notion of diffusion of responsibility is compatible with our behavioral results and the shift in normative beliefs in the *2-Sender* treatments.

Lastly, we would like to point out that our findings do not imply that there are no other explanations for increased antisocial behavior in joint decisions. To isolate the effect of normative beliefs, we designed our experiment so that joint decisions are made without interaction. However, as proposed by Falk and Szech (2013) and Kocher et al. (2018), there are various channels

on empirical beliefs or vice-versa. We do not exogenously vary beliefs (as in Bicchieri et al., 2020), and we use a one-shot setting, which prevents us from analyzing how beliefs are updated (see Bicchieri et al., 2022).

through which interaction between decision-makers can lead to more antisocial behavior.[24]

References

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

Aycinena, Diego, F. B. and Kimbrough, E. O. (2021). Measuring Norms: Assessing Normative Expectation Elicitation Methods. NoBeC Seminar Presentation, September 16.

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.

Bašić, Z. and Verrina, E. (2021). Personal norms - and not only social norms - shape economic behavior. Max Planck Institute for Research on Collective Goods Discussion Paper.

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.

Behnk, S., Hao, L., and Reuben, E. (2022). Replication data for: Shifting normative beliefs: On why groups behave more antisocially than individuals. Inter-university Consortium for Political and Social Research, ICPSR-166122-V1.

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.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57(1):289–300.

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

Bicchieri, C. and Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, 23(2):161–178.

[24] For instance, 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; and the specific rules used to aggregate individual preferences (e.g., majority voting) might result in more antisocial decisions.

Bicchieri, C. and Chavez, A. K. (2013). Norm Manipulation, Norm Evasion: Experimental Evidence. *Economics and Philosophy*, 29(2):175–198.

Bicchieri, C., Diemant, E., Gächter, S., and Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, 132:59–72.

Bicchieri, C., Diemant, E., and Sonderegger, S. (2020). It's Not A Lie if You Believe the Norm Does Not Apply: Conditional Norm-Following with Strategic Beliefs. SSRN Working paper 3326146.

Bicchieri, C., Diemant, E., and Xiao, E. (2021). Deviant or wrong? The effects of norm information on the efficacy of punishment. *Journal of Economic Behavior & Organization*, 188:209–235.

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., Horváth, G., and Nitta, K. (2019). Whistleblowing and diffusion of responsibility: An experiment. *European Economic Review*, 119:287–301.

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.

D'Adda, G., Drouvelis, M., and Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, 62:1–7.

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.

Dear, K., Dutton, K., and Fox, E. (2019). Do 'watching eyes' influence antisocial behavior? A systematic review & meta-analysis. *Evolution and Human Behavior*, 40(3):269–280.

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

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*, chapter 3, pages 39–56. Edward Elgar Publishing, Glos, UK.

Falk, A., Neuber, T., and Szech, N. (2020). Diffusion of Being Pivotal and Immoral Outcomes. *The Review of Economic Studies*, 87(5):2205–2229.

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

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. and Föllmi-Heusi, F. (2013). Lies In Disguise-An Experimental Study On Cheating. *Journal of the European Economic Association*, 11(3):525–547.

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. (2018). Shifting Blame? Experimental Evidence of Delegating Communication. *Management Science*, 64(8):3911–3925.

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.

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. (2019). Teams promise but do not deliver. *Games and Economic Behavior*, 117:420–432.

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 analysis of the online experiment

Appendix A contains the regressions reported in the paper but not fully described there due to space constraints.

Table A1 shows the regressions used to evaluate whether the treatment differences in the subjects' normative beliefs are statistically significant. All normative beliefs range from very unacceptable (1) to very acceptable (5). Therefore, we estimate all coefficients using ordered probit regressions with robust standard errors. In column I, the dependent variable is the senders' acceptability rating of sending the antisocial message. In column II, the dependent variable is the senders' belief of the receivers' acceptability rating of sending the antisocial message. In column III, the dependent variable is the receivers' acceptability rating of sending the

Table A1. Treatment differences in normative beliefs in the online experiment

Note: Ordered probit regressions of the subjects' normative beliefs. Robust standard errors in parentheses. ** and * indicate statistical significance at 0.01 and 0.05.

	I	II	III	IV
Passive-Sender	0.09	0.17	0.00	0.18
	(0.15)	(0.17)	(0.14)	(0.19)
Passive-Sender × sender B	−0.03	0.00		−0.06
	(0.15)	(0.17)		(0.18)
2-Sender-Consensus	0.44**	0.44**	0.33*	0.17
	(0.13)	(0.14)	(0.14)	(0.16)
2-Sender-Unilateral	0.48**	0.54**	0.27	0.07
	(0.13)	(0.14)	(0.15)	(0.17)
Observations	734	734	423	734
χ^2	26.37	21.37	9.17	1.62

Table A2. Treatment differences in empirical expectations in the online experiment

Note: Tobit regressions of the subjects' expected fraction of senders choosing the antisocial message. Robust standard errors in parentheses. ** and * indicate statistical significance at 0.01 and 0.05.

	I	II	III
Passive-Sender	0.02	−0.02	−0.02
	(0.04)	(0.04)	(0.04)
Passive-Sender × sender B	−0.03	0.01	
	(0.04)	(0.04)	
2-Sender-Consensus	−0.02	0.00	−0.08
	(0.03)	(0.04)	(0.04)
2-Sender-Unilateral	−0.02	0.03	−0.02
	(0.04)	(0.04)	(0.04)
Constant	0.45^{**}	0.50^{**}	0.48^{**}
	(0.03)	(0.03)	(0.03)
Observations	734	734	423
F-stat	0.51	0.74	1.35

antisocial message. Finally, in column IV, the dependent variables is the senders' acceptability rating of sending the prosocial message. The dependent variables consist of dummy variables indicating the *2-Sender-Consensus*, *2-Sender-Unilateral*, and *Passive-Sender* treatments and an interaction between *Passive-Sender* and being the passive sender (sender B).

Table A2 shows the regressions used to evaluate whether the treatment differences in the subjects' empirical expectations are statistically significant. We use Tobit regressions with robust standard errors to test for differences across treatments as expectations are censored at 0% and 100%. In column I, the dependent variable is the senders' expected fraction of other senders choosing the antisocial message. In column II, the dependent variable is the senders' belief of the receivers' expected fraction of senders choosing the antisocial message. Finally, in column III, the dependent variable is the receivers' expected fraction of senders choosing the antisocial message. The dependent variables consist of dummy variables indicating the *2-Sender-Consensus*, *2-Sender-Unilateral*, and *Passive-Sender* treatments and an interaction between *Passive-Sender* and being the passive sender (sender B).

Table A3 reports marginal effects from the same probit regressions seen in Table 2 but substituting the senders' normative beliefs with their expectation of the receivers' normative beliefs. In all regressions, the dependent variable equals one if a sender chooses the antisocial message and zero if the sender chooses the prosocial message. We pool senders from the *1-Sender* and *Passive-Sender* treatments and run separate regressions for senders in the *2-Sender-Consensus* and *2-Sender-Unilateral* treatments. Specification I includes the senders' belief of the receivers' acceptability rating of choosing the antisocial message ('expected normative beliefs')

Table A3. Determinants of choosing the antisocial message in the online experiment

Note: The table presents marginal effects of probit regressions in which the dependent variable equals one if the sender chooses the antisocial message and zero otherwise. 'Expected normative beliefs' are the senders' belief of the receivers' acceptability rating of choosing the antisocial message; 'expected antisocial messages' are the senders' expected fraction of other senders choosing the antisocial message; 'belief of receivers' expected antisocial messages' are the senders' belief of the receivers' expected fraction of senders choosing the antisocial message; difference in empirical beliefs' is the difference between the senders' expected fraction of antisocial messages and their belief of the receivers' expected fraction of antisocial messages. Robust standard errors are presented in parentheses. ** and * indicate statistical significance at 0.01 and 0.05.

A. *1-Sender* & *Passive-Sender* treatments ($n = 223$)

	I	II	III	IV	V
Expected normative beliefs	−0.01	0.02	0.00	0.01	0.02
	(0.02)	(0.05)	(0.02)	(0.05)	(0.05)
Expected antisocial messages	0.49**	0.46**			
	(0.09)	(0.11)			
Expected antisocial messages		−0.04			
× expected normative beliefs		(0.07)			
Belief of receivers' expected antisocial messages			0.34**	0.34**	0.48**
			(0.10)	(0.11)	(0.12)
Belief of receivers' expected antisocial messages				−0.01	−0.05
× expected normative beliefs				(0.07)	(0.08)
Difference in empirical beliefs					0.42**
					(0.13)
Difference in empirical beliefs					−0.03
× expected normative beliefs					(0.08)

B. *2-Sender-Consensus* & *2-Sender-Unilateral* treatments ($n = 397$)

	I	II	III	IV	V
Expected normative beliefs	−0.02	0.06	−0.01	0.04	0.05
	(0.02)	(0.03)	(0.02)	(0.04)	(0.04)
Expected antisocial messages	0.50**	0.44**			
	(0.09)	(0.09)			
Expected antisocial messages		−0.17*			
× expected normative beliefs		(0.06)			
Belief of receivers' expected antisocial messages			0.42**	0.36**	0.50**
			(0.09)	(0.10)	(0.10)
Belief of receivers' expected antisocial messages				−0.08	−0.15*
× expected normative beliefs				(0.07)	(0.07)
Difference in empirical beliefs					0.36**
					(0.12)
Difference in empirical beliefs					−0.17*
× expected normative beliefs					(0.08)

and their expected fraction of other senders choosing the antisocial message ('expected antisocial messages'). Specification II adds the interaction between these two variables. Specification III includes senders' expected normative beliefs and their belief of the receivers' expected fraction of senders choosing the antisocial message ('belief of receivers' expected antisocial messages'). Specification IV adds the interaction between the latter two variables. Finally, specification V adds to specification IV the difference between the senders' expected fraction of antisocial messages and their belief of the receivers' expected fraction of antisocial messages ('difference in empirical beliefs').

Appendix B Design and procedures of the lab experiment

The lab experiment contains similar design elements to the online experiment. Specifically, in both experiments, we use sender-receiver games that vary the number of senders and the number of senders involved in the decision. In the *1-Sender* treatment, one sender chooses the message sent to the receiver. In the *2-Sender-Consensus* treatment, two senders jointly choose the message. Finally, there are two senders in the *Passive-Sender* treatment, but only one of them chooses the message. Next, we describe in more detail the lab experiment.

In the *1-Sender* treatment, there is one sender and one receiver. The receiver's task is to choose one out of ten options to determine her and the sender's earnings. There is one prosocial option that pays €10 to each player, 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. As in the online experiment, each of the ten options is randomly labeled with a single letter ranging from A to J. The sender knows how each option is labeled, but the receiver does not.

In the *2-Sender-Consensus* and the *Passive-Sender* treatments, there are two senders (sender A and sender B) and one receiver. The payoff structure for sender A and the receiver are identical to those in the *1-Sender* treatment. 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.

In all treatments, the only information available to the receiver is due to a message. In *1-Sender*, the sender chooses which message is sent. In *2-Sender-Consensus*, the two senders jointly make this choice. In *Passive-Sender*, sender A chooses. There are two available messages. As in the online experiment, the prosocial message identifies the prosocial option. The antisocial message is one of two types. In the *Deception* sessions, the antisocial message points to the antisocial option but claims it is the prosocial option (as in the online experiment). In the *Bitter-pill* sessions, the antisocial message identifies the antisocial option. It is common knowledge that a message always reveals the label of either the prosocial or the antisocial option.

We use the strategy method to measure the senders' willingness to send an antisocial mes-

Table B1. Senders' choice lists in the lab experiment

Note: Message I corresponds to the prosocial message and Message II to the antisocial message. Each row displays the value of x and the sender's earnings (in euros) of each message (if implemented).

List A

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

List B

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

sage. Specifically, senders choose between the prosocial and antisocial messages for 14 different values of x. The rows of List A in Table B1 correspond to the choices of senders in *1-Sender*, senders A in *2-Sender-Consensus*, and senders A in *Passive-Sender*. The rows of List B correspond to the choices of senders B in *2-Sender-Consensus*. After senders make their choices, one row is randomly selected to determine which message is sent (receivers are not informed which row is selected). As in the online experiment, in *2-Sender-Consensus*, the antisocial message is sent only if both senders choose it; otherwise, the prosocial message is sent.

Once a message is chosen, it is displayed on the senders' screen. Senders in *1-Sender* and senders B in *2-Sender-Consensus* and *Passive-Sender* write the message on a piece of paper and then are guided by an experimenter to their receiver's desk. Senders give the paper to receivers and then return to their seat. The experimenter ensures there is no other communication between subjects. Everyone is informed about the delivery process in the instructions. Finally, receivers type in the message they receive before they choose one of the ten options.

B.1 The antisocial premium

We call the minimum monetary compensation a sender must receive to send the antisocial message that sender's antisocial premium. We classify senders who switch messages at a given

x as having an antisocial premium in the interval [€$x - 0.5$, €x]. At the extremes, we classify senders who always choose the prosocial message as having an antisocial premium in the interval [€$7.50, \infty$) if they played in *1-Sender* or as sender A in *2-Sender-Consensus* or *Passive-Sender* and in the interval [€$7.00, \infty$) if they played as sender B in *2-Sender-Consensus*. The analogous intervals for senders who always choose the antisocial message are ($-\infty$, €0.50] and ($-\infty$, €0.00]. We do not classify senders who switched more than once or switched in the wrong direction.

B.2 Normative beliefs

To measure normative beliefs, we use the same method as in the online experiment to elicit the senders' and receivers' acceptability ratings of sending the antisocial message and the prosocial message. We also elicit the senders' belief of the receivers' acceptability ratings. Unlike in the online experiment, senders answer these questions after they deliver the message, which means that, in *2-Sender-Consensus* and *Passive-Sender*, they have some information about one of the other sender's choices.

B.3 Empirical beliefs

As in the online experiment, we elicit the following empirical beliefs:

- The receivers' expected fraction of senders delivering the antisocial message. Receivers earn €0.75 if their answer is correct.

- The senders' belief about the receivers' expected fraction of senders delivering the antisocial message. Senders earn €0.25 if their answer is correct.

- The senders' expected fraction of receivers following the message. Senders earn €0.25 if their answer is correct.

B.4 Guilt

We measure the senders' experienced guilt when they see the option implemented by the receiver and the earnings of all the players with whom they are matched. We ask senders to self-report the intensity at which they experienced guilt on a 7-point Likert scale that ranged from "not at all" (1) to "very intensively" (7). Although guilt is the emotion of interest in our study, we also measured shame, anger, happiness, and gratitude to minimize experimenter demand effects. Evidence that guilt was not excessively salient is that 90% of senders report experiencing at least one emotion with a strictly higher intensity than guilt. We use self-reported measures because, to the best of our knowledge, there are no precise physiological measures of guilt (Adolphs, 2002). 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).

B.5 Procedures

We ran the lab experiment between February and March 2015 at the Laboratory of Experimental Economics (LEE) at University Jaume I in Castellon, Spain. A total of 263 undergraduate students from different faculties participated. We conducted ten sessions, each lasting around 90 minutes.

Upon arrival, the subjects were randomly assigned to desks. After that, the experimenter read aloud the instructions, which are available in Appendix D. Once the experiment ended, subjects were paid in cash. Average earnings were around €15, including a €5 show-up fee.

Appendix C Results of the lab experiment

A total of 140 subjects participated as active senders in the lab experiment. Of all the senders, 8 senders switched more than once and 1 sender switched from the antisocial to the prosocial message as the premium for the antisocial message increased. Since it is not clear what the antisocial premium of these subjects is, we excluded them from the statistical analysis. This exclusion leaves us with 39 senders in the *1-Sender* treatment (19 in *Bitter-pill* and 20 in *Deception*), 71 senders in the *2-Sender-Consensus* treatment (35 in *Bitter-pill* and 36 in *Deception*), and 22 sender A's in the *Passive-Sender* treatment (all of them in *Deception*).

C.1 The antisocial premium

Figure C1 plots the cumulative distributions of the senders' antisocial premiums in *1-Sender* and *2-Sender-Consensus*, pooling the *Bitter-pill* and *Deception* sessions. The first row of Table C1 shows the mean antisocial premium depending on the number of senders and the type of antisocial message. The figures show that many senders are willing to forego profits to act prosocially. More interestingly, having a second sender lowers antisocial premiums. On average, senders in *2-Sender-Consensus* require €1.40 less for sending the antisocial message than senders in *1-Sender* (€1.49 less in *Bitter-pill* and €1.33 less in *Deception*). This difference is substantial, considering that the overall mean antisocial premium across treatment 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 the prosocial to the antisocial message when the latter pays more than €x, then we know that her antisocial premium lies in the interval $[€x - 0.50, €x]$ (Stewart, 1983). At the extremes, senders who always choose the prosocial message have an antisocial premium in the interval $[€7.50, \infty)$ if they played in *1-Sender* or as

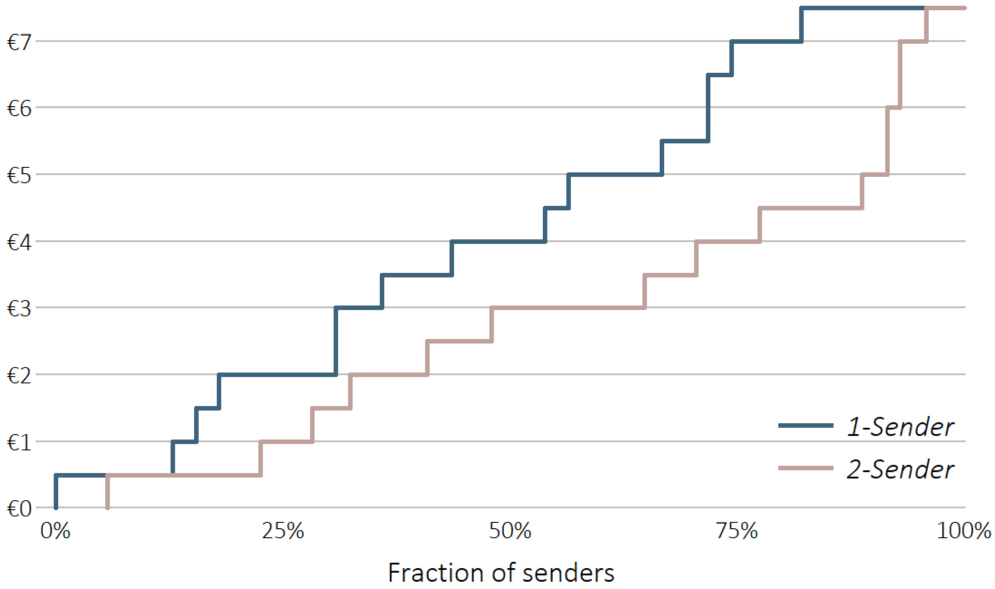


Figure C1. Cumulative distributions of senders' antisocial premiums in the *1-Sender* and *2-Sender-Consensus* treatments in the lab experiment

sender A in *2-Sender-Consensus*. If they played as sender B in *2-Sender-Consensus*, then their antisocial premium is in the interval $[€7.00, \infty)$. For senders who always choose the antisocial message, we classified them as having an antisocial premium in the interval $(-\infty, €0.50]$ if they played in *1-Sender* or as sender A in *2-Sender-Consensus*, or in the interval $(-\infty, €0.00]$ if they played as sender B in *2-Sender-Consensus*. All regressions are estimated using robust standard errors and are found in Table C2. The regressions in columns I and II use data from *1-Sender* and *2-Sender-Consensus*. Column III further includes the data from *Passive-Sender*.

Consistent with the results of the online experiment, we find that antisocial premiums are significantly lower in *2-Sender-Consensus* compared to *1-Sender* ($p = 0.005$ overall; $p = 0.031$ in *Bitter-pill*; $p = 0.045$ in *Deception*). A difference-in-differences test reveals that the difference between *1-Sender* and *2-Sender-Consensus* does not differ between *Bitter-pill* and *Deception* ($p = 0.993$). The difference in antisocial premiums between *Bitter-pill* and *Deception* is close to statistical significance in *1-Sender* ($p = 0.074$) and is significant in *2-Sender-Consensus* ($p = 0.003$).

C.2 Normative beliefs

Next, we test the effect of having a second sender on the subjects' normative beliefs. Table C1 presents the senders' mean acceptability ratings of sending the antisocial message, the senders' mean belief of the receivers' acceptability ratings, and the receivers' mean acceptability ratings. Given that normative beliefs are discrete, ranging from very unacceptable (1) to very acceptable (5), we use ordered probit regressions to test whether treatment differences are statistically

Table C1. Means and standard deviations of selected variables in the lab experiment

	Overall		***Bitter-pill***		***Deception***	
	1-Sender	*2-Sender-Consensus*	*1-Sender*	*2-Sender-Consensus*	*1-Sender*	*2-Sender-Consensus*
	Antisocial premium					
Antisocial premium	4.18	2.78	4.97	3.49	3.42	2.10
	(2.44)	(1.97)	(2.15)	(2.05)	(2.50)	(1.64)
	Normative beliefs of sending the antisocial message					
Senders' normative beliefs	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' expectation of the	1.62	2.42	1.74	2.43	1.50	2.42
receivers' normative beliefs	(1.09)	(1.56)	(1.28)	(1.54)	(0.89)	(1.61)
Receivers' normative beliefs	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 the antisocial message					
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)

significant. The regression coefficients are provided in Table C3. Since the senders' normative beliefs were elicited after the message delivery, we cluster standard errors in *2-Sender-Consensus* on the matched pairs. In columns I through III, the dependent variable is the senders' normative beliefs regarding the acceptability of sending the antisocial message. In columns IV through VI, the dependent variable is the senders' expectations of the receivers' normative beliefs regarding the acceptability of sending the antisocial message. Lastly, in columns VII and VIII, the dependent variable is the receivers' normative beliefs regarding the acceptability of sending the antisocial message. In columns III and VI, we control for subjects' experience up to the point where they reported their normative beliefs. To be precise, we include the following control variables: a dummy variable that equals one if the other sender in *2-Sender-Consensus* chose the antisocial message, a dummy variable that equals one if the message sent to the receiver was the antisocial message, and the sender's earnings if the receiver follows the message. All regressions include data from *1-Sender* and *2-Sender-Consensus.*

The results are in line with those of the online experiment. We find that senders in *2-Sender-Consensus* think it is more acceptable to send the antisocial message than senders in *1-Sender* (2.97 vs. 2.33, $p = 0.008$). Interestingly, the receivers' acceptability ratings are also higher in *2-Sender-Consensus* (3.08 vs. 2.60, $p = 0.102$), as are the senders' beliefs of the receivers' acceptability ratings (2.42 vs. 1.62, $p = 0.007$). This pattern persists when we examine *Bitter-*

Table C2. Treatment differences in antisocial premiums in the lab experiment

Note: Interval regressions of the senders' antisocial premium. Robust standard errors in parentheses. ** and * indicate statistical significance at 0.01 and 0.05.

	I	II	III
2-Sender-Consensus	-1.58^{**}		
	(0.57)		
2-Sender-Consensus × *Bitter-pill*		-1.57^{**}	-1.57^{**}
		(0.74)	(0.73)
2-Sender-Consensus × *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

pill and *Deception* separately. In *Bitter-pill*, the difference in acceptability ratings between *1-Sender* and *2-Sender-Consensus* is 0.69 for senders ($p = 0.021$), 0.69 for the senders' belief of the receivers' acceptability ratings ($p = 0.126$), and 0.35 for receivers ($p = 0.380$). In *Deception*, the difference between *1-Sender* and *2-Sender-Consensus* is 0.59 for senders ($p = 0.100$), 0.92 for the senders' beliefs of the receivers' acceptability ratings ($p = 0.020$) and 0.63 for receivers ($p = 0.116$). Hence, all three measures of the subjects' normative beliefs suggest it is more acceptable to send the antisocial message in *2-Sender-Consensus* compared to *1-Sender.* Finally, note that the control variables in columns III and VI are neither jointly significant in column III ($p = 0.844$) nor column VI ($p = 0.837$).

C.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 beliefs across treatments. Table C4 provides the means and standard deviations of the senders' self-reported emotions dependent on whether the prosocial or the antisocial message was sent to the receiver. In *1-Sender*, 22 prosocial messages and 17 antisocial messages were delivered; while in *2-Sender-Consensus*, 27 prosocial messages and 11 antisocial messages 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 results

Table C3. Treatment differences in normative beliefs in the lab experiment

Note: Ordered probit regressions of the subjects' normative beliefs. Robust standard errors clustered on matched pairs (for senders in *2-Sender-Consensus*) in parentheses. ** and * indicate statistical significance at 0.01 and 0.05.

	Senders'			Senders'			Receivers'	
	beliefs			expected beliefs			beliefs	
	I	II	III	IV	V	VI	VII	VIII
2-Sender-Consensus	0.62**			0.65**			0.39	
	(0.23)			(0.24)			(0.24)	
2-Sender-Consensus × *Bitter-pill*		0.70*	0.75*		0.54	0.48		0.30
		(0.30)	(0.33)		(0.35)	(0.40)		(0.34)
2-Sender-Consensus × *Deception*		0.56	0.59		0.76*	0.68		0.53
		(0.34)	(0.37)		(0.33)	(0.39)		(0.34)
Deception		0.44	0.50		−0.25	−0.20		0.50
		(0.36)	(0.37)		(0.37)	(0.38)		(0.30)
Other sender chose the antisocial message			−0.05			0.10		
			(0.29)			(0.30)		
The antisocial message was sent			0.40			0.19		
			(0.62)			(0.60)		
Earnings if message is followed			0.08			0.06		
			(0.09)			(0.09)		
Observations	110	110	110	110	110	110	79	79
Clusters	77	77	77	77	77	77	79	79
χ^2	6.94	9.08	12.16	7.23	8.17	10.69	2.67	11.76

remain unchanged if we include these observations. 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. 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. The analysis of the other emotions is available upon request. By and large, they are in line with the results for guilt.

On average, senders experience more guilt after sending the antisocial message than after sending the prosocial message. The difference is substantial: 4.23 vs. 1.09 in *1-Sender* and 2.68 vs. 1.62 in *2-Sender-Consensus* (t-tests, $p < 0.004$). More importantly, we also find that senders experience significantly less guilt after sending the antisocial message in *2-Sender-Consensus* compared to *1-Sender* (2.68 vs. 4.23; t-test $p = 0.033$). Hence, the effect of a second sender is not only evident in the senders' behavior and normative beliefs but also in their emotional state. A possible concern may be that these differences are due to using a self-reported measure of guilt. 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

Table C4. Means and standard deviations of the senders' emotions in the lab experiment

	1-Sender		***2-Sender-Consensus***		***Passive-Sender***	
Message	Prosocial	Antisocial	Prosocial	Antisocial	Prosocial	Antisocial
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)

is an unlikely explanation for the treatment differences. That is, it is hard to see how subjects could anticipate that the 'expected' emotional reaction is more guilt for the antisocial message in *1-Sender* than in *2-Sender-Consensus* when subjects took part in only one treatment.

In Table C5, we analyze the association between the senders' guilt and their normative beliefs. We use linear regressions with robust standard errors clustered on matched pairs, and the senders' experienced guilt as the dependent variable. Regressions I and II use data from *1-Sender*. In regressions III and IV, we use data from *2-Sender-Consensus*. 'Delivered the prosocial message' and 'Delivered the antisocial message' are dummy variables indicating the message that was delivered to and followed by the receiver; 'normative beliefs' are the senders' normative beliefs of sending the antisocial message; 'Belief of receivers' expected antisocial messages' is the senders' belief of the receivers' expected probability of receiving the antisocial message. In regressions II and IV, we add interaction effects of the message sent and, respectively, normative beliefs and the senders' belief of receivers' expected fraction of antisocial messages. We find that senders who deliver the antisocial message experience more guilt the more they consider that sending the antisocial message is normatively unacceptable. Interestingly, this effect is stronger in *1-Sender* compared to *2-Sender-Consensus*.

C.4 Empirical beliefs

Now, we turn to senders' belief of the receiver's expected probability of receiving the antisocial message. Table C1 shows the senders' beliefs and the receivers' actual expected probability of receiving the antisocial message. We use Tobit regressions to test for treatment differences as belief responses are censored at 0% and 100%. The regression coefficients are provided in Table C6. In columns I and II, the dependent variable is the senders' belief of the receivers' expected

Table C5. Determinants of experienced guilt in the lab experiment

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

	1-Sender		***2-Sender-Consensus***	
	I	II	III	IV
Delivered the antisocial message	2.99**	4.45	0.97*	1.88
	(0.71)	(2.34)	(0.39)	(1.23)
Delivered the prosocial message × normative beliefs		−0.02		0.02
		(0.12)		(0.13)
Delivered the antisocial message × normative beliefs		−1.15**		−0.22
		(0.24)		(0.18)
Delivered the prosocial message		0.08		0.81
× belief of receivers' expected antisocial messages		(0.34)		(0.52)
Delivered the antisocial message		2.00		0.25
× belief of receivers' expected antisocial messages		(2.35)		(1.24)
Deception	0.49	0.28	0.19	0.17
	(0.53)	(0.56)	(0.37)	(0.41)
Constant	0.87**	0.98**	1.55**	1.11**
	(0.27)	(0.12)	(0.20)	(0.36)
Observations	35	35	69	69
Clusters	35	35	37	37
F-statistic	12.71	18.68	7.14	6.65

probability of receiving the antisocial message. In columns III and IV, the dependent variable is the receivers' expected probability of receiving the antisocial message. All regressions include data from *1-Sender* and *2-Sender-Consensus*. As before, we cluster standard errors on the matched sender pairs in *2-Sender-Consensus*.

On average, senders think that receivers expect to receive the antisocial message with probability 0.57 in *1-Sender* and 0.56 in *2-Sender-Consensus* ($p = 0.628$). The senders' beliefs are reasonably accurate as receivers expect to receive the antisocial message with probability 0.56 in *1-Sender* and 0.50 in *2-Sender-Consensus* ($p = 0.835$). Hence, as in the online experiment, we do not find evidence that senders' belief of the receiver's expected probability of receiving the antisocial message are affected by the involvement of a second sender. In an unreported regression, we look at *Bitter-pill* and *Deception* separately. We find that the senders' belief of the receiver's expected probability of receiving the antisocial message in *2-Sender-Consensus* are not significantly higher in *Bitter-pill* ($p = 0.683$) or *Deception* ($p = 0.513$) than in *1-Sender*.

Table C6. Treatment differences in empirical beliefs in the lab experiment

Note: Tobit regressions of the senders' belief of the receivers' expected probability of receiving the antisocial message and the receivers' expected probability of receiving the antisocial message. Robust standard errors clustered on matched pairs (in *2-Sender-Consensus*) in parentheses. ** and * indicate statistical significance at 0.01 and 0.05.

	I	II	III	IV
2-Sender-Consensus	−0.03		−0.07	
	(0.08)		(0.07)	
2-Sender-Consensus × *Bitter-pill*		−0.05		-0.27^{**}
		(0.10)		(0.10)
2-Sender-Consensus × *Deception*		0.00		0.13
		(0.12)		(0.09)
Deception		0.11		−0.03
		(0.13)		(0.10)
Constant	0.59^{**}	0.54^{**}	0.56^{**}	0.58^{**}
	(0.07)	(0.09)	(0.05)	(0.07)
Observations	110	110	79	79
Clusters	77	77	79	79
F-statistic	0.11	1.51	0.96	6.31

C.5 Determinants of the antisocial premium

In Table C7, we conduct a series of interval regressions of the senders' antisocial premiums. In specification I, we include the senders' normative beliefs, their belief of the receiver's expected probability of receiving the antisocial message, and a dummy variable indicating whether it is a *Deception* or *Bitter-pill* session. Inspired by models of guilt aversion, in specification II, we add the interaction term of senders' normative beliefs with their belief of the receiver's expected probability of receiving the antisocial message. 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 *2-Sender-Consensus.* We standardized the control variables so that the constant is comparable across specifications II and III. In all regressions, we cluster standard errors on matched sender pairs in *2-Sender-Consensus.*

We see a similar pattern across *1-Sender* and *2-Sender-Consensus* that is broadly consistent with those seen in the online experiment. First, normative beliefs and beliefs of the receivers' expectations both have a negative effect on antisocial premiums. Second, the interaction between normative beliefs and beliefs of the receivers' expectations is positive, but it is statistically significant only in *2-Sender-Consensus.* Hence, beliefs of the receivers' expectations have a stronger effect on the behavior of senders who think sending the antisocial message is unacceptable compared to senders who think that sending the antisocial message is acceptable. We

Table C7. Determinants of antisocial premiums in the lab experiment

Note: Interval regressions of the senders' antisocial premium. Robust standard errors clustered on matched pairs (in *2-Sender-Consensus*) in parentheses. ** and * indicate statistical significance at 0.01 and 0.05.

	1-Sender			***2-Sender-Consensus***		
	I	II	III	I	II	III
Normative beliefs	-0.39	-0.98	-1.03	-0.20	-1.06^{**}	-1.00^{*}
	(0.33)	(0.50)	(0.55)	(0.25)	(0.39)	(0.39)
Belief of receivers' expected antisocial messages	-6.32^{**}	-9.15^{**}	-9.39^{**}	-3.01^{**}	-7.25^{**}	-7.50^{**}
	(1.48)	(2.68)	(3.02)	(0.91)	(1.80)	(1.83)
Belief of receivers' expected antisocial messages × normative beliefs		1.13	1.20		1.44^{**}	1.31^{*}
		(0.93)	(1.01)		(0.53)	(0.53)
Deception session	-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

also conducted regressions substituting the senders' normative beliefs with their expectation of the receivers' normative beliefs. We present the results in Table C8. We find that the general pattern is similar to the one reported above. However, the interaction between normative beliefs and beliefs of the receivers' expectations is even weaker in *1-Sender*.

C.6 Active participation

Like in the online experiment, we also ran additional sessions using a *Passive-Sender* treatment. In this *Passive-Sender* treatment, sender B is present, delivers the message, and receives the same payoffs as in the *2-Sender-Consensus* treatment of the lab experiment, but has no say on the content of the message. The message is picked by sender A using the same procedure as in the *1-Sender* treatment.

Table C9 shows the mean antisocial premium in *Passive-Sender*. Like before, we use interval regressions to evaluate statistical significance. We find that antisocial premiums in *Passive-Sender* are close to those in *1-Sender* (€3.95 vs. €3.43 on average; $p = 0.558$) and significantly higher than antisocial premiums in *2-Sender-Consensus* (€3.95 vs. €2.10 on average; $p = 0.004$). We observe a similar pattern when we compare normative beliefs across the three treatments. Table C9 also shows the mean acceptability ratings of sending the antisocial message and the senders' belief of the receivers' acceptability ratings. By and large, we see that the senders'

Table C8. Determinants of antisocial premiums in the lab experiment (other specifications)

Note: Interval regressions of the senders' antisocial premium. Robust standard errors clustered on matched pairs (in *2-Sender-Consensus*) in parentheses. ** and * indicate statistical significance at 0.01 and 0.05.

	1-Sender			***2-Sender-Consensus***		
	I	II	III	I	II	III
Expected normative beliefs	−0.10	0.95	0.91	−0.05	−0.95**	−0.88**
	(0.28)	(0.88)	(1.04)	(0.16)	(0.28)	(0.31)
Belief of receivers' expected antisocial messages	−6.36**	−3.40	−3.26	−3.11**	−7.29**	−7.55**
	(1.59)	(3.30)	(2.88)	(0.95)	(1.37)	(1.54)
Belief of receivers' expected antisocial messages		−2.24	−2.34		1.68**	1.53**
× expected normative beliefs		(1.84)	(1.98)		(0.39)	(0.43)
Deception	−1.16	−1.23	−1.39	−1.15*	−1.14**	−0.93**
	(0.76)	(0.74)	(0.74)	(0.45)	(0.42)	(0.36)
Constant	8.37**	6.94**	7.14**	4.87**	7.23**	9.72**
	(1.08)	(1.73)	(1.77)	(0.82)	(1.03)	(1.39)
Additional controls	No	No	Yes	No	No	Yes
Observations	39	39	39	71	71	71
χ^2	23.42	37.59	26.95	27.15	48.61	63.72

acceptability ratings of the active and passive senders in *Passive-Sender* are similar to those of senders in *1-Sender* and below those of senders in *2-Sender-Consensus*. Once again, we use ordered probit regressions to evaluate whether differences are statistically significant. There are no statistical differences between senders in *1-Sender* and senders A in *Passive-Sender* (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 *Passive-Sender* (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 *2-Sender-Consensus* tend to view sending the antisocial message to be more acceptable than senders A in *Passive-Sender* (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 *Passive-Sender* (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 *Passive-Sender* are similar to those of receivers in *1-Sender* (2.73 vs. 2.90, $p = 0.337$) but not for senders in *2-Sender-Consensus* (2.73 vs. 3.53, $p = 0.048$). We also find higher levels of experienced guilt if the receiver follows the antisocial message in *Passive-Sender* compared to *2-Sender-Consensus*, and we do not see noticeable differences in empirical beliefs. These regressions are available upon request.

Table C9. Means and standard deviations for *Passive-Sender* in the lab experiment

	Sender A	*Sender B*	*Receiver*
Antisocial premium	3.95		
	(2.34)		
Normative beliefs of sending the antisocial message	2.57	2.23	2.73
	(1.40)	(1.15)	(1.24)
Expectation of the receivers' normative beliefs	1.67	1.91	
	(1.02)	(1.34)	
Belief that the antisocial message is sent			0.54
			(0.29)
Expectation of the receivers' belief that the antisocial	0.62	0.65	
message is sent	(0.31)	(0.27)	

Appendix D Instructions

This appendix contains a sample of the instructions used in the two experiments. Specifically, we provide the instructions from the *2-Sender-Consensus* treatment using *deceptive* antisocial messages. The instructions used in the *Bitter-pill* sessions and other treatments are almost identical and are available from the authors upon request. Section D.1 contains the instructions for the lab experiment and Section D.2 for the online experiment.

D.1 Instructions for the lab experiment

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"

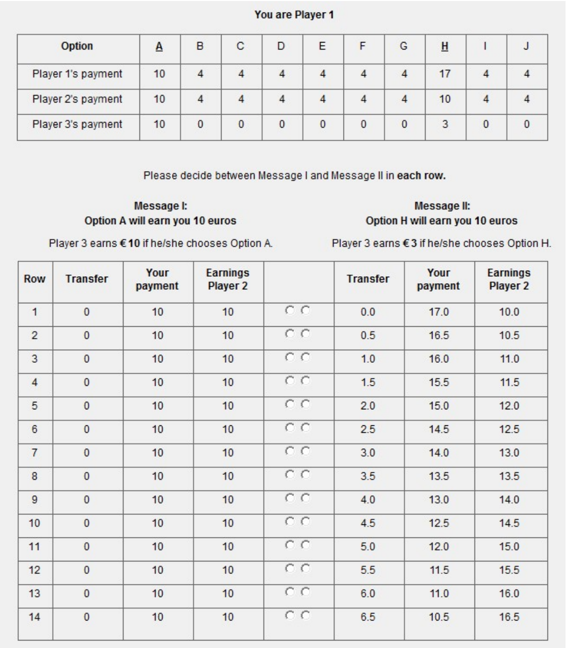
Player 1 cannot deliver the message to Player 3. Only Player 2 is able to deliver the message for him/her. If the option mentioned in the message coincides with the option subsequently chosen by Player 3, then Player 1 transfers between €0 and €6.50 to Player 2 for delivery. The screens below will be 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.*

Decisions of Player 1



You are Player 1

Option	<u>A</u>	B	C	D	E	F	G	<u>H</u>	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

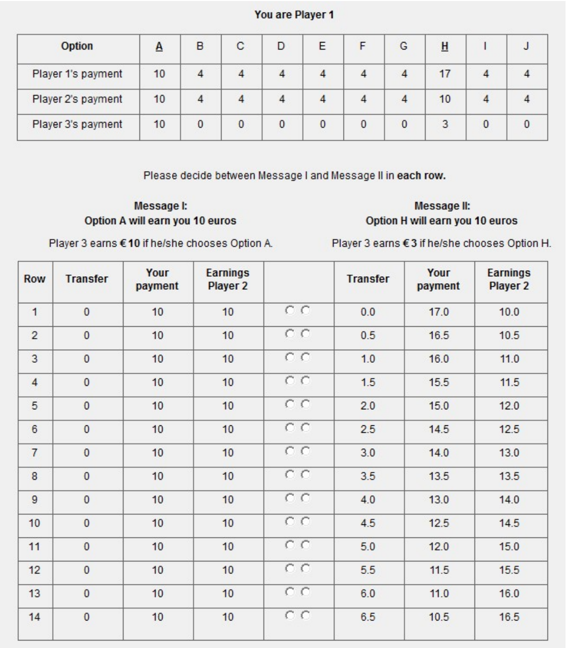
Player 3 earns **€ 10** if he/she chooses Option A.

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	<u>A</u>	B	C	D	E	F	G	<u>H</u>	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

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. 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.

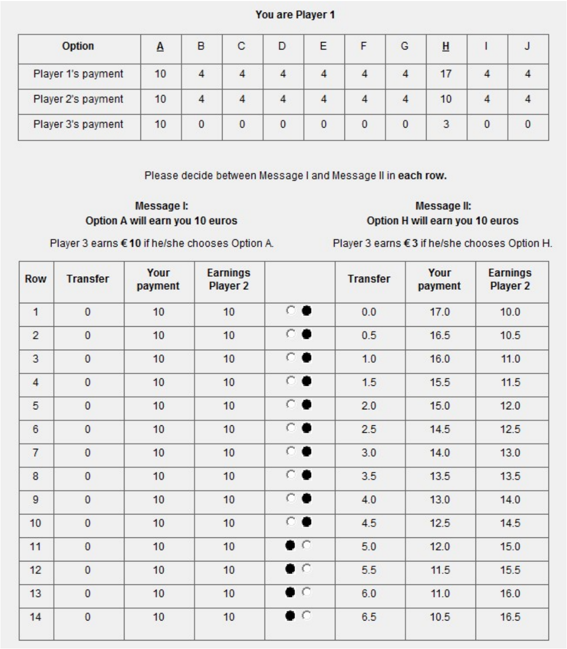
To summarize, Message II is delivered to Player 3 only when both Player 1 and Player 2 choose Message II in the selected row; otherwise Message I is delivered.

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

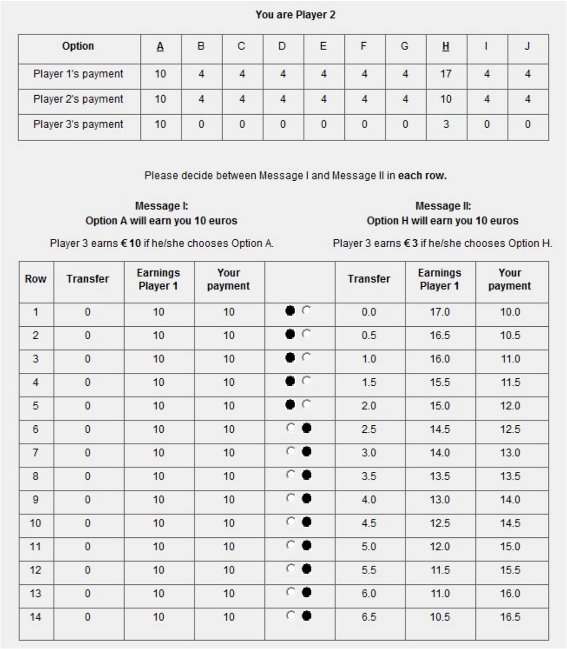
Player 3 earns **€10** if he/she chooses Option A.

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 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

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	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:

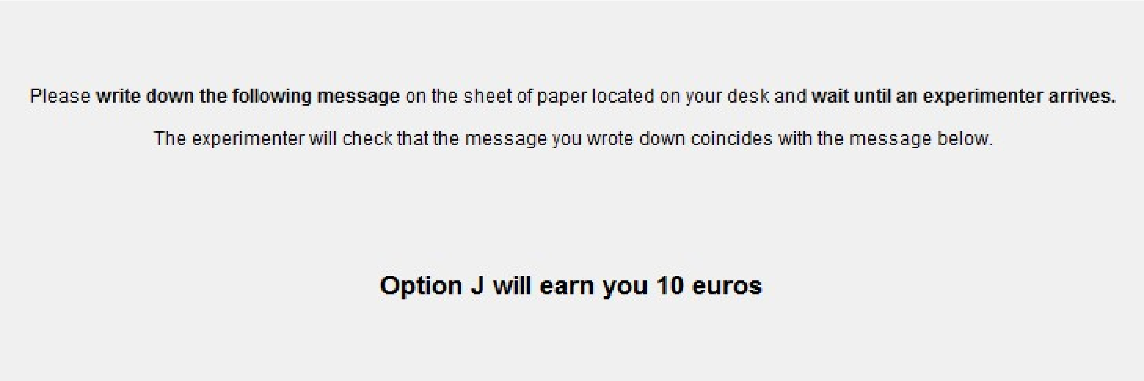
- *Row 4:* Since Player 1 chose Message II but Player 2 disagreed in favor of Message I, then Player 2 delivers Message I. Thereafter, if Player 3 chooses the option corresponding to Message I, then Player 1, Player 2, and Player 3 all earn €10.

- *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.



Please **write down the following message** on the sheet of paper located on your desk and **wait until an experimenter arrives.**

The experimenter will check that the message you wrote down coincides with the message below.

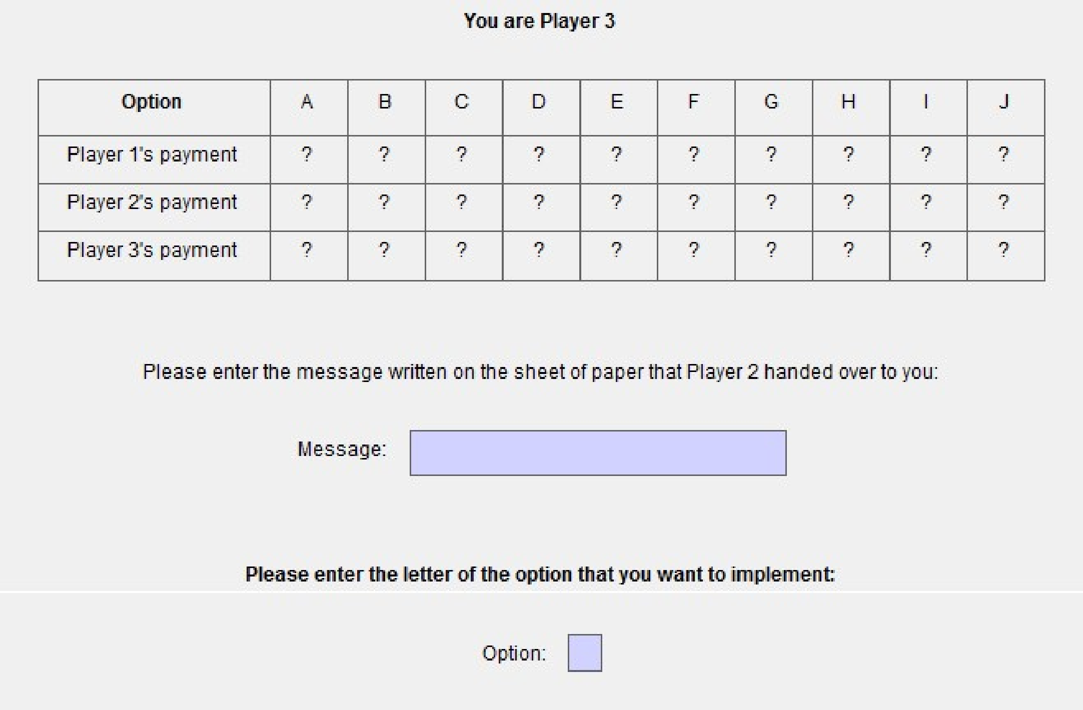
Option J will earn you 10 euros

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.



You are Player 3

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

Please enter the message written on the sheet of paper that Player 2 handed over to you:

Message:

Please enter the letter of the option that you want to implement:

Option:

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.

D.2 Instructions for the online experiment

You are participating in a study on economic decision-making. The study takes around 30 minutes to complete. For completing the study, you will receive *$3.33* (£2.50). In addition, you will be able to earn a *bonus payment.* You will be paid only if you complete the entire study. The study is anonymous. Hence, your identity will not be revealed to others and the identity of others will not be revealed to you.

Next, you will see the instructions. *Please read the instructions carefully as they describe how you can earn the bonus payment.* You will be asked questions to confirm that you have read the instructions. *If you answer these questions incorrectly, you will be excluded from the study and you won't be eligible for payment.*

By continuing to the next screen, you consent to participate in this study. For more details about your consent, click on "See consent form".

Specific instructions

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

The setting

There are ten options, each labeled with a unique letter: A, B, C, D, E, F, G, H, I, or J.

- The computer will randomly assign one option to pay a bonus payment of *$5.00 to Player 1, $5.00 to Player 2, and $5.00 to Player 3.*
- It will also randomly assign another option to pay a bonus payment of *$6.75 to Player 1, $6.75 to Player 2, and $1.50 to Player 3.*
- The remaining eight options pay a bonus payment of *$2.00 to Player 1, $2.00 to Player 2, and $0.00 to Player 3.*

Importantly, *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 bonus 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	2	2	5	2	6.75	2	2	2	2	2
Player 2's payment	2	2	5	2	6.75	2	2	2	2	2
Player 3's payment	0	0	5	0	1.50	0	0	0	0	0

By contrast, *Player 3 will not know which options provide which bonus payments.* The table below shows what Player 3 will see instead.

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 is a *message chosen by Player 1 and Player 2.* After receiving the message, Player 3 chooses one of the ten options. *The option chosen by Player 3 determines the bonus payment of all three players.*

Understanding Question #1 (if you answer incorrectly, you will be excluded from the study)

- Does Player 3 know how the computer assigned the payments to the options?

Player 1 and Player 2 choose a message

Both, Player 1 and Player 2 choose *one message* for Player 3. There are *two available messages.*

- *Message I* corresponds to the option that pays $5 to Player 3. It reads "*Option [here goes the letter of option that pays $5 to Player 3] will earn you $5*".
- *Message II* corresponds to the option that pays $1.50 to Player 3. It reads "*Option [here goes the letter of option that pays $1.50 to Player 3] will earn you $5*".

Note that neither Player 1 nor Player 2 can 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 $5 or $1.50, but he/she can be certain that the option does not pay him/her $0.

Example

Suppose that the computer randomly assigns bonus payments to options as shown in the table below. In this case, Player 1 and Player 2 can choose one of the following two messages to be sent to Player 3:

- *Message I:* "Option F will earn you $5"

- *Message I:* "Option D will earn you $5"

Option	A	B	C	D	E	F	G	H	I	J
Player 1's payment	2	2	2	6.75	2	5	2	2	2	2
Player 2's payment	2	2	2	6.75	2	5	2	2	2	2
Player 3's payment	0	0	0	1.50	0	5	0	0	0	0

What message is sent?

Player 1 and Player 2 choose a message *simultaneously.* Thereafter, the message sent to Player 3 is determined in the following way:

- If *Player 1 chooses Message I*, then regardless of Player 2's choice, *Message I is sent.*
- If *Player 2 chooses Message I*, then regardless of Player 1's choice, *Message I is sent.*
- If *both Player 1 and Player 2 choose Message II*, then *Message II is sent.*

Player 3 will not be informed of the individual choices of Player 1 and Player 2.

Understanding Question #2 (if you answer incorrectly, you will be excluded from the study)

- If Player 1 chooses Message II and Player 2 chooses Message I, which message is sent to Player 3?
- If both Player 1 and Player 2 choose Message II, which message is sent to Player 3?

Player 3 chooses an option

After seeing the message, Player 3 chooses one of the ten options to determine the bonus payments of all players.

Understanding Question #3 (if you answer incorrectly, you will be excluded from the study)

For this question, suppose that labels are assigned to the various options as indicated by the table below.

Option	A	B	C	D	E	F	G	H	I	J
Player 1's payment	2	5	2	2	2	2	6.75	2	2	2
Player 2's payment	2	5	2	2	2	2	6.75	2	2	2
Player 3's payment	0	5	0	0	0	0	1.50	0	0	0

- Suppose that Message II "Option G will earn you $5" is sent to Player 3 and Player 3 implements Option G. What is the bonus payment of each player?

- Suppose that Message I “Option B will earn you \$5” is sent to Player 3 and Player 3 implements Option B. What is the bonus payment of each player?
- Suppose that Message I “Option B will earn you \$5” is sent to Player 3 and Player 3 implements Option E. What is the bonus payment of each player?