Received: 24 October 2022 | Revised: 13 February 2023 | Accepted: 22 March 2023

DOI: 10.1002/bdm.2328

RESEARCH ARTICLE

WILEY

# Preferences for honesty can support cooperation

Aron Szekely[1,2] | David Bruner[3] | Sven Steinmo[4] | Arpad Todor[5] | Clara Volintiru[6] | Giulia Andrighetto[1,7,8]

[1]Institute of Cognitive Sciences and Technologies, National Research Council of Italy, Rome, Italy

[2]Collegio Carlo Alberto, Turin, Italy

[3]Department of Economics, Appalachian State University, Boone, North Carolina, USA

[4]Department of Political Science, University of Boulder Colorado, Boulder, Colorado, USA

[5]Political Science Department, National University for Political Studies and Public Administration, Bucharest, Romania

[6]Faculty of International Business and Economics, Bucharest University of Economy Studies (ASE), Bucharest, Romania

[7]The Institute for Analytical Sociology, Linköping University, Norrköping, Sweden

[8]Institute for Futures Studies, Stockholm, Sweden

**Correspondence**
Aron Szekely, Collegio Carlo Alberto, Turin, Italy.
Email: aron.szekely@carloalberto.org

**Funding information**
Horizon 2020 Framework Programme, Grant/Award Number: 699824; Knut and Wallenberg Grant, Grant/Award Number: 2016.0167; European Union's Seventh Framework Programme, Grant/Award Number: 295675

[Correction added on 25 May 2023, after first online publication: The given names of Sven Steinmo and Giulia Andrighetto were previously omitted and have been added in this version.]

## Abstract

Many collective action problems are inherently linked to honesty. By deciding to behave honestly, people contribute to solving the collective action problem. We use a laboratory experiment from two sites ($n = 331$ and $n = 319$) to test whether honest preferences can drive cooperation and whether these preferences can be differentially activated by framing. Subjects participate in an asymmetric information variant of the public goods game in one of two treatments that vary only in their wording: The Contribution Frame uses a standard public good game framing, while in the Honesty Frame, words aimed to trigger honesty are used. We measure subjects' honesty in three ways using the (i) sender–receiver task, (ii) the die-roll task, and (iii) self-reported honesty levels and account for other-regarding preferences and social norms to disentangle key alternative motives. We find that all three measures of honesty preferences robustly predict contributions, as do other-regarding preferences and empirical expectations but not normative expectations. Additionally, honesty preferences predict contributions in the Honesty Frame but not in the Contribution Frame, although the difference between these is not consistently significant. Finally, we find no differences in average cooperation across the treatments.

**KEYWORDS**
cooperation, honesty, laboratory experiment, social norms, social preferences

## 1 | INTRODUCTION

Groups of people face social dilemmas whenever they have to cooperate to obtain a shared benefit (Dawes, 1980; Kollock, 1998). Classic examples are worker protests (Olson, 1965), the maintenance of shared resources (Hardin, 1968), and the provision of public services by paying taxes (Alm, 1991). The fundamental difficulty is that individual and collective interests conflict, since cooperation is individually costly but the benefits are socially shared. Despite this tension, people often cooperate and solve the social dilemmas that they face.

Why people do so, despite the individual cost, is the topic of much research (Rand & Nowak, 2013). One widespread view is that

cooperation can be supported by a trade-off between self and other-regarding preferences (Andreoni, 1989, 1990; Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999; Murphy et al., 2011). Such other-regarding preferences are dispositional-based accounts which argue that (some) people consider other's outcomes in their own utility function (e.g., Fehr & Schmidt, 1999) and this can allow cooperation to emerge if individual payoffs to cooperating are sufficiently high. Indeed, extensive empirical research shows that other-regarding preferences can indeed increase cooperation (Balliet et al., 2009; Murphy et al., 2011). Another influential perspective proposes a socially mediated view of cooperation and argues that social norms drive cooperation (Andrighetto et al., 2013; Bicchieri, 2006; Elster, 2009; Fehr & Fischbacher, 2004; Schram & Charness, 2015).[1] Social norms can be defined as shared behavioral rules that proscribe or prescribe ways of acting that are followed because of reciprocal expectations and, in some cases, social punishment (Bicchieri, 2006; Cialdini et al., 1990; Elster, 1989). Many studies now confirm the important role that social norms have in cooperation (Bicchieri, 2006; Bicchieri et al., 2022; Bicchieri & Chavez, 2010; Bicchieri & Xiao, 2009; Isler & Gächter, 2022; Kölle & Quercia, 2021; Przepiorka et al., 2022; Szekely et al., 2021) while others also show the close link between social norms and traits (i.e., the dark factor of personality) that are relevant to prosocial behavior (Hilbig et al., 2022).

Here, we argue and test that preferences for honesty are another important driver of cooperation. Specifically, our first research question is

RQ1: Are people's preferences for honesty associated with cooperation in social dilemmas?

Existing evidence regarding this question comes from studies that consider whether the Honesty–Humility factor in the HEXACO model of personality (Ashton & Lee, 2007, 2008)—which is defined by the facets of sincerity, fairness, modesty, and greed avoidance (Lee & Ashton, 2004, p. 334)—predicts cooperation. Consistent with this idea, multiple studies find that Honesty–Humility is positively associated with cooperation in social dilemmas (Heck et al., 2018; Hilbig et al., 2012, 2014, 2018; Hilbig & Zettler, 2009; Hilbig, Zettler, Leist & Heydasch, 2013; Mischkowski & Glöckner, 2016; Zettler et al., 2013; Zhao et al., 2016). Indeed, this association is now well-established (Hilbig, 2022).

Our contribution to this literature is fourfold. First, we use multiple measures of honesty, two of which are incentivized and based on behavior. Existing studies, since they focus on Honesty–Humility, do not rely on incentivized measures of honesty. While these are instructive—Honesty–Humility has been shown to predict behavior in one kind of incentivized honesty task (Heck et al., 2018; Hilbig, 2022)—it is nevertheless important to test whether other incentivized measures, which can help overcome social desirability bias, also show the association between honesty and cooperation (see also Hilbig et al., 2014). Second, we implement cooperation in our study using an incentivized public goods game. Many prior studies use hypothetical payoffs in economic games (e.g., Hilbig et al., 2012; Hilbig, Zettler, Moshagen, & Heydasch, 2013; Zhao et al., 2016) and ask participants to imagine a set of payoffs without actually paying subjects for their decisions. Moreover, apart from one exception (Mischkowski & Glöckner, 2016), the studies we are aware of (see above) consider two-person interactions (e.g., the dictator game or ultimatum game) and not *N*-person social dilemmas like the public goods game. Third, we measure and control for other-regarding preferences and social norms of cooperation, two key alternative mechanisms used to explain cooperation. While some existing studies account for other-regarding preferences (e.g., Hilbig & Zettler, 2009; Mischkowski & Glöckner, 2016), none to our knowledge account for social norms. Fourth, we collect data from substantial sample sizes across two study sites to maximize the power and robustness of our results.

Our second research question aims to understand whether this potential relationship between honesty and cooperation can be affected by the framing of the social dilemma. More specifically, we aim to test whether the framing of a social dilemma—by framing here we mean the specific language used to describe the social dilemma to subjects—can differentially activate honest preferences and in turn affect cooperative behavior. We conjecture that preferences for honesty are likelier to be expressed when the framing of a situation contains honesty cues. Thus, in our second research question, we ask the following:

RQ2: Can honesty motivations be differentially activated based on the framing of a social dilemma?

Additionally, if honesty preferences are expressed while other motives (e.g., altruistic or social norms) remain unchanged and already support cooperation to an extent, then we should see that promoting the expression of honesty translates into higher overall contributions. So, in our final research question, we ask the following:

RQ3: Does implementing honesty framing increase contributions in social dilemmas?

Prior experiments demonstrate that framing can have important effects on behavior (e.g., Andreoni, 1995; Capraro & Vanzo, 2019; Ellingsen et al., 2012; Engel & Rand, 2014; Liberman et al., 2004). The mechanism we posit, a situational activation of preferences, differs from those typically used to explain why framing shapes behavior (beliefs about other's actions, group identity theories, and social norms) (Dreber et al., 2013; Ellingsen et al., 2012; Gerlach & Jaeger, 2017). While the mechanism we test is different to those typically associated with framing, it is consistent with the affordances account of personality–situation interactions (de Vries et al., 2016). Known as domain specific situational affordances, the idea is that "[e]ach situation has a potentially distinct affordance … that is, a different opportunity to express behaviour, and, consequently, to express (or constrain) aspects of personality" (p. 412). One component of this explanation is a "trait activation mechanism" that argues that situational cues influence the probability that a trait is expressed. This in turn draws on trait activation theory (Tett & Burnett, 2003; Tett & Guterman, 2000), the core principle of which is that "personality

[1]The conceptual relationship between social norms and other-regarding preferences is fairly well-established and it is possible to distinguish these empirically (see Bicchieri, 2006). Put bluntly, the former are behavioral rules—that is, they can motivate behavior irrespective of outcome—while when motivated by other-regarding preferences, individuals take actions that achieve outcomes for themselves and others. In contrast, the relationship between honesty and social norms is less understood. Conceptually, and consistent with Abeler et al. (2019), the two should relate. Internalized social norms, or moral norms, can provide the direct costs associated with lying, while social norms can also influence the incentives associated with maintaining the appearance of honesty.

traits are expressed as response to trait-relevant situational cues" (Tett & Burnett, 2003, p. 502).

We can also view of our second and third research questions through the lens of personality theories that combine honesty and cooperativeness in the same underlying trait: the Honesty–Humility factor of HEXACO and the dark factor of personality, "the tendency to maximize one's individual utility— disregarding, accepting, or malevolently provoking disutility for others—accompanied by beliefs that serve as justifications" (Moshagen et al., 2018, p. 656).[2] By triggering the honesty part of these factors through framing a decision situation, one should also trigger the other facets of the underlying factor, thereby promoting cooperation. While in absence of the honesty trigger, the underlying factor should remain partly dormant.

We study our three questions by measuring preferences for honesty and comparing cooperation levels and associations in two asymmetric information variants of linear public goods games that are identical in payoffs but differ in their framing. One treatment, the "Honesty Frame" (HF), is designed to include subtle honesty cues, while the other treatment, the "Contribution Frame" (CF), uses the language of a standard public goods game (e.g., Fehr & Gächter, 2000). The framing of the former makes it possible to be honest or not; it is not simply a decision to cooperate or not. Rather, people face a request to report a state of affairs accurately. In contrast, the latter avoids this possibility.

Our experiment uses an asymmetric information variant of the public goods game in which group members have unequal endowments and do not know the specific allocations. Consequently, there is uncertainty about the extent of others' contributions. Moreover, we frame the cooperative situation using honesty cues in our Honesty Frame treatment. We contend that both asymmetric information and honesty cues are frequent features of real-life social dilemmas, and their combination can implicate honesty as a component of cooperation. For instance, tax payment is structurally a social dilemma in which people do not know the extent of each other's contributions and it also contains honesty cues: The decision of how much to report implies that one can either be honest and report everything or dishonest and report something less than one's full earnings. This in turn affects contributions to the public good. The same logic applies to other settings, including teamwork (people do not know how much time and effort someone has available and they can report this honestly or dishonestly which consequently affects whether they are cooperative or not) and scientific conduct; scientists writing a paper know more about their analyses results than do readers (asymmetric information), and *p*-hacking (Simmons et al., 2011) is dishonest behavior about the selection of these analyses and results that implies personal gain, from the publication of results, but defection toward the scientific community in that it muddies the common pool of scientific knowledge. As a consequence, it is important to study how honesty relates to cooperation in such a setting.

Our questions are also partly motivated by an empirical puzzle found in experimental papers that study cooperation problems framed from a taxation perspective (see Alm, 2019, for review). These use a variant of the public good game in which people decide how much of their income, which may be earned through a real-effort task or simply endowed, to "report." Reported earnings are "taxed" at a set percentage and are audited and fined at known rates. Strikingly, in experimental treatments in which *all tax revenues are destroyed* (and subjects know this), people cooperate at substantial levels and report much of their incomes (Andrighetto et al., 2016; Guerra & Harrington, 2018; Steinmo & D'Attoma, 2021; Zhang et al., 2016). To explain this result, we postulated that the taxation framing of the social dilemma allowed people's preferences for honesty to be expressed, thereby motivating their contribution decisions, even when instrumentally "contributing" did not provide benefits. Fulfilling their intrinsic preferences for honesty—either avoiding direct lying costs or for reputational concerns—would drive behavior in this case and not the lack of material incentives.

The trait activation perspective that we adopt on framing would account for the above puzzle, but two of the standard mechanisms, beliefs about other's actions and group identity theories, cannot explain it. The former argues that framing changes people's expectations of how others behave, which, coupled with other-regarding preferences, changes cooperative behavior. The latter contends that social preferences themselves are flexible and influenced by the framing of the interaction (see Gerlach & Jaeger, 2017). Identifying with a group changes individuals' utility functions to consider the payoffs of group members (e.g., Akerlof & Kranton, 2000; Bacharach, 1999). Since both accounts postulate that behavior is motivated by others' outcomes, they cannot explain why people would contribute to a public good that is subsequently destroyed, while social norms, a third often invoked mechanism for framing effects, may account for our puzzle, and we measure these.

## 2 | METHODS

We conducted two experiments with identical procedures and protocols to study our questions and test the robustness of our results across samples in diverse contexts. One was conducted in Italy, Bologna (Bologna Laboratory for Experiments in Social Science, May–June 2017), while the other was conducted in Romania, Bucharest (National University for Political Studies and Public Administration and Bucharest University of Economic Studies, May 2017).

### 2.1 | Experimental procedure

Our experiment consists of six stages, the order of which was fixed (see Table 1).[3] All of the interactions are anonymous. In *Stage 1*, subjects earn their endowment via a real-effort clerical task in which they

[2]We wish to thank an anonymous reviewer for raising this possibility.

[3]Using a fixed ordering of the experimental stages allows us to reduce wealth effects by limiting subjects' knowledge about their earnings (i.e., by placing the die-roll task last) and to increase the probability of accurate responses in the most important tasks by placing them early in the study. Yet, there is also a disadvantage: There could be particular carry-over effects from one stage to the next (e.g., certain social preferences may be triggered in the SVO slider task by having subjects participate in it after the sender–receiver task).

**TABLE 1** Experimental protocol summary.

| Stage | Task |
|---|---|
| 1 | Real-effort clerical task (Andrighetto et al., 2016; Zhang et al., 2016) |
| 2a | Cooperation decisions in Honesty Frame or Contribution Frame (five rounds) |
| 2b | Social norm elicitation (Bicchieri et al., 2014) |
| 3 | Sender–receiver task (Gneezy et al., 2013) |
| 4 | Social value orientation slider measure (Murphy et al., 2011) |
| 5 | Die-rolling honesty elicitation (Fischbacher & Föllmi-Heusi, 2013) |
| 6 | Questionnaire including self-reported honesty |

*Note*: Stage 2a was repeated five times with stranger matching. Stage 2b was presented to subjects once after their first reporting decision.

input data into the computer from a printed document. For every correctly copied row of data, subjects receive 20 tokens. At the end of the experiment, every one token is converted into 0.01 euros. The task requires little skill, based on copying printed letters and numbers onto the computer, but entails focus and effort. It has been previously used in multiple studies (Andrighetto et al., 2016; Guerra & Harrington, 2018; Steinmo & D'Attoma, 2021; Zhang et al., 2016). Subjects were paid the sum of their earnings in Stages 2a, 3, 4, and 5. They did not receive feedback about their earnings in the stages until the end of the experiment (except in Stage 5 which we put last to avoid wealth effects). We use the real-effort task to create variation in subjects' earnings and informational asymmetries such that subjects know more about their own earnings than others in their groups, both of which allow them to report honestly or not their earnings, and to increase subjects' motivation.

In *Stage 2a*, subjects participate in a public goods game in groups of four. Whatever they put into the common pot is doubled and split evenly among the group members, irrespective of their decisions. Exactly how they receive this information depends on which of the two treatments they are in. In the Contribution Frame treatment, the words "contribute" and "withhold" are used, while in the Honesty Frame treatment, the words "report" and "hide" are used instead. Thus, the treatment manipulation consists of using contribute or report and withhold or hide (the relevant variants of these words are also used, e.g., contributed, reported, and hid; see Data Availability Statement for instructions). For instance, on their first decision screen, subjects observe the following (square brackets enclose the between-subjects' treatment wording):

> Based on your performance in the clerical task, your income is X tokens.
> Choose how much of your income to [*contribute/report*] below.
> Income that you [*withhold/hide*], you keep for yourself.

In the Contribution Frame treatment, we used only words associated with cooperation; thus, honesty preferences should not be expressed. Consider that one's decision to "contribute" or "withhold" is not about being honest: They are simply subjects' decisions to do as they wish. In contrast, the Honesty Frame treatment "report" and "hide" were chosen specifically to trigger honesty preferences. When subjects are asked to report their earnings, they can be dishonest and not comply with the request (which then has the consequence of determining the tokens in the collective pot). We use such a minimal change in framing because we want to change honesty motivations precisely, without affecting other variables, between the treatments. Additionally, we specifically chose the wording so that the strength of their positive or negative associations are balanced between the treatment. Subjects make a total of five cooperation decisions in the public goods game, receiving feedback after each decision. Every time they are grouped with others whom they have not played with before, a perfect strangers design, and they know this.[4] At the end of their fifth decision, we include three manipulation check questions.[5] Subjects were paid for one randomly chosen round in Stage 2a that was only revealed to them at the end of the experiment.

The rest of the stages are included to identify why subjects contribute. Our measures concern social norms (Stage 2b), honesty as measured through the sender–receiver task (Stage 3), other-regarding preferences (Stage 4), the die-rolling task (Stage 5), and self-reported honesty (Stage 6). We include measures of civic norms and multiple demographic questions in the post-experimental questionnaire (Stage 6).

After their first decision in the public goods game, subjects are asked to respond to three questions in *Stage 2b* concerning the cooperation decision they just made that elicit personal normative beliefs, empirical expectations, and normative expectations (Bicchieri, 2006; Bicchieri et al., 2014).[6] Bicchieri's (2006) influential account of social norms proposes that social norms are behavioral rules that are followed due to sufficiently high empirical and normative expectations. The former are beliefs about whether other people follow a behavioral rule (e.g., contribute to the pot), and the latter is a second-order belief held by an individual about whether other people think that one ought to follow the behavioral rule. Only when a behavior is motivated by both empirical and normative expectations can it be plausibly considered to be a social norm. Thus, we measure both components of Bicchieri's theory of social norms. We also measure personal normative beliefs, which are people's own views on whether a behavior is socially appropriate or not. This gives an indication of the presence of a behavioral rule and is also a key confounder of social norms. These allow us to measure, and distinguish between actions taken because of personal normative beliefs, what people like or think they should

[4]Given the size of the laboratories, subjects could only be matched with perfect strangers five times.

[5]These were "Please think back to the [*contributing/reporting*] decisions you just made. Indicate how much you agree with the following statements." In general, when I made my decisions: 1. "I felt like I should be honest" (Honest); 2. "I wanted to make the most money for myself" (Money); and 3. "I felt like I should contribute to the common good" (Common Good).

[6]They are asked to (i) "Please indicate what percentage of income the members of your group, including you, should have [*contributed/reported*]?"; (ii) "How many of the other three members of your group [*contributed/reported*] the following percent of their income?" with three input boxes for "Less than 50%," "50%," and "More than 50%"; and (iii) "How do the other three members of your group respond to question 1?" with three input boxes for "Less than 50%," "50%," and "More than 50%."

do, descriptive norms, what people expect others to do, and actions taken because of social norms—what people expect others to do and they expect others to think that they ought to do (Bicchieri, 2006). Following these questions, subjects make their four remaining decisions in the public goods game.

*Stage 3* elicits subjects' aversion to lying using Gneezy et al.'s (2013) modified sender–receiver task. Subjects are randomly paired and allocated a random integer between 1 and 6. One of the pair is assigned the role of *sender* and the other the *receiver*. The sender learns the value of the random integer and has to send a message about this value to the receiver (i.e., "The number is $Z$" with $Z \in \{1,2,3,4,5,6\}$). The receiver is told about the message and decides whether to believe the sender. The sender's payoff from the task increases linearly in the reported number. Hence, if the sender only cares about their payoff, they should always send $Z = 6$. The receiver's payoff depends on whether they believe the sender or not and whether the message is accurate or not. If the receiver believes the sender and the message is accurate, they receive 100 tokens. If they believe the sender and the message is inaccurate, they receive zero tokens. If they do not believe the sender, they receive 50 tokens, regardless of the accuracy of the message. So, the receiver faces a risky decision. Decisions are elicited using the strategy method: Both subjects in a pair are asked to decide for each of the six possible numbers they could be assigned what message they would like to send. At the end of the experiment, subjects are informed of their role, the randomly assigned number, and their resulting payoff.

*Stage 4* measures subjects' other-regarding preferences using the social value orientation (SVO) slider measure (Murphy et al., 2011). Subjects are randomly paired and both decide how to split tokens in 15 dictator games. Each dictator game presents the chooser with a different trade-off between keeping and giving some tokens. Subjects are informed that only one person in each pair will be randomly selected to have their decisions implemented. This task allows subjects to be classified according to how much they consider the outcome of others' in their payoff function. Subjects are paid for one of their randomly chosen decisions, or their partner's decisions, at the end of the experiment.

*Stage 5* elicits subjects' honesty using the die-roll reporting task (Fischbacher & Föllmi-Heusi, 2013). Subjects are each given a six-sided die and asked to roll it in private and then input the outcome of their roll on the computer screen. Subjects are paid 20 tokens times the reported number of their roll, unless they report 6, in which case they receive zero. Subjects in this stage know how much they earn as their earning depends entirely on their own choice. To prevent subjects' earnings from affecting their other decisions, we put this stage—in which they knew their earnings—as the last opportunity for subjects to earn money.

In the final part of the experiment, *Stage 6*, subjects complete a post-experimental questionnaire. We asked subjects to tell us how honest they believe that they are ("Please indicate how honest you think you are") and included questions on demographic information and further potential motivations (see Data S1 for screenshots). After completion of the survey, participants were paid individually in private.

Overall, we collected three measures of honesty: sender's message in the sender–receiver task, reporting in the die-rolling task, and self-reported honesty. Each measure has advantages and disadvantages (Gerlach et al., 2019; Soraperra et al., 2019). The sender–receiver task measures honesty using an incentivized approach and captures individual-level behavior. Nevertheless, there are also plausible downsides to it. Subjects may correctly believe that the experimenters can tell how much they are lying and as such behave more normatively than if their decision could not be observed (Gneezy et al., 2013, p. 294). Additionally, the task implicates other-regarding preferences since lying potentially impose a cost on the recipient of the lie. The die-roll task, meanwhile, is incentivized, it is not possible for experimenters to identify lying at the individual level so subjects make genuinely private decisions, and it does not directly implicate other-regarding preferences (albeit concern for the outcomes of the experimenter may still play a role). However, as a consequence, the die-roll task cannot provide precise individual-level information about dishonesty as a subject may report a different number that they genuinely rolled. Finally, self-reported honesty is a simple measure, responses to it cannot implicate other-regarding preferences—in contrast to the two incentivized measures (Isler & Gächter, 2022; Soraperra et al., 2019), since there are no material consequences to responses, and it is likely to capture self or public image concerns of honesty—a key component of honest behavior (Abeler et al., 2019). At the same time, and crucially, self-reported honesty is not incentivized and responses can be individually identified the experimenter. Thus, social desirability bias may dominate responses. Since all three measures have plausible advantages and disadvantages and may have different links with cooperation, we study each measure separately but aim to identify common patterns across them. We also create a single composite honesty variable using exploratory factor analysis on the three measures. As a robustness check on this composite variable, we also create an alternative that is the mean normalized response across the three honesty measures. Both of these composite variables are exploratory.

## 2.2 | Analytic strategy

To examine whether honesty predicts cooperation (RQ1), we start by displaying bivariate associations between each of the three honesty measures and cooperation using correlations (one observation per subject against average cooperation in the public goods game) and bivariate regressions with cluster robust standard errors at the subject level. We then expand the regression models by adding control variables for the two alternative mechanisms (other-regarding preferences and social norms) and, in another set of models, adding an extensive set of controls to test the robustness of the associations. These controls are a dummy variable for study, personal normative beliefs, three questions from the WVS concerning civic norms that justify cheating (on taxes, claiming government benefits, and avoiding paying a fare on public transport), whether most people try to take

advantage of you, left–right political orientation, gender, age, self-reported risk preferences, previous experience in experiments, and perceived performance in the clerical task. We also conduct exploratory analyses by replicating the aforementioned analyses using the factor-analysis generated composite honesty variable. Then, we use the same regression model configurations to analyze whether honesty motivations can be differentially activated by framing (RQ2) but include an interaction between the honesty measure and treatment to capture the difference in slopes across treatments. We also conduct exploratory analyses by running these analyses with the composite honesty predictor. To answer whether an honesty framing increases contributions (RQ3), we use linear regressions with cluster robust standard errors at the individual level and five linear regressions, one for each decision.

## 3 | RESULTS

Subjects for the study in Italy were recruited using the Online Recruitment System for Experimental Economics (ORSEE) (Greiner, 2015), and those in Romania were recruited through online announcements at the respective universities. The sessions were programmed and conducted using z-Tree (Fischbacher, 2007). We collected data on 331 subjects in Italy ($mean_{age} = 23.2$, $sd_{age} = 3.13$, 54.4% female, and 89.7% students; Table S1) and 319 subjects[7] in Romania ($mean_{age} = 20.8$, $sd_{age} = 2.95$, 58.8% female, and 99.7% students; Table S2). Both reached our target number of subjects based on power calculations (see Data S1). We conducted 25 experimental sessions in total (12 in Italy and 13 in Romania), and each session lasted approximately 60 min. The mean number of rows that subjects were able to copy in the real-effort task was 8.77 ($sd = 1.96$) in Italy and 8.44 ($sd = 2.71$) in Romania. As we did not find any meaningful differences between the two studies and they are exact replications, differing only in the location and language of the instructions, we analyze them together and include session dummies in the analyses that test robustness. In Data S1 (Italy only analyses and Romania only analyses) we also present results separated by study. The experimental code, instructions, data, and analyses are available for download at https://osf.io/ckbzf (DOI 10.17605/OSF.IO/CKBZF).

### 3.1 | Data preparation

To analyze the sender–receiver task, we consider senders' deviations away from perfect truth-telling. Following Gneezy et al. (2013), we take the absolute difference between the message Z sent by each sender and the number allocated to the pair. So, a fully honest sender, who sends the same message as the number that he or she observes, receives a score of 0. While a fully dishonest sender, who deviates as much as possible from that number, receives a score of 24 (Figure 1a).[8] For the die-roll task, we make only one change before analysis. We recode the reported number "6" into the number "0." This is because reporting 6 gains the subject 0 euros; the lowest possible amount so reporting 6 is the answer that is most likely to be honest (Figure 1b). Regarding self-reported dishonesty, this runs from 0 to 9 with 0 being very honest and 9 being very dishonest (Figure 1c). Thus, for each of these variables, a higher value means more dishonesty.

Personal normative beliefs, empirical expectations, and normative expectations are all continuous variables from 0% to 100%, with 100 indicating full reporting (moral beliefs, expectations of what will happen, and normative expectations, respectively) and 0 indicating no reporting in one's group. While for SVO, people with a higher angle are more other-regarding in their preferences and those with a lower angle are less other-regarding.

### 3.2 | Honesty measures

We start by checking the relationship between the three honesty variables and find that they are all positively associated (sender–receiver and die-rolling: $r = .156$, $p < .001$; sender–receiver and self-reported dishonesty: $r = .173$, $p < .001$; die-rolling and self-reported dishonesty: $r = .098$, $p = .014$; Figure 1d-f). This suggests that all three tasks are, more or less, valid measures of honesty. Yet, consistent with our expectations and prior work, the sender–receiver task is the most closely correlated with social value orientation (sender–receiver and social value orientation: $r = -.296$, $p < .001$) although so are the other two measures (die-rolling and social value orientation: $r = -.258$, $p < .001$; self-reported dishonesty and social value orientation: $r = -.193$, $p < .001$). Additionally, die-rolling appears to be a somewhat noisy individual-level measure with subjects who were entirely honest in the sender–receiver task reporting all die-roll numbers (from 0 to 5), partly because some of them will have truly earned those amounts (Figure 1d), while self-reported dishonesty, unsurprisingly given that lack of incentivization, is highly right skewed with the majority of subjects (71.9%) ratings themselves between 0 and 2 on dishonesty (Figure 1c). This suggests that self-reported dishonesty is partly a reflection of self or public image concerns.

To check the demand effect in the sender–receiver task, we compare the proportion of completely honest subjects in the die-roll task (Figure 1a,b). If the observability of their dishonesty by the experimenter is a deterrent to lying because subjects want to appear to be honest, then the proportion of honest subjects should be higher in the sender–receiver task than the die-roll task. Following Fischbacher and Föllmi-Heusi (2013), we use the 6.46% of subjects who earned 0 euros in the die-roll task to estimate that 38.76% of subjects were

[7]Due to a procedural error, 24 subjects did not answer the end questionnaire in one of the Romanian sessions.

[8]This occurs when a sender chooses the message 6 for the numbers 1, 2, and 3, and the message 1 for 4, 5, and 6 (leading to differences of 5, 4, 3, 3, 4, and 5 respectively). For completeness, and consistent with Gneezy et al. (2013), we do not remove 33.1% of senders who, at least in one case, send a message that is lower than their assigned number—a counterproductive lie. Nevertheless, excluding these counterproductive senders from the relevant analysis does not substantively change the results (Tables S3 and S4). The majority (60.3%) of senders who make a counterproductive lie only do so either once or twice.

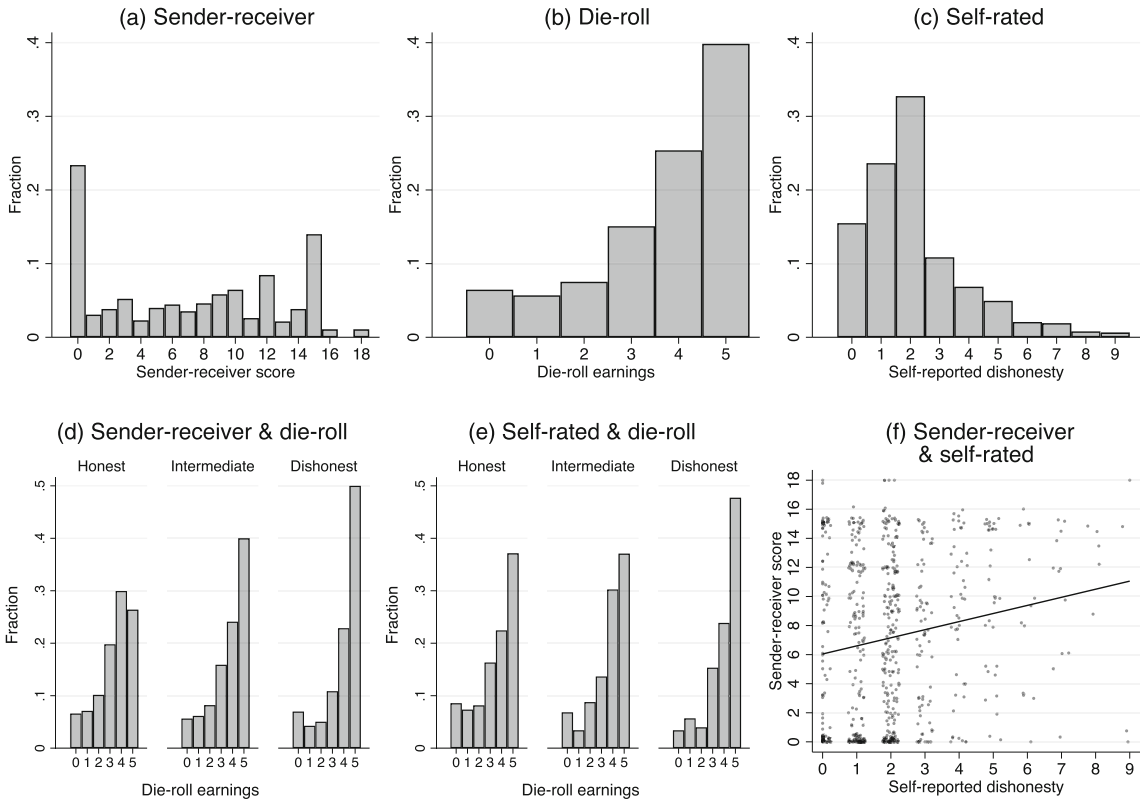


**FIGURE 1** Honesty measures: Their distributions and associations. The marginal distributions are displayed for the (a) sender–receiver task scores, the (b) die-roll task earnings, and (c) self-rated dishonesty. The joint distributions are shown for (d) die-roll earnings broken down into honest, intermediate honest, and dishonest sender–receiver messages and (e) die-roll earnings broken down into honest, intermediate honest, and dishonest self-rated dishonest. (f) The association between the sender–receiver score and self-rated dishonesty with an OLS line of best fit overlaid.

completely honest about that task.[9] By contrast, 23.38% of subjects were completely honest in the sender–receiver task. Hence, we find no evidence of a demand effect associated with the sender–receiver task. If anything, the data suggest the sender–receiver task may elicit spiteful dishonesty among subjects.

## 3.3 | RQ1: Are people's preferences for honesty associated with cooperation in social dilemmas?

We find that all three measures of honesty robustly and significantly predict contributions: More dishonesty is associated with lower contributions (Figure 2a–c and Table 2). This is found when analyzing honesty and cooperation using bivariate correlations ($r_{sender-receiver} = -.198$, $p < .001$; $r_{die-roll} = -.170$, $p < .001$; $r_{self-report} = -.179$, $p < .001$), bivariate regression models ($b_{sender-receiver} = -.906$, $p < .001$; $b_{die-roll} = -2.898$, $p < .001$; $b_{self-report} = -2.604$, $p < .001$), multiple regression models with the two alternative mechanisms ($b_{sender-receiver} = -.529$, $p = .002$; $b_{die-roll} = -1.578$, $p = .013$; $b_{self-report} = -1.230$, $p = .028$), and multiple regression models with extensive controls ($b_{sender-receiver} = -.559$, $p < .001$; $b_{die-roll} = -1.295$, $p = .040$; $b_{self-report} = -1.165$, $p = .035$; Table 2).

We also replicated the above analyses with the composite honesty variable based on factor analysis and find that there is a significant negative association between the composite variable and cooperation in every analysis (Table S5). This is the case with a bivariate correlation ($r_{composite\ \ honesty} = -.277$, $p < .001$), a bivariate regression model ($b_{composite\ \ honesty} = -7.066$, $p < .001$), a regression controlling for social value orientation and social norms ($b_{composite\ \ honesty} = -4.149$, $p < .001$), and a regression model with extensive controls ($b_{composite\ \ honesty} = -4.121$, $p < .001$). Identical results are found with the alternative composite honesty variable based on a normalized mean of the three separate measures (Table S7).

Concerning the other two theoretically identified mechanisms of other-regarding preferences and social expectations, we find that social value orientation and empirical expectations always significantly and positively associated with contributions while normative expectations are only significantly positively associated with contributions in the simple models. Once additional controls are

[9]Assuming subjects are unconditionally honest and do not underreport their roll implies that among the one sixth of subjects who truly rolled 6 (and hence earned 0), 6.46% reported this honesty. Thus, [0.0646 / (1/6)] * 100 = 38.76% of reported numbers were honest.

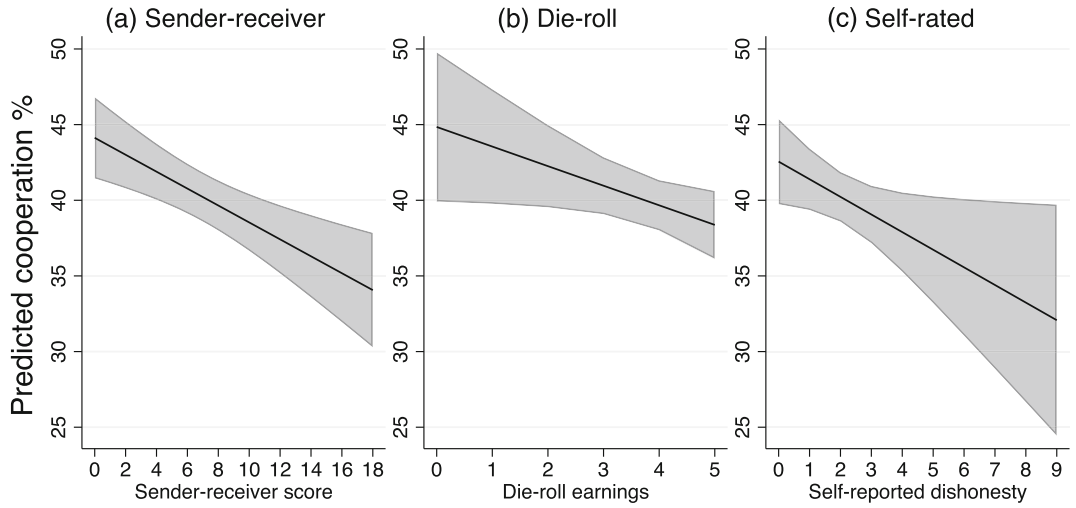


**FIGURE 2** Cooperation across the three honesty measures. Plots (a)–(c) display the associations between cooperation and the (a) sender–receiver score, (b) die-roll earnings, and (c) self-rated dishonesty. Lines represent predicted values, and shaded areas represent 95% CIs accounting for clustering at the individual level (626 subjects; 24 subjects did not complete the end questionnaire). Predicted values and confidence intervals derived from the full multiple linear regression models are displayed in Table 2 (models 2).

**TABLE 2** Predictors of contribution.

| | Sender–receiver task | | Die-roll task | | Self-reported dishonesty | |
|---|---|---|---|---|---|---|
| | Model 1 | Model 2 | Model 1 | Model 2 | Model 1 | Model 2 |
| Sender–receiver score | −0.529** | −0.559*** | | | | |
| | (0.169) | (0.157) | | | | |
| Die-roll earnings | | | −1.578* | −1.295* | | |
| | | | (0.636) | (0.631) | | |
| Self-reported dishonesty | | | | | −1.230* | −1.165* |
| | | | | | (0.559) | (0.552) |
| SVO angle | 0.461*** | 0.367*** | 0.479*** | 0.398*** | 0.488*** | 0.401*** |
| | (0.064) | (0.061) | (0.063) | (0.061) | (0.063) | (0.060) |
| Empirical expectations | 0.250*** | 0.222*** | 0.247*** | 0.221*** | 0.230*** | 0.213*** |
| | (0.043) | (0.041) | (0.043) | (0.042) | (0.044) | (0.042) |
| Normative expectations | 0.109** | 0.019 | 0.111** | 0.023 | 0.107** | 0.017 |
| | (0.038) | (0.038) | (0.039) | (0.038) | (0.039) | (0.038) |
| Additional controls | No | Yes | No | Yes | No | Yes |
| Constant | 19.882*** | −6.799 | 21.484*** | −7.157 | 18.648*** | −7.304 |
| | (2.488) | (7.227) | (3.275) | (7.944) | (2.556) | (7.718) |
| Observations | 3250 | 3130 | 3250 | 3130 | 3130 | 3130 |

*Note*: Coefficients are displayed and standard errors are in parentheses. All models are linear regression models with standard errors clustered according to subjects (either 650 or 626; the difference is because 24 subjects did not complete the end questionnaire). Additional controls are dummy variable for study, personal normative beliefs, three questions from the WVS justifying cheating (on taxes, claiming government benefits, and avoiding paying a fare on public transport), whether most people try to take advantage of you, left–right political orientation, gender, age, self-reported risk preferences, previous experience in experiments, and perceived performance in the clerical task.
$^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, and $^{***}p < .001$.

included—specifically personal normative beliefs—the relationship between normative expectations and cooperation becomes nonsignificant (Table 2).

**Result 1**: Preferences for honesty positively predict contributions even when accounting for other-regarding preferences, social expectations, and a range of additional controls.

### 3.4 | RQ2: Can honesty motivations be differentially activated based on the framing of a social dilemma?

Before analyzing the between-treatment predictors of cooperation, we consider our manipulation checks. Unexpectedly, we find no between-treatment differences (Money: $b = .049$, $p = .592$; Common good: $b = -.023$, $p = .810$; Honest: $b = -.155$, $p = .091$). This null finding is likely explained by a combination of two factors: Our framing manipulation was subtle, and the checks were placed *after* all five contribution decisions were made. Given that subjects receive feedback after each decision, the effect of framing on the manipulation checks may have already been overridden by experience at the time we include the checks. Despite this, and as we show below, we think there is good reason to think that our manipulations had effects.

Were honesty motivations triggered to a greater extent in the Honesty Frame than in the Contribution Frame? Across all models and for all three measures, dishonesty predicts lower contributions in the Honesty Frame (Figure 3a–c and Table 3). This is the case in regression models with no controls ($b_{sender-receiver} = -1.128$, $p < .001$; $b_{die-roll} = -3.804$, $p < .001$; $b_{self-report} = -4.902$, $p < .001$), multiple regression models with the alternative mechanisms ($b_{sender-receiver} = -.767$, $p = .001$; $b_{die-roll} = -2.556$, $p = .002$; $b_{self-report} = -3.395$, $p < .001$), and multiple regression models with extensive controls ($b_{sender-receiver} = -.767$, $p < .001$; $b_{die-roll} = -2.101$, $p = .007$; $b_{self-report} = -3.178$, $p < .001$; Table 3).

In the Contribution Frame meanwhile, honesty measures do not consistently and significantly predict contributions (Figure 3a–c and Table 3). In the regression model with no controls, the sender–receiver task is significantly associated with contribution ($b_{sender-receiver} = -.655$, $p = .015$) but not the other two measures ($b_{die-roll} = -1.781$, $p = .087$; $b_{self-report} = -1.244$, $p = .116$), and already once the controls for the alternative mechanisms are included, the associations become far from significant ($b_{sender-receiver} = -.254$, $p = .320$; $b_{die-roll} = -.393$, $p = .689$; $b_{self-report} = .040$, $p = .955$) suggesting that the sole exception is due to confounding. In the multiple regression models with further controls, similarly no associations are found ($b_{sender-receiver} = -.342$, $p = .153$; $b_{die-roll} = -.393$, $p = .685$; $b_{self-report} = -.022$, $p = .974$).

While the difference between the regression slopes is consistently in the expected direction, such that the association is stronger in the Honesty Frame than in the Contribution Frame, the interaction term does not always reach significance (Figure 3a–c and Table 3). In the basic regressions with no controls, the interaction coefficient is not significant for the sender–receiver task and the die-roll task ($b_{sender-receiver} = .472$, $p = .206$; $b_{die-roll} = 2.024$, $p = .148$), but it is significant for self-reported honesty ($b_{self-report} = 3.659$, $p = .003$). The same pattern holds in the regression models with alternative mechanism controls ($b_{sender-receiver} = .513$, $p = .122$; $b_{die-roll} = 2.164$, $p = .085$; $b_{self-report} = 3.435$, $p = .001$) and with extensive controls ($b_{sender-receiver} = .425$, $p = .173$; $b_{die-roll} = 1.708$, $p = .156$; $b_{self-report} = 3.200$, $p = .002$).

Reproducing the above analyses with the composite honesty measure leads to similar, and even stronger, results (Table S6). It predicts cooperation in the Honesty Frame in every model: a regression without controls ($b_{composite\ honesty} = -11.215$, $p < .001$), a regression with the alternative mechanism controls ($b_{composite\ honesty} = -8.129$, $p < .001$), and the regression with additional controls ($b_{composite\ honesty} = -7.584$, $p < .001$). While in the Contribution Frame, composite honesty does not predict cooperation in the regression model with social value orientation and social norms

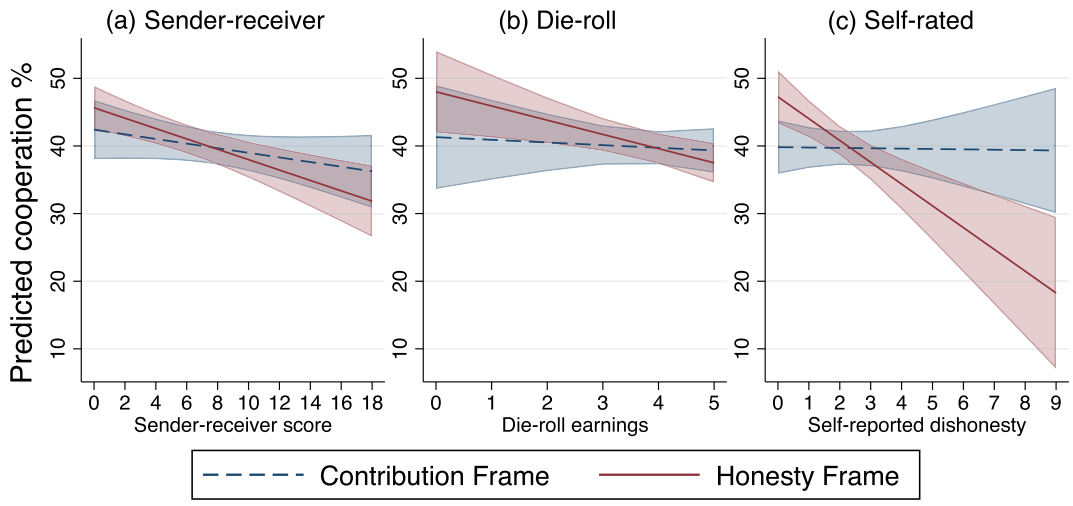


**FIGURE 3** Cooperation by treatment across the three honesty measures. Plots (a)–(c) display the associations between cooperation and the (a) sender–receiver score, (b) die-roll earnings, and (c) self-rated dishonesty according to treatment. Lines represent predicted values, and shaded areas represent 95% CIs accounting for clustering at the individual level (626 subjects; 24 subjects did not complete the end questionnaire). Predicted values and confidence intervals derived from the full multiple linear regression models are displayed in Table 3 (models 2).

**TABLE 3** Predictors of contribution across the Honesty Frame and Contribution Frame.

| | Sender–receiver task | | Die-roll task | | Self-reported dishonesty | |
|---|---|---|---|---|---|---|
| | Model 1 | Model 2 | Model 1 | Model 2 | Model 1 | Model 2 |
| Sender–receiver score | −0.767*** | −0.767*** | | | | |
| | (0.221) | (0.206) | | | | |
| Die-roll earnings | | | −2.556** | −2.101** | | |
| | | | (0.810) | (0.775) | | |
| Self-reported dishonesty | | | | | −3.395*** | −3.178*** |
| | | | | | (0.822) | (0.817) |
| Contribution frame | −4.410 | −3.240 | −8.904+ | −6.706 | −7.787* | −7.235** |
| | (2.928) | (2.775) | (5.045) | (4.881) | (3.038) | (2.797) |
| Contribution frame # sender–receiver score | 0.513 | 0.425 | | | | |
| | (0.331) | (0.312) | | | | |
| Contribution frame # die-roll earnings | | | 2.164+ | 1.708 | | |
| | | | (1.255) | (1.204) | | |
| Contribution frame # self-reported dishonesty | | | | | 3.435** | 3.200** |
| | | | | | (1.069) | (1.024) |
| SVO angle | 0.468*** | 0.373*** | 0.478*** | 0.398*** | 0.486*** | 0.400*** |
| | (0.064) | (0.061) | (0.063) | (0.061) | (0.063) | (0.060) |
| Empirical expectations | 0.247*** | 0.218*** | 0.247*** | 0.222*** | 0.234*** | 0.216*** |
| | (0.043) | (0.041) | (0.043) | (0.042) | (0.043) | (0.042) |
| Normative expectations | 0.110** | 0.021 | 0.112** | 0.024 | 0.101* | 0.012 |
| | (0.038) | (0.038) | (0.039) | (0.038) | (0.040) | (0.039) |
| Additional controls | No | Yes | No | Yes | No | Yes |
| Constant | 21.804*** | −4.313 | 25.513*** | −3.872 | 23.656*** | −6.645 |
| | (2.732) | (7.437) | (3.877) | (8.049) | (3.223) | (7.466) |
| Observations | 3250 | 3130 | 3250 | 3130 | 3130 | 3130 |

*Note*: Coefficients are displayed and standard errors are in parentheses. All models are linear regression models with standard errors clustered according to subjects (either 650 or 626; the difference is because 24 subjects did not complete the end questionnaire). The reference treatment group is the Honesty Frame. Additional controls are dummy variable for study, personal normative beliefs, three questions from the WVS justifying cheating (on taxes, claiming government benefits, and avoiding paying a fare on public transport), whether most people try to take advantage of you, left–right political orientation, gender, age, self-reported risk preferences, previous experience in experiments, and perceived performance in the clerical task.
$^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, and $^{***}p < .001$.

($b_{composite\ honesty} = -1.252$, $p = .377$) and the regression with further controls ($b_{composite\ honesty} = -1.575$, $p = .229$) although there is an association in the basic regression ($b_{composite\ honesty} = -4.009$, $p = .009$). Moreover, the difference between the slopes is significant in every model ($b$ range: 6.009 to 7.206, always $p < .001$) showing that honesty is more closely associated with contributions in the Honesty Frame than in the Contribution Frame. The alternative composite honesty variable shows identical results (Table S8).

**Result 2**: Contributions are positively associated with honesty in the Honesty Frame. There is no significant association between honesty and contributions in the Contribution Frame. The difference between the two is sometimes significant.

## 3.5 | RQ3: Does implementing honesty framing increase contributions in social dilemmas?

We find no overall difference in cooperation between the two treatments. Subjects cooperate on average 39.53% in the Contribution Frame and 41.33% in the Honesty Frame ($b = 1.800$, $p = .375$).

When we analyze cooperation broken down by decision number (Figure 4), we find that cooperation in the Honesty Frame is significantly higher than in the Contribution Frame in decision 1 (45.98 vs. 40.61: $b = 5.370$, $p = .023$) but not in decision 2 (44.38 vs. 40.64: $b = 3.744$, $p = .130$), decision 3 (41.09 vs. 41.28: $b = -0.191$, $p = .942$), decision 4 (37.34 vs. 40.13: $b = -2.796$, $p = .286$), nor decision 5 (37.84 vs. 34.97: $b = 2.872$, $p = .266$). The

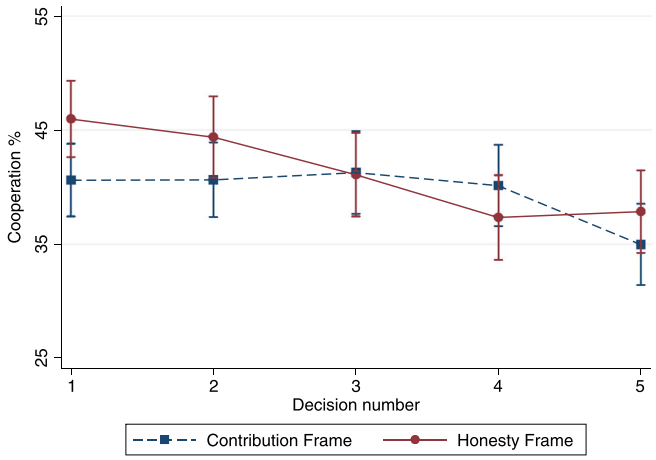


**FIGURE 4** Average cooperation by round and treatment. Error bars represent 95% CIs accounting for clustering at the individual level (320 subjects in the Contribution Frame and 330 subjects in the Honesty Frame).

difference between the treatments in decision 1 should be treated with caution as we do not find significant difference in cooperation when analyzing the studies separately (see Data S1, Italy only analyses and Romania only analyses).

**Result 3**: There are no overall differences in contributions to the public good between the Contribution Frame and the Honesty Frame treatments.

## 4 | DISCUSSION AND CONCLUSIONS

We find clear evidence that honesty predicts cooperation even when accounting for other-regarding preferences, social norms of cooperation, and a wide range of additional controls. This highlights the important link between preferences for honesty and contributing to social dilemmas. Our finding, across all three measures of honesty and the composite measures, is consistent with the previous papers that have looked at whether Honesty–Humility predicts cooperation (Hilbig et al., 2012, 2018; Hilbig & Zettler, 2009; Mischkowski & Glöckner, 2016; Zettler et al., 2013). Thus, honesty predicts cooperation whether the measure of honesty is based on the HEXACO personality inventory, the sender–receiver task, the die-roll task, or even self-reported honesty. Showing this link also in a fully incentivized setting, including behavioral measures and an incentivized public goods game, across two sample locations, and accounting for two key alternative mechanisms, reinforces the previous literature that honesty and cooperation are associated.

In addition to honesty, we show that, as expected based on extensive past work, other-regarding preferences, personal normative beliefs, and empirical expectations—but not normative expectations—all predict contribution. These point to the crucial roles of preferences but also social expectations, which are only now widely acknowledged to be important drivers of cooperative behavior (Bicchieri, 2006, 2017; Ostrom, 2005).

Yet, we also find evidence that the association between honesty and contribution is only consistently found in the honesty-framed treatment and not in the honesty-suppressing treatment. This points to an interesting possibility: That the features of a situation affect whether honesty preferences are expressed or not in the decision. This would have two curious implications. First, some real-life situations are framed in such a way that they plausibly allow concerns for honesty to manifest themselves in cooperative behavior (like in our honesty-framed treatment) while others do not. Whether honesty preferences are relevant for understanding why people contribute is thus also partly a feature of the interaction situation. Second, if honesty shapes cooperation, then the standard neutrally framed public good game may not be an accurate model of the many real-world cooperation situations in which there are honesty cues.

It is important to keep in mind however that this finding—differential activation of honesty preferences—does not receive unequivocal support in our data. While in the honesty-framed treatment, there are consistent associations across each honesty measure and cooperation, but not in the Contribution Frame, the difference between the two associations is sometimes not significant. Thus, this should be taken cautiously. Indeed, and turning to the broader literature on honesty, while subtle cues are sometimes found to affect honesty-linked behavior (Bateson et al., 2006; Ernest-Jones et al., 2011), other (high-powered) studies find no effect (Kristal et al., 2020). Moreover, it is unclear whether the same finding would also apply to the standard public goods games setting instead of our asymmetric information variant.

Finally, and contrary to our expectations, we find no overall difference in cooperation between the Honesty Frame and the Contribution Frame. How can we understand this result? One possibility is that feedback based on others' actions overpowers the framing effect by decision 2. Another is that by implicating honesty preferences, not only do subjects with strong honesty preferences contribute more but also those with low honesty contribute less than they would have done in the Contribution Frame. Based on the distribution of honesty in the sample, this offsets the anticipated increases.

### ACKNOWLEDGMENTS

The work presented in this paper has been supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant Agreement no. (295675), Knut and Wallenberg Grant "How do human norms form and change?" 2016.0167, and the Horizon 2020 Framework Programme Project PROTON "Modelling the Processes leading to Organised crime and Terrorist Networks" under Grant Agreement No. 699824 coordinated by Giulia Andrighetto. The funders had no role in study design, data collection and analysis, decision to publish,

or preparation of the manuscript. We want to thank Alexandra Diaconescu, Daniela Panica, George Stefan, and Andrada Nimu for their efforts in organizing the Bucharest experiments. Open Access Funding provided by Consiglio Nazionale delle Ricerche within the CRUI-CARE Agreement.

## CONFLICT OF INTEREST STATEMENT

None.

## DATA AVAILABILITY STATEMENT

The experimental code, instructions, data, and analyses are available for download at https://osf.io/ckbzf (DOI 10.17605/OSF.IO/CKBZF).

## ORCID

*Aron Szekely* https://orcid.org/0000-0001-5651-4711

## REFERENCES

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115–1153. https://doi.org/10.3982/ECTA14673

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, *115*(3), 715–753. https://doi.org/10.1162/003355300554881

Alm, J. (1991). A perspective on the experimental analysis of taxpayer reporting. *The Accounting Review*, *66*(3), 577–593.

Alm, J. (2019). What motivates tax compliance? *Journal of Economic Surveys*, *33*(2), 353–388. https://doi.org/10.1111/joes.12272

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy*, *97*(6), 1447–1458. https://doi.org/10.1086/261662

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, *100*(401), 464–477. https://doi.org/10.2307/2234133

Andreoni, J. (1995). Warm-glow versus cold-prickle: The effects of positive and negative framing on cooperation in experiments. *The Quarterly Journal of Economics*, *110*(1), 1–21. https://doi.org/10.2307/2118508

Andrighetto, G., Brandts, J., Conte, R., Sabater-Mir, J., Solaz, H., & Villatoro, D. (2013). Punish and voice: Punishment enhances cooperation when combined with norm-signalling. *PLoS ONE*, *8*(6), e64941. https://doi.org/10.1371/journal.pone.0064941

Andrighetto, G., Zhang, N., Ottone, S., Ponzano, F., D'Attoma, J., & Steinmo, S. (2016). Are some countries more honest than others? Evidence from a tax compliance experiment in Sweden and Italy. *Frontiers in Psychology: Cognitive Science*, *7*, 472. https://doi.org/10.3389/fpsyg.2016.00472

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*(2), 150–166. https://doi.org/10.1177/1088868306294907

Ashton, M. C., & Lee, K. (2008). The HEXACO model of personality structure and the importance of the H factor. *Social and Personality Psychology Compass*, *2*(5), 1952–1962. https://doi.org/10.1111/j.1751-9004.2008.00134.x

Bacharach, M. (1999). Interactive team reasoning: A contribution to the theory of co-operation. *Research in Economics*, *53*(2), 117–147. https://doi.org/10.1006/reec.1999.0188

Balliet, D., Parks, C., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes & Intergroup Relations*, *12*(4), 533–547. https://doi.org/10.1177/1368430209105040

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*(3), 412–414. https://doi.org/10.1098/rsbl.2006.0509

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. (2017). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190622046.001.0001

Bicchieri, C., & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, *23*(2), 161–178. https://doi.org/10.1002/bdm.648

Bicchieri, C., Dimant, E., Gächter, S., & Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, *132*, 59–72. https://doi.org/10.1016/j.geb.2021.11.012

Bicchieri, C., Lindemans, J. W., & Jiang, T. (2014). A structured approach to a diagnostic of collective practices. *Frontiers in Psychology*, *5*, 1418. https://doi.org/10.3389/fpsyg.2014.01418

Bicchieri, C., & Xiao, E. (2009). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making*, *22*(2), 191–208. https://doi.org/10.1002/bdm.621

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166–193. https://doi.org/10.1257/aer.90.1.166

Capraro, V., & Vanzo, A. (2019). *The power of moral words: Loaded language generates framing effects in the extreme dictator game* (SSRN Scholarly Paper ID 3186134). Social Science Research Network. https://papers.ssrn.com/abstract=3186134

Cialdini, R., Reno, R. R., & 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. https://doi.org/10.1037/0022-3514.58.6.1015

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, *31*(1), 169–193. https://doi.org/10.1146/annurev.ps.31.020180.001125

de Vries, R. E., Tybur, J. M., Pollet, T. V., & van Vugt, M. (2016). Evolution, situational affordances, and the HEXACO model of personality. *Evolution and Human Behavior*, *37*(5), 407–421. https://doi.org/10.1016/j.evolhumbehav.2016.04.001

Dreber, A., Ellingsen, T., Johannesson, M., & Rand, D. G. (2013). Do people care about social context? Framing effects in dictator games. *Experimental Economics*, *16*(3), 349–371. https://doi.org/10.1007/s10683-012-9341-9

Ellingsen, T., Johannesson, M., Mollerstrom, J., & Munkhammar, S. (2012). Social framing effects: Preferences or beliefs? *Games and Economic Behavior*, *76*(1), 117–130. https://doi.org/10.1016/j.geb.2012.05.007

Elster, J. (1989). *The cement of society: A survey of social order*. Cambridge University Press.

Elster, J. (2009). Norms. In P. Hedström & P. Bearman (Eds.), *The Oxford handbook of analytical sociology*. Oxford University Press.

Engel, C., & Rand, D. G. (2014). What does "clean" really mean? The implicit framing of decontextualized experiments. *Economics Letters*, *122*(3), 386–389. https://doi.org/10.1016/j.econlet.2013.12.020

Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: A field experiment. *Evolution and Human Behavior*, *32*(3), 172–178. https://doi.org/10.1016/j.evolhumbehav.2010.10.006

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, *90*(4), 980–994. https://doi.org/10.1257/aer.90.4.980

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151

Fischbacher, U. (2007). Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178. https://doi.org/10.1007/s10683-006-9159-4

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—An experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547. https://doi.org/10.1111/jeea.12014

Gerlach, P., & Jaeger, B. (2017). Cooperation needs interpretation. *A meta-analysis on context frames in social dilemma games*. https://doi.org/10.17605/OSF.IO/27U8Y

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*, 1–44. https://doi.org/10.1037/bul0000174

Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior & Organization*, *93*, 293–300. https://doi.org/10.1016/j.jebo.2013.03.025

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125. https://doi.org/10.1007/s40881-015-0004-4

Guerra, A., & Harrington, B. (2018). Attitude–behavior consistency in tax compliance: A cross-national comparison. *Journal of Economic Behavior & Organization*, *156*, 184–205. https://doi.org/10.1016/j.jebo.2018.10.013

Hardin, G. (1968). The tragedy of the commons. *Science*, *162*(3859), 1243–1248. https://doi.org/10.1126/science.162.3859.1243

Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making*, *13*, 356–371. https://doi.org/10.1017/S1930297500009232

Hilbig, B. E. (2022). Personality and behavioral dishonesty. *Current Opinion in Psychology*, *47*, 101378. https://doi.org/10.1016/j.copsyc.2022.101378

Hilbig, B. E., Glöckner, A., & Zettler, I. (2014). Personality and prosocial behavior: Linking basic traits and social value orientations. *Journal of Personality and Social Psychology*, *107*(3), 529–539. https://doi.org/10.1037/a0036074

Hilbig, B. E., Kieslich, P. J., Henninger, F., Thielmann, I., & Zettler, I. (2018). Lead us (not) into temptation: Testing the motivational mechanisms linking honesty–humility to cooperation. *European Journal of Personality*, *32*, 116–127. https://doi.org/10.1002/per.2149

Hilbig, B. E., Moshagen, M., Thielmann, I., & Zettler, I. (2022). Making rights from wrongs: The crucial role of beliefs and justifications for the expression of aversive personality. *Journal of Experimental Psychology: General*, *151*, 2730–2755. https://doi.org/10.1037/xge0001232

Hilbig, B. E., & Zettler, I. (2009). Pillars of cooperation: Honesty–humility, social value orientations, and economic behavior. *Journal of Research in Personality*, *43*(3), 516–519. https://doi.org/10.1016/j.jrp.2009.01.003

Hilbig, B. E., Zettler, I., & Heydasch, T. (2012). Personality, punishment and public goods: Strategic shifts towards cooperation as a matter of dispositional honesty–humility. *European Journal of Personality*, *26*(3), 245–254. https://doi.org/10.1002/per.830

Hilbig, B. E., Zettler, I., Leist, F., & Heydasch, T. (2013). It takes two: Honesty–humility and agreeableness differentially predict active versus reactive cooperation. *Personality and Individual Differences*, *54*(5), 598–603. https://doi.org/10.1016/j.paid.2012.11.008

Hilbig, B. E., Zettler, I., Moshagen, M., & Heydasch, T. (2013). Tracing the path from personality—via cooperativeness—to conservation. *European Journal of Personality*, *27*, 319–327. https://doi.org/10.1002/per.1856

Isler, O., & Gächter, S. (2022). Conforming with peers in honesty and cooperation. *Journal of Economic Behavior & Organization*, *195*, 75–86. https://doi.org/10.1016/j.jebo.2021.12.026

Kölle, F., & Quercia, S. (2021). The influence of empirical and normative expectations on cooperation. *Journal of Economic Behavior & Organization*, *190*, 691–703. https://doi.org/10.1016/j.jebo.2021.08.018

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, *24*(1), 183–214. https://doi.org/10.1146/annurev.soc.24.1.183

Kristal, A. S., Whillans, A. V., Bazerman, M. H., Gino, F., Shu, L. L., Mazar, N., & Ariely, D. (2020). Signing at the beginning versus at the end does not decrease dishonesty. *Proceedings of the National Academy of Sciences*, *117*(13), 7103–7107. https://doi.org/10.1073/pnas.1911695117

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Liberman, V., Samuels, S. M., & Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and Social Psychology Bulletin*, *30*(9), 1175–1185. https://doi.org/10.1177/0146167204264004

Mischkowski, D., & Glöckner, A. (2016). Spontaneous cooperation for prosocials, but not for proselfs: Social value orientation moderates spontaneous cooperation behavior. *Scientific Reports*, *6*, 21555. https://doi.org/10.1038/srep21555

Moshagen, M., Hilbig, B. E., & Zettler, I. (2018). The dark core of personality. *Psychological Review*, *125*, 656–688. https://doi.org/10.1037/rev0000111

Murphy, R. O., Ackerman, K. A., & Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771–781. https://doi.org/10.1017/S1930297500004204

Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups* (Revised ed.). Harvard University Press.

Ostrom, E. (2005). Understanding institutional diversity. Understanding Institutional Diversity. http://digitalcommons.usu.edu/unf_research/54

Przepiorka, W., Szekely, A., Andrighetto, G., Diekmann, A., & Tummolini, L. (2022). How norms emerge from conventions (and change). *Socius*, *8*. https://doi.org/10.1177/23780231221124556

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425. https://doi.org/10.1016/j.tics.2013.06.003

Schram, A., & Charness, G. (2015). Inducing social norms in laboratory allocation choices. *Management Science*, *61*(7), 1531–1546. https://doi.org/10.1287/mnsc.2014.2073

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Soraperra, I., Weisel, O., & Ploner, M. (2019). Is the victim Max (Planck) or Moritz? How victim type and social value orientation affect dishonest behavior. *Journal of Behavioral Decision Making*, *32*(2), 168–178. https://doi.org/10.1002/bdm.2104

Steinmo, S., & D'Attoma, J. (2021). *Willing to pay? A reasonable choice approach*. Oxford University Press. https://doi.org/10.1093/oso/9780198796824.001.0001

Szekely, A., Lipari, F., Antonioni, A., Paolucci, M., Sánchez, A., Tummolini, L., & Andrighetto, G. (2021). Evidence from a long-term experiment that collective risks change social norms and promote cooperation. *Nature Communications*, *12*(1), 5452. https://doi.org/10.1038/s41467-021-25734-w

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *The Journal of Applied Psychology*, *88*(3), 500–517. https://doi.org/10.1037/0021-9010.88.3.500

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, *34*(4), 397–423. https://doi.org/10.1006/jrpe.2000.2292

Zettler, I., Hilbig, B. E., & Heydasch, T. (2013). Two sides of one coin: Honesty–humility and situational factors mutually shape social dilemma decision making. *Journal of Research in Personality*, *47*(4), 286–295. https://doi.org/10.1016/j.jrp.2013.01.012

Zhang, N., Andrighetto, G., Ottone, S., Ponzano, F., & Steinmo, S. (2016). "Willing to pay?" Tax compliance in Britain and Italy: An experimental analysis. *PLoS ONE*, *11*(2), e0150277. https://doi.org/10.1371/journal.pone.0150277

Zhao, K., Ferguson, E., & Smillie, L. D. (2016). Prosocial personality traits differentially predict egalitarianism, generosity, and reciprocity in economic games. *Frontiers in Psychology*, *7*, 1137. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01137

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Szekely, A., Bruner, D., Steinmo, S., Todor, A., Volintiru, C., & Andrighetto, G. (2023). Preferences for honesty can support cooperation. *Journal of Behavioral Decision Making*, *36*(4), e2328. https://doi.org/10.1002/bdm.2328