Trust and Trustworthiness in the Villain's Dilemma: Collaborative Dishonesty with Conflicting Incentives?

27th September 2024

Giulia Andrighetto[1,2,3], Andrej Angelovski[4], Daniela di Cagno[5], Francesca Marazzi[6], and Aron Szekely[1,7*]

[1] Institute of Cognitive Sciences and Technologies, Italian National Research Council, Rome, Italy
[2] Institute for Futures Studies, Stockholm, Sweden
[3] Malardalens University, Vasteras, Sweden
[4] Middlesex University, London, UK
[5] LUISS Guido Carli, Rome, Italy
[6] University of Rome Tor Vergata, Rome, Italy
[7] Collegio Carlo Alberto, Turin, Italy
* Corresponding author: aron.szekely@carloalberto.org; telephone: +39 011 15630843

Abstract

Wrong-doers may try to collaborate to achieve greater gains than would be possible alone. Yet potential collaborators face two issues: they need to accurately identify other cheaters and trust that their collaborators do not betray them when the opportunity arises. These concerns may be in tension, since the people who are genuine cheaters could also be the likeliest to be untrustworthy. We formalise this interaction in the "villain's dilemma" and use it in a laboratory experiment to study three questions: what kind of information helps people to overcome the villain's dilemma? Does the villain's dilemma promote or hamper cheating relative to individual settings? Who participates in the villain's dilemma and who is a trustworthy collaborative cheater? We find that information has important consequences for behaviour in the villain's dilemma. Public information about actions is important for supporting collaborative dishonesty, while more limited sources of information lead to back-stabbing and poor collaboration. We also find that the level of information, role of the decision maker, and round of the experiment affect whether dishonesty is higher or lower in the villain's dilemma than in our individual honesty settings. Finally, individual factors are generally unrelated to collaborating but individual dishonesty predicts untrustworthiness as a collaborator.

Keywords: Honesty, Corruption, Coordination, Trust game, Villain's dilemma

The replication material for the study is available at https://doi.org/10.17605/OSF.IO/62TYP

Giulia Andrighetto ORCID: https://orcid.org/0000-0002-3896-1363
Andrej Angelovski ORCID: https://orcid.org/0000-0003-3011-8002
Francesca Marazzi ORCID: https://orcid.org/0000-0002-1229-0661
Aron Szekely ORCID: https://orcid.org/0000-0001-5651-4711

1. Introduction

Imagine that you are a bank cashier and you want to swindle your employer. Ideally, you would do this alone, yet, you cannot. You need help to access customer details, create and approve fraudulent payments, and bypass security protocols. Aside from these hurdles, banks also take specific precautionary measures against cheating by employing the "four eyes principle", requiring that two employees approve the same decision or transaction.[1] So how do you find a cheater to collaborate with and avoid those who are unlikely to help? And, if you do find a willing partner in crime, how can you trust that they won't double-cross you and take all the loot for themselves?

This scenario is not only an imaginary one. Insider fraud is surprisingly common in a wide range of industries including banking and finance and is often carried out by more than one perpetrator. A recent report from the Association of Certified Fraud Examiners (ACFE, 2020) estimated that $3.6billion, likely a vast underestimate, are lost annually from internal fraud. Additionally, a majority of these frauds (51.4%) were conducted by collaborators and banking and finance constitute the single largest sector in which their cases occur (15.4% of cases). Also consistent with this are KPMG's (2016) findings of 750 analysed fraudsters, 62% of whom colluded with others.

Consider now again the situation faced by our bank cashier, but this time through an analytical lens. Would-be corrupt collaborators have to overcome two issues. First is *selection*: they need to identify genuine cheaters as partners and avoid unwitting rule-abiding citizens on whom, at best, their efforts are wasted or, at worst, would report them to authorities. Second is *incentives*: there are clear incentives for collaborators to cheat each other whenever possible to get all of the windfall. Worst of all, for collaborators, the solutions to these two issues are potentially in opposition. The very people likeliest to be genuine cheaters, and thus willing to collaborate, may also be the same people who are likeliest to cheat the other.

Gambetta identifies these factors as the "villain's paradox" and convincingly argues that it is a central problem that criminals working together need to overcome (Gambetta, 2009a, p. 30). We test whether this tension also holds in the less severe instances of rule-breaking. To do so,

[1] A physical implementation of this is with dual locks: locks that require two people to operate with separate codes or keys. Reportedly, dual padlocks were developed in Soviet Russia, requiring two separate keys to unlock, in attempt to reduce the rampant stealing and corruption.

we propose a game, the *villain's dilemma*, which captures the essence of the villain's paradox—players collaborate to cheat the experimenter but have incentives to double-cross each other—and use it in a laboratory experiment to study three related research questions.

Our first research question tackles the villain's paradox directly and seeks to understand some of the conditions, focusing on the role of information, under which people can solve the problem successfully. Specifically, we ask:

> *Are people able to solve the villain's dilemma and cooperate with little information about each other or is extensive information necessary to facilitate collaboration?*

We also use our experiment to study two other, broader, research questions. Prior research indicates that having the opportunity to cheat collaboratively, as opposed to alone, increases cheating (Gross et al., 2018; Leib et al., 2021; Weisel & Shalvi, 2015). Yet this result rests on non-conflicting incentives. That is, it puts to one side the key issue that collaborators who cheat the system have incentives to cheat each other thereby adding risk and a problem of trust into the interaction. Here we ask if this result still holds given the more realistic situation that corrupt collaborators face:

> *Does the villain's dilemma promote or hamper cheating relative to individual settings?*

Finally, little is known about the characteristics of people involved in collaborative cheating. There are some associations between demographic factors and collaborative cheating (e.g. men are overrepresented amongst fraudsters), yet it is unclear whether this is due to differences in incentives, selection (e.g. men being caught more often), or whether such associations represent causal relationships. Similarly, little is known about the characteristics of trustworthy collaborators. Thus, we also ask:

> *Who participates in the villain's dilemma and who is an untrustworthy collaborative cheater?*

Our experiment consists of four stages and begins by putting subjects through two well-established tasks used to measure honesty, or a willingness to cheat, in individual contexts: the die-roll task (Fischbacher & Föllmi-Heusi, 2013) and the sender-receiver task (Gneezy, 2005). Subjects are then allocated into fixed groups of six and participate in the villain's dilemma over multiple rounds during which we observe how frequently and with what outcome, collaboration occurs.

In three between-subjects treatments we vary the amount of information that potential collaborators have about each other in the villain's dilemma. As such, we vary the institutional setting in which participants interact. In the treatment with lowest information available, matched subjects are informed about the prior reported die-roll of their current partner. This information set does not allow subjects to build cumulative knowledge about specific people, but nevertheless gives them some information about how (dis)honest their current partner is likely to be (*Dyadic history no ID*) and allows them to develop a general sense of how trustworthy people in their group are. In the intermediate information treatment, subjects are informed about their current partner's prior action *and* their identifier. Here group members can accumulate knowledge about individual-specific honesty profiles every time they participate in the villain's dilemma (*Dyadic history with ID*). In the third and most information rich setting, participants are shown the prior reported die-roll and identifier of all of their group members (*Public history*). All treatments are complemented by a final questionnaire regarding participants' demographic and personal characteristics.

2. Literature review

2.1. Solving the villain's dilemma with reputation

In addition to identifying and highlighting the villain's paradox, Gambetta (2009a) proposes potential solutions. To solve the issue of selection, he argues that reliable signs and signals of one's "criminal-ness", should be used by those looking for accomplices (Gambetta, 2009b; Spence, 1973, 1974). And to solve the issue of backstabbing and lack of trust he considers two understudied solutions. One is displays of incompetence: showing that one lacks other, more legal, possibilities. The other is the mutual exchange of compromising information. By exchanging information that could harm each other, criminals can threaten each other and shape each other's incentives to collaborate and thus ensure trust. Indeed, a fascinating experiment finds evidence that students use the exchange of compromising information to ensure trust in a variant of the trust game (Gambetta & Przepiorka, 2019).

But there is another classic, more widely known solution, that is not peculiar to the criminal context: reputation. Reputation is an evaluation of other individuals based on their skills and past actions (Giardini et al., 2019; Milinski, 2016; Romano et al., 2021; Számadó et al., 2021). Individuals are frequently motivated to gain and maintain a good reputation as it is seen as a "universal currency" for future social exchange (Milinski, 2016) and reputation is known to be a powerful motivator of cooperation both through indirect reciprocity (Rand & Nowak, 2013) and in partner choice (Roberts et al., 2021). As Gambetta highlights, reputation is also the most straight-forward way of solving the paradox: criminals should *behave well* and live up to their promises to establish a reputation for trustworthiness just as an ordinary business person would do. By doing so, interests will become aligned to good practice, and one can stop worrying about good character (Gambetta 2009b, p. 39).[2] However, reputations for cooperation and for dishonest collaboration differ. Reputation for honest collaboration signals one's willingness to sacrifice individual utility to confer benefits on others (Barclay, 2016). This attracts help from others, even from strangers or out group members (Milinski, 2016; Nowak & Sigmund, 2005; Wu et al., 2016). While a reputation for dishonest collaboration may differ since it could also reveal features that have to do with the dark side of personality, such as "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). Holding a reputation for dishonest collaboration may thus send a mixed message to a potential partner in a dishonest activity.

2.2. Does the villain's dilemma promote or hamper dishonesty?

Multiple past experiments have found that collaborating increases cheating or dishonesty (e.g. Conrads et al., 2013; Gross et al., 2018; Sutter, 2009; Weisel & Shalvi, 2015). Weisel and Shalvi's (2015) seminal paper, puts subjects in a sequential die-rolling task in which the collaborators' earn money by cheating the experimenter by both reporting the same die-roll, and finds that collaborative corruption dominates as matched die-rolls are reported vastly more than by chance. Moreover, dishonesty is higher than in an individual variant of the same task indicating that collaboration can "liberate people to lie more than when they work alone" (p. 10653). Similarly, Gross et al.'s (2018) experiment on "ethical free riding" allows participants to select partners in a setting in which collaborative dishonesty is possible. They find that both

[2] Quantitative work on illicit drug markets backs up this notion, with reputation facilitating cooperation (Przepiorka et al., 2017).

honest and dishonest individuals abuse collaborative dishonesty by attempting to partner, or to remain partnered with, dishonest individuals.[3] Yet, here too, incentives among collaborators are aligned (i.e. partners match die-rolls to earn that amount) and no incentives to cheat each other. We are unaware of studies implementing incentive conflict among potential collaborating cheaters (see Leib et al., 2021 for a review); the consequences of cheaters having incentives to cheat each other has remained understudied. As such, whether this will also happen in the villain's dilemma is unclear since a fear of being cheated may decrease collaborative dishonesty.

From the classical economics framework, in which individuals have purely self-regarding preferences, people in our collaborative setting should avoid collaborative cheating as they should expect their collaborator to cheat (see Section 3.2 for more details). As such, incentives are for people to be entirely dishonest in the individual setting and stay out of dishonest tasks in the collaborative settings.

Behaviourally-motivated theories instead propose, and find evidence, that people trade-off monetary incentives with internal costs of lying and honest image concerns (Abeler et al., 2019; Cohn et al., 2019; Weisel & Shalvi, 2021). A pure lying-cost approach would imply no change in dishonesty between individual and collaborative corruption settings since everyone can make similarly consequential lies. While an image concern framework could imply both higher or lower dishonesty: subjects may prefer to be seen as non-corrupt, yet they may also want a reputation for corruption since they are likely to only match other corrupt individuals.

2.3. Who participates in collaborative corruption and who is a trustworthy corrupt collaborator?

Little is systematically known about the predictors of collaborative corruption. To help us, we start with what we know from experimental studies. While many structural factors have been found to shape corrupt collaboration, e.g. similarity of interactants (Irlenbusch et al., 2020) or sequential vs. simultaneous decision (Rilke et al., 2021), or the availability of the same participants from round to round (see, for example, Abbink, 2004; Bühren, 2020), individual

[3] In case of the honest individuals this is done by honest first movers who are matched with dishonest second movers (as per the Weisel & Shalvi, 2015 design). The honest first movers report truthfully the rolled number but do not change partners when the dishonest second movers match their dice roll, i.e. honest individuals seem to be "ethically free-riding".

factors remain little studied. Drawing on a recent meta-analysis of collaborative dishonesty (Leib et al., 2021), the only evidence we have concerns gender and age. Women, or women-only groups, are mildly more honest than men, or men only and mixed groups, in collaborative corruption settings (Conrads et al., 2013; Muehlheusser et al., 2015). While for age, Conrads et al. (2013) report that older subjects are less dishonest, however, this association is not robust after controlling for personality characteristics. There is also recent evidence showing that Honesty-Humility, a factor in the HEXACO model of personality (Ścigała et al., 2019; Thielmann et al., 2024; Zettler et al., 2020), is negatively correlated with individual dishonesty while no robust association is found with the Big Five measures (Hilbig, 2022).

Concerning colluders, KPMG's (2016) case analysis finds that fraudsters who collude tend to be more senior employees and to have worked longer at the company than the solo fraudsters. This suggests that older employees are likelier to collude which may be due to time spent at the organisation rather than their age and is in line with the experimental work on staff-rotation as an anti-corruption tool (Abbink, 2004). KPMG's data matches the experimental results for gender: men are likelier to collaborate than women (66% vs. 45% respectively). An ACFE 2020 survey[4] similarly finds that (*i*) managers and owners/executives comprise a majority of the fraudsters (55%), (*ii*) the modal time (46%) that fraudsters had worked for a company was 1-5 years, (*iii*) young and old perpetrators are least represented in the survey while intermediate ages (≈31-50) are most represented, and (*iv*) that men are overrepresented relative to women (72% vs. 28%). But there are important caveats with these data: we do not know the base rates in the larger population, making it difficult to know whether these simply reflect composition or if there is genuine selection. Moreover, it remains unclear whether the cases reflect real differences in behaviour, which may be driven by different incentives, or simply differences in being caught. Our experiment avoids these issues by controlling incentives and monitoring all participants equally.

Our third source of insight comes from the extensive literature on individual dishonesty (Rosenbaum et al., 2014) where a slight tendency for women to behave more honestly than men has been found (e.g. Dreber & Johannesson, 2008; Gibson et al., 2013), although not unanimously (e.g. Fries et al., 2021; Gylfason et al., 2013; Hanna & Wang, 2017). Moreover,

[4] The ACFE survey does not separate between colluders and non-colluders, but, since the former comprise a large proportion of their sample even the overall statistics can give us some indications.

whatever differences there are may be affected by subtle changes such as stake dependence (Childs, 2012) or the consequences of the lie, whether harming or helping others (Erat & Gneezy, 2012). Conversely, there is little evidence that age is systematically associated with honesty. We do not know of any direct evidence about who makes a trustworthy collaborator.

3. Materials and methods

3.1. Overview

Subjects in our study participated in four stages (Table 1; see Supplementary Material for instructions and screenshots). They received the instruction for each stage only at the end of the preceding one. In Stage 1, participants repeatedly play (10 times) the die-rolling task (Fischbacher & Föllmi-Heusi, 2013). Subjects roll a 6-sided die privately and are told to report the number that comes up, with higher numbers leading to higher payoffs (i.e. 1 = 1 ECU, 2 = 2 ECU, 3 = 3 ECU, 4 = 4 ECU, 5 = 5 ECU, and 6 = 6 ECU). By asking subjects to make 10 separate decisions we generate extensive information about their individual (dis)honesty decisions and reduce the noise that is inherent in this task. While they know that they are paid for one randomly drawn round from Stage 1, they are only told which decision was chosen at the end of the experiment.

Stage 2 implements another individual honesty task: the sender-receiver task (Gneezy, 2005). The computer randomly pairs subjects and assigns one the role of sender and the other the receiver. The sender receives private information about a payoff matrix in which only the actions of the receiver can influence the resulting outcome. One of the receiver's actions benefits the sender (who earns 2 ECU) at a cost to the receiver (who earns 1 ECU) while the reverse is true for the other (sender earns 1 ECU and the receiver earns 2 ECU). The sender chooses one of two messages, one which is true and the other is false, to send to the receiver about what action he or she should take. Crucially, the sender has incentives to deceive the receiver and the receiver knows this.

Stage 3 is the villain's dilemma, which we describe in detail below. Participants played this for 30 rounds. In the final phase, Stage 4, subjects answer a questionnaire in which we elicit their demographics, self-reported trust and cheating measures[5], self-reported risk preferences[6] cognitive reflection (Frederick, 2005), and Big Five personality characteristics using the 10-item inventory (Rammstedt & John, 2007).[7]

Table 1. Experimental protocol summary.

Stage	Task
1	Die-rolling honesty elicitation x 10 (Fischbacher & Föllmi-Heusi, 2013)
2	Sender-receiver task (Gneezy et al., 2013)
3	The villain's dilemma x 30 rounds
4	Questionnaire

Notes: The questionnaire contained items about demographics, self-reported risk preferences, the Cognitive Reflection Test (Frederick, 2005), and the Big Five (Rammstedt & John, 2007).

We use the first two stages of our experiment (the die-roll task and sender-receiver task) to measure subjects' individual-level behavioural tendencies concerning honesty and to identify cheating in non-collaborative honesty tasks. Specifically, we use the information gleaned from them to test whether cheating is higher in the villain's dilemma or in the individual settings and to identify predictors of participation in the dilemma and trustworthiness as a collaborative cheater. We use two measures since each have distinct features (Gerlach et al., 2019; Soraperra et al., 2019). Die-rolling is an incentivised measure of (dis)honesty in which it is impossible for the experimenters to identify lying at the individual-level, and hence subjects should not be worried about being caught. Additionally, dishonesty imposes costs on the experimenter and not on other subjects. The sender-receiver task is also incentivised but, it is possible to identify dishonesty at the individual-level which may shape subjects' behaviour, and the consequence of lying impose costs on other subjects (instead of the experimenter). By measuring and studying both, we can gain a broader picture of the relationship between individual dishonesty and collaborative dishonesty in the villain's dilemma. Moreover, since dishonest collaboration

[5] These questions ask: "You left your watch in a toilet, do you think you are going to find it there?", "you are having trouble solving an exercise during an exam" in both known and unknown contexts (at their university and in the airport; peeking at the exam of your friend and an unknown student). We do not use these variables in the analyses however, as they show little variation.

[6] Subjects are asked the following question "Which amount of money makes you indifferent between receiving that amount of money for sure and participating in a lottery where you can win 0 with 50% probability and 100 with 50% probability?" and have three possible alternatives: *a)* 50, *b)* an amount higher than 50, *c)* an amount lower than 50.

[7] At the time we designed our experiment, results concerning Honesty-Humility and collaborative dishonesty were not yet widely available.

in the villain's dilemma is costly to the experimenter and not to other subjects, like in the die-rolling task, this allows us to make clearer comparisons across the two.

The experiment was programmed in z-Tree software (Fischbacher, 2007). The experimental design and procedures are compliant with LUISS University's rules and it received ethical approval from the CESARE lab (Supplementary Materials, Section 4). Written consent was obtained from all participants.

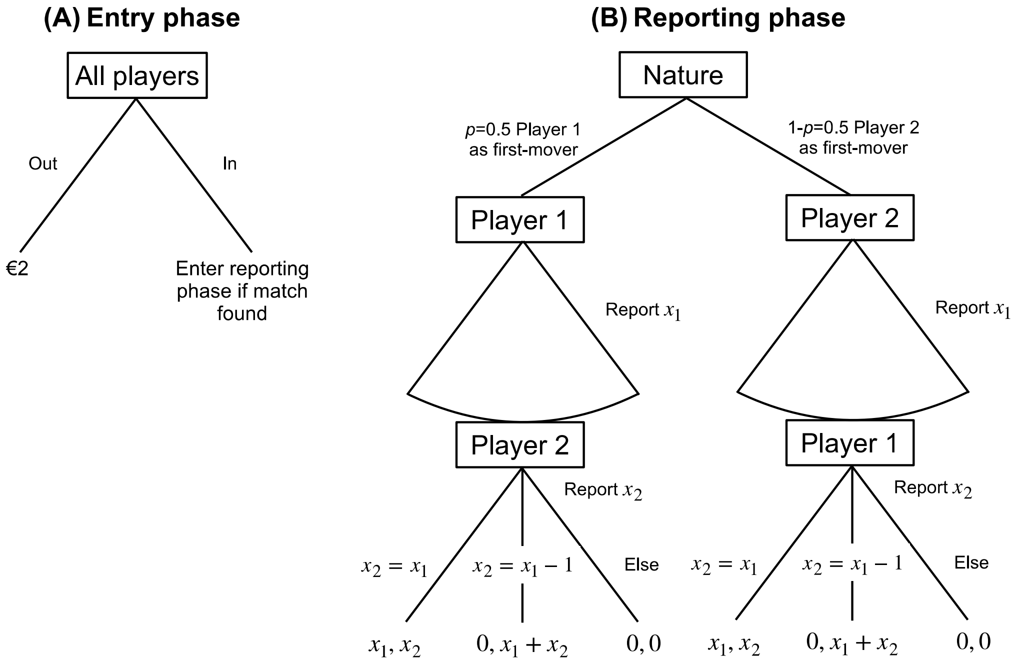
3.2. The villain's dilemma

The villain's dilemma is implemented in two phases: an entry phase and a reporting phase. Participants first decide (Figure 1A) whether to participate in the villain's dilemma, in which case they can earn between 0 ECU and 11 ECU—depending upon the outcome of the interaction—or to stay out, in which case they earn a fixed amount of 2 ECU. If participants decide to enter, they are paired based on their preference ordering (described below), and one of the two is randomly assigned the role of first mover (FM) and the other the role of second mover (SM) (Figure 1B). In the reporting phase, the FM rolls a six-sided die and is asked to report the number x_1, where $x_1 \in \{1,6\}$, that he or she rolls. The SM then observers the number reported by the FM and rolls their six-sided die and is asked to report the number x_2, where $x_2 \in \{1,6\}$, rolled. If the SM's reported number matches the report of the FM ($x_2 = x_1$) then they each earn the amount they reported. If the SM's reported number undercuts the FM's number by 1 ($x_2 = x_1$-1) then the SM keeps the total of their earnings (x_1+x_2) and the FM gets nothing. For any other combination of reported numbers, both players earn nothing.

Since by staying out they earn only 2 ECU, it is attractive for participants to enter the villain's dilemma. Yet whether or not this is truly the case depends upon their, and their partner's, intentions. If they intend to be honest and expect that their partner is also honest then their earnings, in expectation, from the villain's dilemma is a measly 1.07 ECU, less than what they could earn by staying out. Moreover, if we consider the game strategically, and assume standard self-regarding risk neutral preferences, then the subgame perfect Nash equilibrium is also {Stay out 1, Stay out 1}. This is because the FM anticipates that the SM will choose to undercut by 1 for every reported roll, since this maximises the SM's earnings, and so would decide to report 1, which cannot be undercut (reporting any other number by the SM would get both participants the lowest possible earnings of 0 ECU). This, would leave both the FM and the SM with earnings of 1 ECU. Consequently, participants should stay out and thereby earn 2

ECU. Put differently, it only makes instrumental sense to participate in the villain's dilemma if one expects that their partner is likely to be both dishonest, over-reporting high numbers, and trustworthy by matching numbers.

Figure 1. The villain's dilemma.



Participants play the villain's dilemma for 30 rounds in fixed groups of six. Each person within a group is allocated a shape (e.g. star, circle, triangle) and they keep this for the duration of the experiment. In every round, they have the possibility to be matched with another participant from their group. To decide matching, we elicit their preferred matching rank for the other group members, allowing the possibility of considering two or more equally suitable participants in the same group (in case of a tie, one of the participants was randomly chosen). Then, they indicate whether, if a match was found, they would want to collaborate. Put differently, we elicit their preference ranking and then ask whether they want to put their ranking "into action" or to keep their ranking dormant. The reason for eliciting the ranking for all participants is because it allows us to have the same steps for all of them and to avoid a potential demand effect due to being inactive while opting-in participants were stating their ranking.[8]

[8] Sometimes referred to as *action bias* (see Patt & Zeckhauser, 2000 for an experimental investigation).

After stating their preference for collaboration, a random order is selected among the participants who opted to enter, and the participant who is drawn first gets to collaborate with the first person on her ranked list (given that this person also decided to opt in for collaboration, otherwise the second ranked participant is attempted etc.), followed by the second drawn participant who decided to go in, and so on. If a participant wanted to collaborate but no match was possible, (s)he received the 2 ECU flat fee. At the end of the experiment the computer randomly selects for payment one of the 30 rounds and the individual payoff for this stage corresponds to the payoff of that round.

In contrast to the standard collaborative die-rolling scenarios (e.g. Weisel & Shalvi, 2015), our entry phase allows for selection between different would-be collaborators, and, our reporting phase includes the possibility of back-stabbing, thereby adding the component of trust. Yet, and in contrast to a standard trust game (e.g. Berg et al., 1995), the villain's dilemma puts two motivations in tension. By collaborating with another player and earning high amounts, players cheat the experimenter and are thus behaving immorally. However, by collaborating with another player they are also behaving cooperatively, or in a trustworthy way as a SM.

3.3. Experimental treatments

We implemented three between-subjects treatments to observe participants in three institutional settings that vary in the amount of available information. Specifically, we modify the amount of information that participants know about (potential) collaborators in the villain's dilemma and on which they can choose with whom to establish a collaboration. The three treatments listed according to the volume of available information (from lowest to highest), are:

(1) *Dyadic history no ID (Dyadic no ID)*. Participants who decide and actually enter a collaboration receive information about the number their collaborator reported in the prior period but are not aware of his or her past role (i.e. first or second mover), before deciding what to report in the current period. But they do not know the identity (shape) of their collaborator. So, while participants transmit a limited form of history to their partner every time that they are matched, they cannot build up individual-specific behavioural profiles about the others in their group. They are instead limited to estimating a distribution at the group-level or inferring individual behaviour from only the prior round (e.g. assuming that a partner is likely trustworthy if (s)he reported a 6

in the previous round or whether (s)he possibly undercut the previous partner if a 5 was reported).

(2) *Dyadic history with ID (Dyadic ID)*. Participants receive the same information as in Dyadic no ID about the roll that their collaborator reported in the prior period. Furthermore, they are informed about the identifier (shape) of their collaborator. Thus, over time, group members can slowly build up individual-specific behavioural profiles of each other that they take it into account when ranking potential partners. While participants in the Dyadic ID may not be able to perfectly remember the entire history of play (although they could physically keep a tab as a pen and paper were provided) participants do get a general sense of the behaviour of others, e.g. square was cooperative or non-cooperative. We specifically used shapes, instead of numeric identifiers, to help with this. By comparison, in Dyadic no ID it is impossible for participants to associate multi-round behaviour with specific people and so can only make limited individual inferences or update their beliefs about the group.

(3) *Public history (Public)*. Participants receive the same information as in Dyadic ID, but for *all* group members who decided and succeeded in entering collaboration in the previous round. Hence die-rolls are observed publicly and with each identifier. Since the matching between group members is not specified, it is unclear what outcome actually occurred; whether a group member was trustworthy or untrustworthy. Yet subjects' willingness to report high numbers is perfectly clear.

Information might influence (mis)behaviour through multiple mechanisms, among them reputational concerns but also self-reflection or social norms. The main aim of our experiment is to study the consequences of different informational environments and not to disentangle the specific pathways through which these environments shape behaviour.

3.4. Analytic strategy

To understand how much information people need to solve the villain's dilemma, our first research question, we look at three outcomes concerning the villain's dilemma (Stage 3): choosing to enter into collaboration (instead of staying out), the die-rolls that people report once they enter, and the outcomes that emerge from their collaborative interactions. We use two sample *t*-tests, in which each group provides one observation (i.e. 20 to 22 observations per treatment), as a conservative approach to testing differences in means. Additionally, for each outcome we conducted regression modelling in which we include extensive control covariates (Table 4, Table 5, Table A8).

We next attempt to understand whether the villain's dilemma promotes or hampers dishonesty, our second research question, in multiple ways. First, we compare the percentage of fully honest people in stages 1 (individual die-rolling) and 2 (the sender-receiver task) to the percentage of people who decide not to enter the villain's dilemma in Stage 3. Only people who intend to behave, at least somewhat, dishonestly should enter the villain's dilemma, while fully honest people should stay out as staying out gains them 2 ECU while entering and being honest gains them 1.07 ECU in expectation. We test these using paired *t*-tests run on individual averages. Second, we compare the reported die-roll of subjects in Stage 1 relative to their die-rolls in Stage 3 as first movers and second movers. While incentives diverge between Stage 1 die-rolling and Stage 3 die-rolling, our aim is to understand whether the set of factors implemented in the villain's dilemma, and the various treatments that subjects participate in, shape dishonesty relative to individual settings. We test these differences using paired *t*-tests on the individual-level frequencies of reporting 6. We do not use multiple regression analyses here since all comparisons are within subject at different stages of the experiment and are balanced by implication. We study our second research question from multiple angles because the incentives between individual die-rolling and collaborative die-rolling in the villain's dilemma, which we focus on, are not identical. As such, we consider in detail which combination of factors in the villain's dilemma promotes or hampers dishonesty. Importantly, while the incentives are not identical, they are comparable: individually reporting a number in the Stage 1 and partners reporting a number in Stage 2, earns the exact same amount. Moreover, meta-analytic evidence suggests that small differences in stake sizes does not shape dishonesty in the die-rolling task (Gerlach et al., 2019).

Finally, we study individual predictors for participating in collaborative corruption and being an untrustworthy corrupt collaborator, our third research question, using random effects probit regressions with standard errors clustered at the group level for the choice to opt into collaboration in Stage 3 or undercutting as the dependent variable (Table 4, Table 5). For each, we present the results of five different specifications, moving from the simplest one (Model 1), which includes only game-related covariates, to the most complex (Model 5), where we account for extensive individual-level characteristics gathered with our final questionnaire. The aim of this procedure is to both test for individual predictors and to robustly check whether the treatment differences observed in the previous sections (and confirmed with the simplest models) survive when we account for different and increasingly complex set of covariates.

4. Results

We ran our experiment at the LUISS CESARE Lab (Rome, Italy) in presence with student participants recruited with ORSEE (Greiner, 2015) and collected data on 378 subjects (44.97% female, mean age = 21.99, SD=2.56): 120 in Dyadic no ID, 132 in Dyadic ID, and 126 in Public. These translate into 20, 22, and 21 groups respectively. Participants were in undergraduate or postgraduate programs in Economics, Law, or Political Science.[9] Each session lasted around two hours. No subject participated in more than one session. The average payment for each participant was €16.7 euros including a participation fee of €5.

Before turning to our research questions, we briefly describe the individual honesty results of our study. In the individual die-rolling task (Stage 1), we find that people over-report higher die-rolls and under-report lower die-rolls, but, many are not fully income maximisers (see Figure A1). In the sender-receiver task (Stage 2), we also find a mix between honesty and dishonesty: 63% of senders sent an honest message while 37% lied. Correspondingly, 66.1% of receivers trusted the message and 33.9% didn't follow the message. All of this is broadly consistent with existing results (Abeler et al., 2019; Fischbacher & Föllmi-Heusi, 2013; Gneezy, 2005; Rosenbaum et al., 2014).

However, we also find some unexpected variation: somewhat more individually honest decisions are reported in the Dyadic no ID and Dyadic ID treatments than in the Public treatment. The mean reported die-rolls are 3.93 (SD=1.66), 3.95 (SD =1.62), and 4.49 (SD =1.55) respectively (Kolmogorov-Smirnov two-sample tests: Public vs. Dyadic ID: p<0.001; Public vs. Dyadic no ID: p<0.001; Dyadic ID vs. Dyadic no ID: p=0.596; Figure A2). While in the sender-receiver task 70%, 65.2%, and 54% of the messages are honest, respectively

[9] Sessions started in 2019 and were suspended due to the COVID-19 pandemic. An attempt to resume the sessions was made in October 2020 (four sessions, with only two groups each to allow for physical distancing in the laboratory; however, the experimental subjects in these sessions differed substantially in terms of pre-treatment characteristics from subjects of previously ran sessions, and therefore the sessions were suspended again and the eight groups were dropped from the analysis. In Section 4 of the Supplementary Material we replicate the analyses also including the COVID sessions and find the same results. Sessions were later resumed and completed in March 2022. In order to account for differences in behaviour between the 2019 and 2022 sessions, we also present an additional analysis in the Supplementary Material, Section 2, where we restricted our analyses to only the pre-COVID sessions. Our analysis shows that the results are consistent with the full analyses presented later in the paper, see Tables A4 and A5.

(Public vs. Dyadic ID: OR=0.63, p=0.20; Public vs. Dyadic no ID: OR=0.50, p=0.069; Dyadic ID vs. Dyadic no ID: OR=0.80, p=0.56). Although this was unanticipated, we believe that these differences are unlikely to cause issues for inference in the rest of the experiment for five reasons. First, we identify the primary source driving the differences: even though allocation into treatments was randomised, by chance, more experienced subjects participated in the Public treatment sessions (5%, 11.4%, and 35.7% in Dyadic no ID, Dyadic ID, and Public respectively). This led to the lower individual honesty measures as we describe above. Second, the treatments are well-balanced on most other covariates (Table A2). Third, we control for experience and other covariates statistically in multiple regression models including Stage 1 and 2 dishonesty (Table 4, Table 5, Table A8). Fourth, our between-treatment results (which could have been potentially affected) are robust to pre-treatment variation in experience. We do this by checking what *would* happen if we were to make the treatments comparable in terms of pre-stage 3 characteristics. We do this purely as an exercise to test robustness; all analyses in the paper contain the full sample.[10] Fifth, we show later that Stage 1 dishonesty is associated with undercutting in Stage 3 (Section 4.3.2, Table 5). If differences in Stage 1 dishonesty were driving our results then we should observe higher undercutting in Public than in the other two treatments. Yet, we see the exact opposite, with undercutting being the lowest in Public (Section 4.1.3, Figure 5). We now turn to our first substantive research question.

4.1. Are people naturally able to solve the villain's dilemma and cooperate or is credible information necessary to facilitate collaboration?

4.1.1. Entering the villain's dilemma

Across all rounds, 93.7% of participants opted into collaboration in Public, 81.9% in Dyadic ID, and 84% in Dyadic no ID (Table 2). These levels of choosing to collaborate are high and may be explained by an ambiguity over what the appropriate choice is. Put differently, cheating the experimenter in a collaborative setting may seem less immoral and more appropriate as what is taken from the experimenters partly goes to another subject who may be perceived as more in need. The difference in opting into collaborations between Public and Dyadic ID (Opt

[10] Tables A3 and A5 in the Supplementary Material show the results of a robustness test to check what would happen were we to make the treatments comparable in terms of pre-stage 3 characteristics. To this end, in A3 and A5, we remove the groups who, on aggregate, behaved most dishonestly in Stage 1 of the Public treatment and we find identical results from our regression analyses (see Tables A3 and A5). As a threshold to identify groups who behaved most dishonestly, we first compute the fraction of 6s reported in Stage 1 at the group level. We then drop all the groups in treatment Public that have such fraction higher than the maximal fraction in the remaining two treatments, which leads us to remove a total of 8 groups. We then perform the same regression analysis on such restricted sample (13 out of 21 groups for Public and the full sample for the other two treatments).

$\text{in}_{\text{Public}}$-Opt $\text{in}_{\text{Dyadic ID}}$=0.118, p<0.001, d=1.654) and Public and Dyadic no ID (Opt $\text{in}_{\text{Public}}$-Opt $\text{in}_{\text{Dyadic no ID}}$=0.097, p=0.002, d=1.058) are both significant and substantively meaningful. While the difference between the two dyadic treatment (Opt $\text{in}_{\text{Dyadic ID}}$-Opt $\text{in}_{\text{Dyadic no ID}}$=-0.021, p=0.485, d=-0.218) is not significant and not substantive. We find the same results when using probit regressions that control for an extensive range of covariates (Table 4).

Table 2. Frequencies of opting in and of actually realized collaboration

	Public (1)		Dyadic ID (2)		Dyadic no ID (3)		t-test	t-test	t-test
Variable	N	Mean	N	Mean	N	Mean	(1)-(2)	(1)-(3)	(2)-(3)
	[n]	[SD]	[n]	[SD]	[n]	[SD]	[*d*]	[*d*]	[*d*]
Opt in	3780	0.937	3960	0.819	3600	0.840	0.118***	0.097***	-0.021
	[21]	[0.243]	[22]	[0.385]	[20]	[0.367]	[1.654]	[1.058]	[-0.218]
Realized	3780	0.802	3960	0.680	3600	0.716	0.122***	0.085**	-0.036
	[21]	[0.399]	[22]	[0.467]	[20]	[0.451]	[1.385]	[0.773]	[-0.348]

Note: *N* identifies the total number of observations and *n* the number of (independent) groups. t-tests on between-treatment differences are run on group-level averages (thus with *n* observations) to preserve the independence of observations. *d* indicates Cohen's *d*. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

These differences in opting into collaboration translate into actual collaboration differences implying that satisfactory matches are found: 80.2% of subjects in Public enter the villain's dilemma, 68% do so in Dyadic ID, and 71.6% enter in Dyadic no ID.[11] The differences between Public and Dyadic ID ($\text{Realized}_{\text{Public}}$-$\text{Realized}_{\text{Dyadic ID}}$=0.122, p<0.001, d=1.385) and Public and Dyadic no ID ($\text{Realized}_{\text{Public}}$-$\text{Realized}_{\text{Dyadic no ID}}$=0.085, p=0.018, d=0.773) are significant and substantive while the difference between Dyadic ID and Dyadic no ID is not significantly different ($\text{Realized}_{\text{Dyadic ID}}$-$\text{Realized}_{\text{Dyadic no ID}}$=-0.036, p=0.267, d=0.348).

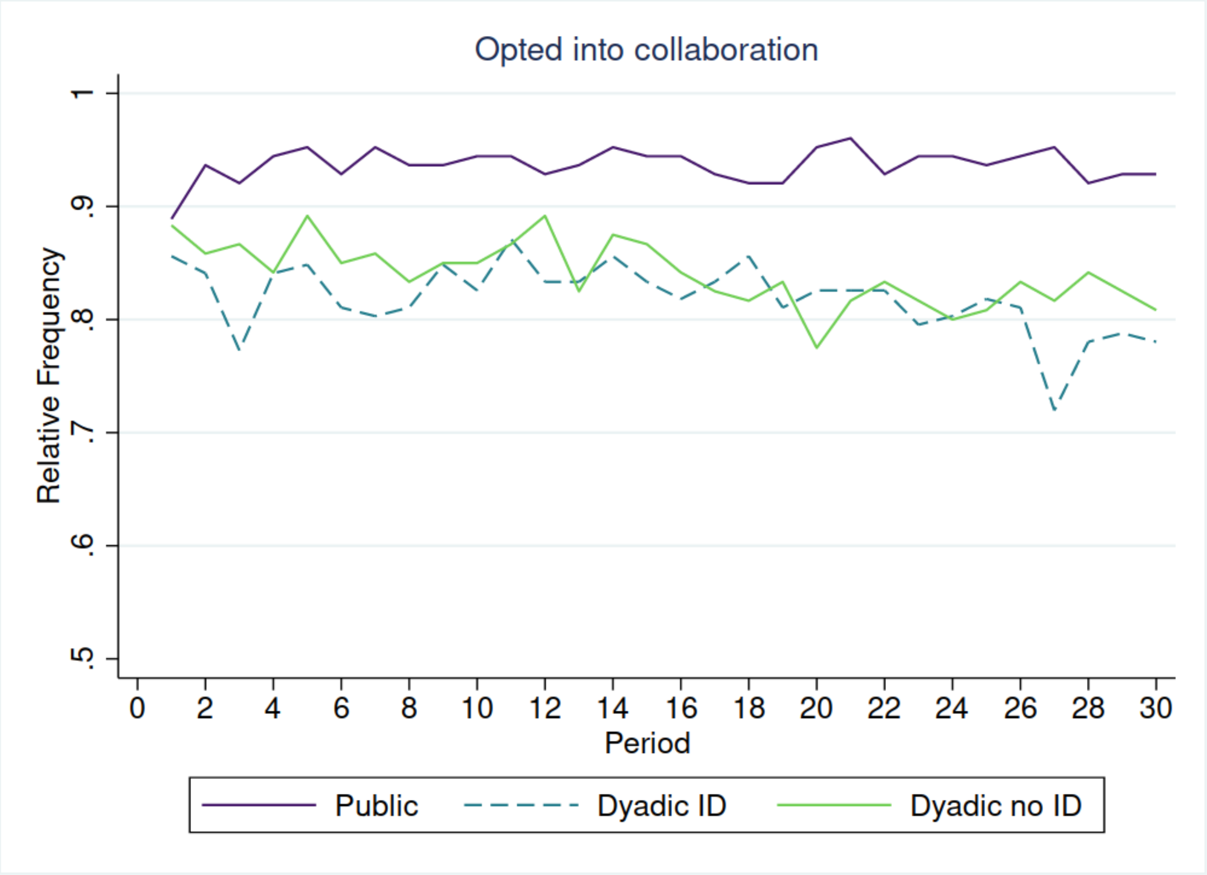
In dynamics too, these between-treatment differences are clear (Figure 2). Choosing to collaborate remains high and stable in Public, while it starts at high levels (albeit a little lower than in Public) in the other treatments and then declines slowly over time, although it is unclear if this would stabilise or continue declining further.[12] This pattern is also reflected in actually entering into collaboration (Figure A3). Although the decline of choosing to collaborate in the

[11] The overall frequency of participants wanting to collaborate but not finding a match is equal to 15.35%.

[12] We find, with Kendall's rank correlations, a declining trend between round and mean opting in for both Dyadic no ID (p= 0.003) and Dyadic ID (p= 0.012), while none in Public (p= 0.619).

Dyadic ID and Dyadic no ID treatments is slow, the Public treatment seems to be more effective in supporting consistently high collaborative corruption (Table A12).

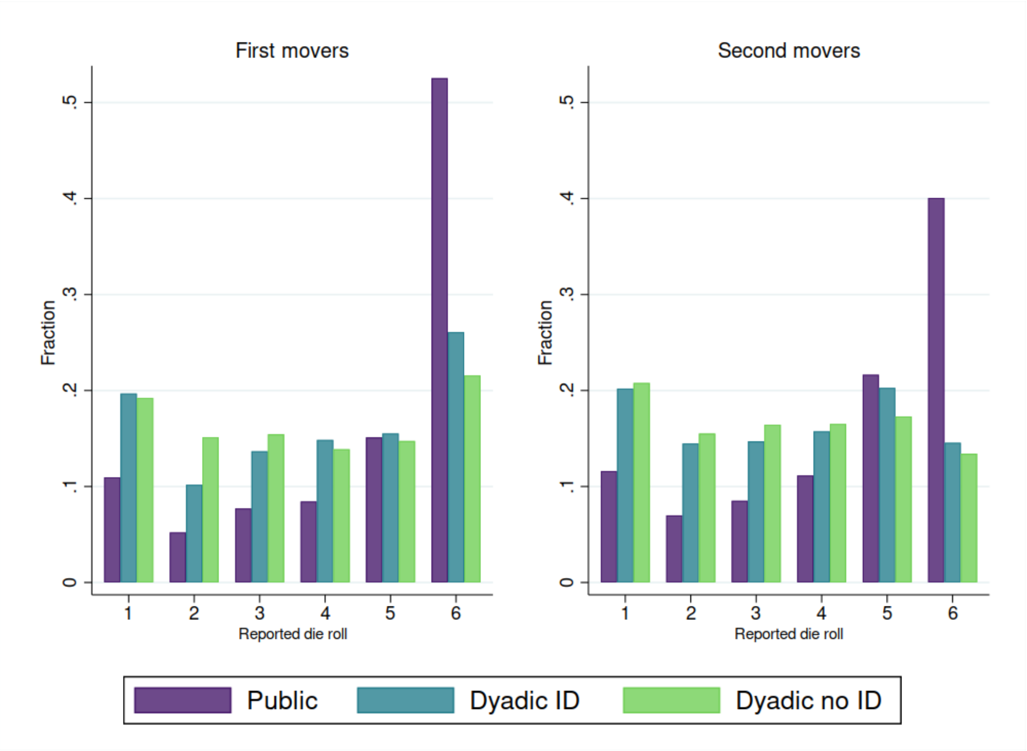
Figure 2. Choosing to collaborate according to treatment



4.1.2. Reported die-rolls in the villain's dilemma

Why is this the case? To understand why collaboration differs across the treatments, consider the actions of both FM (Figure 3, left panel) and SM (Figure 3, right panel; Figure 4) in the villain's dilemma and the subsequent outcomes that emerged (Figure 5). All three figures clearly show what is happening.

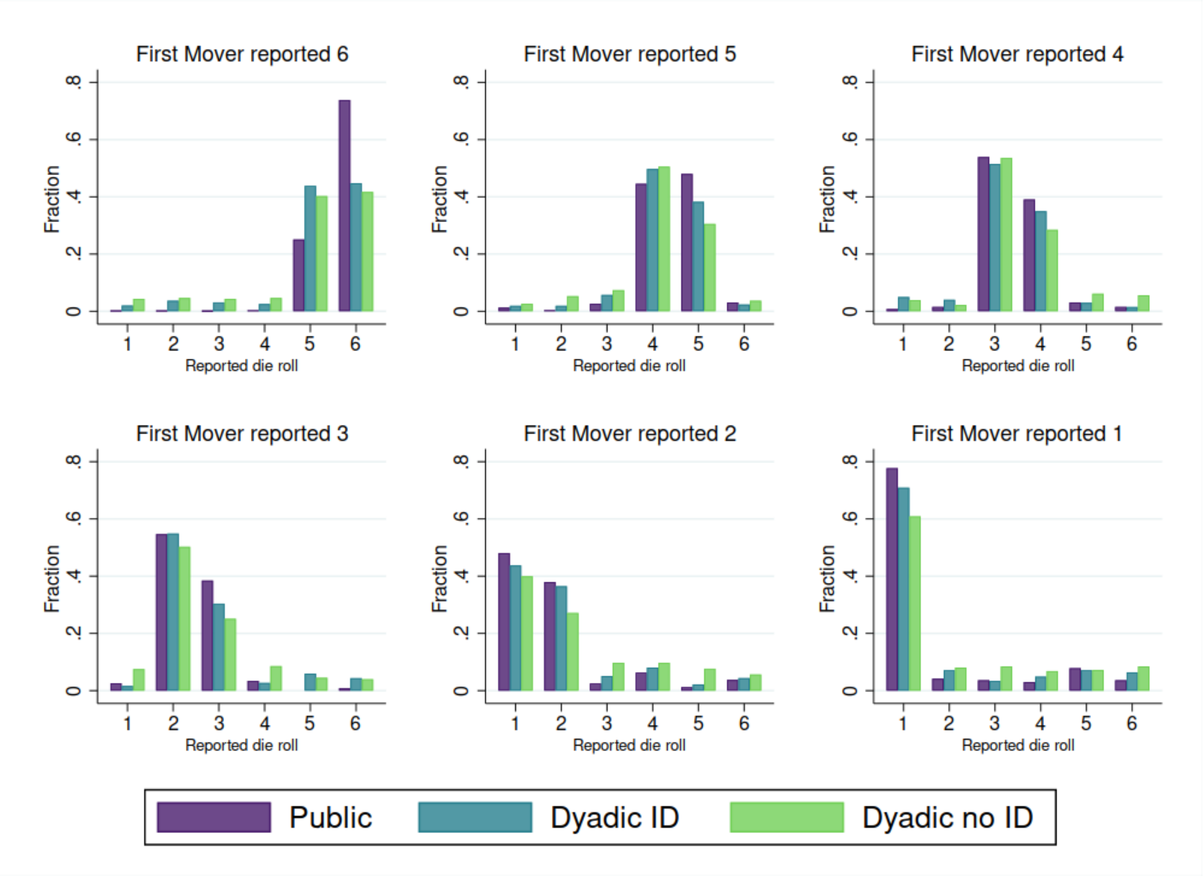
Figure 3. Reported die-roll for first (left panel) and second (right panel) mover by treatment



In the Public treatment, 52.54 % of the reported die-rolls by first movers are 6 and this is largely reciprocated by second movers, among whom 40.07% also report 6 (Figure 3). While there are few reports of 1: 10.96% from the FM and 11.62 from the SM. In contrast to Public, in the Dyadic no ID and Dyadic ID treatments, first movers only report the highest die-roll 21.57% (Public-Dyadic no ID=30.97, *p*<0.001, *d*=2.083) and 26.08% (Public-Dyadic ID=26.46 *p*<0.001, *d*=1.946) of the time and report lower numbers in larger proportions (Figure 3). Indeed, 1s are reported 19.24% (Public-Dyadic no ID=-8.28, *p*=0.002, *d*=-1.032) and 19.69% (Public-Dyadic ID=-8.73, *p*=0.002 *d*=-1.016) of the time, which is substantially higher than in Public and even slightly above what would be expected by chance. Second movers too report 6 infrequently and less than in Public, at 13.42% (Public-Dyadic no ID=-1.8, *p*<0.001, *d*=2.196) and 14.56% (Public-Dyadic ID=-2.94, *p*<0.001, *d*=2.192) and report low numbers in substantial and higher proportions than in Public, with 20.79% (Public-Dyadic ID=-9.17, *p*=0.001, *d*=-1.156) and 20.21% (Public-Dyadic no ID=-8.59, *p*=0.001, *d*=-10.54) reporting 1s.

We find further between treatment differences when we look more carefully at the second movers' die-rolls by separating their reports conditional on first movers' choices. This is particularly evident when the first movers report 6 or 1 (Figure 4). Conditional on first movers reporting 6, 73.87% of second movers in Public reciprocate by reporting 6, while only 44.73% (Public-Dyadic ID=29.14, *p*<0.001, *d*=2.023) and 41.73% (Public-Dyadic no ID=32.14, *p*<0.001, *d*=2.700) do so in the Dyadic ID and Dyadic no ID respectively. Moreover, there is substantially less undercutting by second movers in Public when the first mover reports 6 at 25.13% than in the Dyadic ID at 43.87% (Public-Dyadic ID=-18.74, *p*<0.001, *d*=-1.300) and 40.29% in the Dyadic no ID (Public-Dyadic no ID=-15.16, *p*<0.001, *d*=-1.33). Conditional on first movers reporting 1, 77.71% of second movers in the Public reciprocate with 1, while fewer do so in the dyadic treatments: 70.94% do so in Dyadic ID (Public-Dyadic ID=6.77, *p*=0.552, *d*=0.183) and 60.89% in the Dyadic no ID (Public-Dyadic no ID=16.82, *p*=0.067, *d*=0.588). The remaining second movers decide to report a higher number that leads to a mismatch and gains collaborators 0. We return to this seemingly odd outcome in the following section (p. 23). Regression analyses, which pool the FM and SM die-rolls but controls for extensive covariates, find substantively the same results (Table A8).

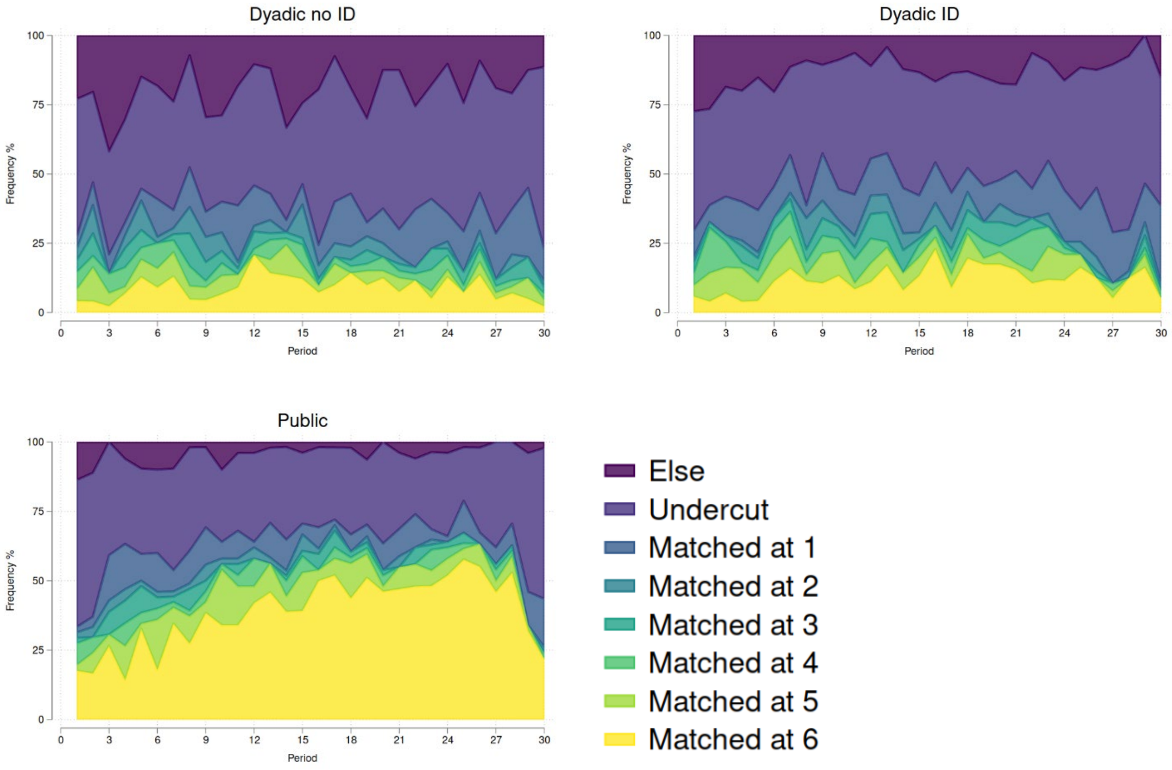
Figure 4. Distribution of second movers' choice (reported die-roll when entering the villain's dilemma) conditional on first movers' choice



4.1.3. Outcomes of the villain's dilemma

These differences in trusting and trustworthiness also clearly come across when considering outcomes (Figure 5). Across the 30 periods, collaborators in Public are able to match on 6 in 38.81% of the interactions. By contrast, matching at 6 is rare in the dyadic treatments: 11.66% for Dyadic ID and 9.0% for Dyadic no ID (differences for both comparisons relative Public: p<0.001 using two sample t tests run on group-level observations and d>2). Matching on 1 is fairly similar across the treatments at 8.51% in the Public treatment while it is 13.97% in Dyadic ID (difference relative to Public, p=0.09, d=-0.531) and 11.71% in Dyadic no ID (difference relative to Public, p=0.361, d=-0.289). Although even in Public, there is far from full trustworthiness and there is a real risk of back-stabbing, with almost a third (31.2%) of the outcomes end up with undercutting, this risk is highest in the dyadic treatments in which undercutting happens 38.8% in Dyadic ID (difference relative to Public, p=0.010, d=-0.825) and 37.39% in Dyadic no ID (difference relative to Public, p=0.034, d = -0.686). We find the same results when analysing undercutting using a probit regression (Table 5).

Figure 5. Outcomes in the villain's dilemma according to treatment



Note: Areas display the frequency of collaboration outcomes, distinguishing between matched and not-matched collaborations. Else includes all instances not characterized by matching or undercutting (i.e. second mover reports higher number than first mover or lower number by 2).

There are also interesting differences in dynamics across the rounds. Undercutting in round 1 is similar across the three treatments, i.e. 45.10% in Public, 35.29% in Dyadic ID and 39.58% in Dyadic no ID, and, while some divergence does appear, it does not clearly and stably diverge round-by-round (Table A11). By contrast, matching on 6 is already different in the first round. In Public, this is 17.65% while it is 5.88% in Dyadic ID (difference relative to Public, p=0.066, d=0.368) and 4.17% in the Dyadic no ID (difference relative to Public, p=0.033, d=0.435). This highlights the important role of information in facilitating profitable collaborative cheating thereby making entry into the dilemma attractive.

Therefore, what emerges from this analysis is that receiving more information on a potential collaborator (as in the Public treatment) seems to increase the trustworthy collaborations via both reducing undercutting the partner and increasing the joint-payoff maximizing choices.

Turning to the outcomes of “Else” (Figure 5), on first glance these seem puzzling. Why would any SM mismatch with a FM in such a way that they both get the worst possible outcome (0, 0)? One simple explanation is that some second movers make mistakes in their reports, misclicking or not understanding the scenario. Another is that some SM report their die-rolls honestly, which, in a majority of cases leads to the Else outcome. Yet there are also three more intriguing possibilities. First, in the Dyadic ID and Public treatments, in which it is possible to track individuals’ actions, a SM could take revenge and retaliate against a previous partner’s betrayal by inflicting costs on both of them. Second, a SM whose FM partner reports a low number may want to signal cooperativeness to their future partners by reporting a high number. Even though this imposes costs on them in the current round, reporting a high number may make their future partners likelier to trust them in collaborative dishonesty. Third, a SM may decide to impose costly punishment on the FM because that FM has reported a too low a number.

While we cannot cleanly separate between these possibilities, based on the design of our study and hints in the data, we believe that the most likely explanations are signalling and costly punishment. These are the only possibilities that can account for two patterns in the data and are not implausible based on the design (see Supplementary Materials, Section 1.3).

4.2. Does the villain’s dilemma promote or hamper dishonesty relative to individual settings?

4.2.1. Individual honesty and entering the villain’s dilemma

From the individual die-rolling reports (Stage 1), we can estimate using Fischbacher and Föllmi-Heusi's approach (2013, p. 533) that the percentage of entirely honest reports are 53.6% [=(8.94/(16.67))*100] across all treatments and 35.7% [=(5.95/16.67)*100] in Public, 57.3% [=(9.55/16.67)*100] in Dyadic ID, and 68.5% in Dyadic no ID [=(11.42/16.67)*100]. Similarly, in the sender-receiver task (Stage 2), we find that 63.0% of the senders are truthful (senders in the Stage 2 can only be fully honest or fully dishonest) (paired test on lying in Stage 2 vs. optin in Stage 3, $p < 0.001$). In contrast, across all treatments and rounds, only 13.5% of decisions in Stage 3 were to stay out—substantially lower than individual honesty (one-sample *t* test on difference between frequency of opting-out decision against 53.6%, $p < 0.001$). The same difference can be seen when considering treatments separately: 6.3% of decisions were to stay out in Public (one-sample *t* test on difference between frequency of opting-out decision against 35.7%, $p < 0.001$), 18.1% stayed out in Dyadic ID (one-sample *t* test on difference between frequency of opting-out decision against 57.3%, $p < 0.001$), and 16.0% in Dyadic no ID (one-sample *t* test on difference between frequency of opting-out decision against 68.5%, $p < 0.001$). These percentages are all far away from the individual honesty levels found. This suggests that the villain's dilemma encourages subjects' intentions of being dishonest. In the next subsection, we check whether these intentions turn into behaviour.

4.2.2. Individual die-rolling and villain's dilemma die-rolling

We further compare reported die-rolls as individuals relative to die-rolling once inside the villain's dilemma and find that the effect of the villain's dilemma depends upon (*i*) treatment, (*ii*) the role that a subject is in (FM or SM), and (*iii*) the round of the villain's dilemma (i.e. round 1 or average across all rounds) (Table 3).

Table 3. Frequency of reporting six by treatment, role, and round in the Stage 3

First Movers

	Stage 1	Stage 3, round 1	Stage 3, all	St1-St3r1	St1-St3
Public	0.335	0.255	0.521	0.08	-0.186***
Dyadic ID	0.194	0.137	0.245	0.057	-0.051
Dyadic no ID	0.223	0.125	0.242	0.098*	-0.019

Second Movers

	Stage 1	Stage 3, round 1	Stage 3, all	St1-St3r1	St1-St3
Public	0.429	0.235	0.405	0.194**	0.024
Dyadic ID	0.253	0.137	0.166	0.116***	0.087***
Dyadic no ID	0.217	0.063	0.128	0.154***	0.089***

Notes: frequencies of players' reporting 6, by treatment and by stage. First and second mover roles refer to the player's role in the first round of Stage 3. Between-stage comparisons (last two columns) are tested via paired *t* tests run on individual-level frequencies of reporting 6. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Start with the reporting of 6s by FMs in each treatment of Stage 3 and their Stage 1 reporting of 6s. In the Public treatment, 0.335 of Stage 1 reported die-rolls were 6s, in round 1 of Stage 3 this was similar at 0.255 (difference: p=0.190), but across all rounds reporting of 6s increased to 0.521 (difference: p<0.001). In Dyadic ID, Stage 1 reporting of 6s was 0.194, which is comparable to Stage 3 reporting in both round 1 at 0.137 (difference: p=0.290) and across all rounds at 0.245 (difference: p=0.178). And, in Dyadic no ID reporting of 6s in Stage 1 was 0.223 while it was lower in round 1 of Stage 3 at 0.125 (difference: p=0.063) and comparable across all rounds of Stage 3 at 0.242 (difference: p=0.329). Taken together, this means that, for first movers, dishonest behavior in the villain's dilemma is higher than individual die-rolling when there is public history and sufficient rounds have been played.[13]

Turn now to the reporting of 6s by the SMs. In Public, 0.429 of the Stage 1 reported rolls were 6s while in Stage 3 round 1 this was lower at 0.235 (difference: p=0.016) but had reached comparable levels across all rounds 0.405 (difference: p=0.910). In Dyadic ID, Stage 1 reporting of 6s was 0.253 while it was lower in Stage 3 in round 1 at 0.137 (difference: p=0.047) and across all rounds at 0.166 (difference: p<0.001). Likewise, in Dyadic no ID, Stage 1 reporting of 6s was 0.217, but this was reduced to 0.063 in round 1 of Stage 3 (difference: p=0.001) and remained lower at 0.128 across all rounds (difference: p<0.001). For second movers, these results mean that the villain's dilemma generally reduces dishonesty and only in one case—when there is public history and more rounds had been played—is there comparable levels of dishonesty.

4.3. Who participates in collaborative corruption and who is an untrustworthy corrupt collaborator?

4.3.1. Participating in collaborative corruption

Curiously, individual-level honesty—Stage 1 reported die-rolls (Models 1-5) and lying when being a sender in Stage 2 (Model 2, sample restricted to Sender participants)—are entirely

[13] Tests for between-treatment differences on individual vs. collaborative dishonesty are presented in Table A12 in the Supplementary Material. Using the same rationale as in Table 3, we rely on the frequency of reporting six in Stages 1 and 3 as a proxy for dishonesty and use the difference between these frequencies to indicate whether dishonest behaviour changes in the individual vs group setting.

unpredictive of entering collaboration in the villain’s dilemma (Table 4). People who are likelier to be dishonest in individual contexts are no likelier to choose collaboratively dishonest. Indeed, almost none of the individual-level factors are unhelpful in predicting entry into collaboration: neither trusting the message as a receiver in Stage 2 (Model 3, sample restricted to Receiver participants), score on the cognitive reflection test, experience with laboratory experiments, self-reported risk attitude, age, gender, extraversion, conscientiousness, nor neuroticism predict opting in. The only two exceptions are the dimensions of agreeableness and openness from the Big Five. Agreeableness is positively associated with opting in (Model 5, AME: +1.3%, std.err = 0.043, p = 0.023) while openness is negatively associated with it in (Model 5, AME: -1.6%, std.err = 0.040, p = 0.003). Above all, the strongest and consistent predictors are treatments.

Table 4. Opting to collaborate in the villain’s dilemma

	Model 1	Model 2	Model 3	Model 4	Model 5
Period	-0.001*	-0.001	-0.001	-0.001*	-0.001*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Ref. Cat: Dyadic ID					
Public	0.134***	0.093***	0.161***	0.127***	0.124***
	(0.019)	(0.028)	(0.031)	(0.019)	(0.020)
Dyadic no ID	0.024	0.016	0.033	0.024	0.021
	(0.027)	(0.035)	(0.039)	(0.027)	(0.026)
Experienced (Lab)	0.017	-0.010	0.040	0.008	0.005
	(0.033)	(0.043)	(0.057)	(0.037)	(0.035)
Mean Dice Stage 1		0.018	0.013	0.019	0.020
		(0.020)	(0.020)	(0.013)	(0.013)
Lied in Stage 2		0.014			
		(0.023)			
Trusted in Stage 2			-0.032		
			(0.031)		
Ref. Cat: risk seeking					
Risk neutral		0.010	-0.032	-0.013	-0.001
		(0.037)	(0.043)	(0.028)	(0.032)
Risk averse		0.007	-0.026	-0.010	-0.000
		(0.033)	(0.041)	(0.025)	(0.029)
Age				0.001	0.001
				(0.004)	(0.005)
Female				0.007	0.022
				(0.020)	(0.021)
Extraversion					0.003
					(0.005)

Agreeableness					0.013** (0.006)
Conscientiousness					-0.009 (0.007)
Neuroticism					-0.002 (0.005)
Openness					-0.016*** (0.005)
Cognitive Reflection Score					0.008 (0.009)
N	11340	5670	5670	11340	11340

Notes: Average marginal effects from random effects probit models with random intercepts at the individual level and standard errors clustered at the group level (reported in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Subjects were classified as experienced if they had participated in more than 5 prior experiments.

We also checked if there are interactions between individual dishonesty and treatment predict opting into collaboration (Table A13). Apart from a negative interaction between Public and mean Stage 1 die-rolling—indicating that more individually dishonest subjects opt to enter less than individually honest subjects in the Public treatment relative to the other treatments—there are no substantive heterogeneous treatment effects.

4.3.2. Untrustworthy corrupt collaborators

To understand what makes a collaboration fail or flourish, we now turn to undercutting behaviour (Table 5). Like for opting to collaborate, the strongest predictors of undercutting one's partner's die-roll in the villain's dilemma is the treatment, with Public triggering most frequently honest behaviour among corrupt collaborators (Models 1-5). Yet unlike for choosing to collaborate, individual honesty here matters. Reporting higher values in Stage 1 is positively associated with betraying the partner (Model 5, AME: +5.6%, std.err = 0.056, $p < 0.001$). That is, participants exhibiting a higher propensity to be dishonest individually with die-rolling, are also more likely to undercut their partner to obtain higher financial gains (Models 2-5). Additionally, experience with laboratory experiments seems to be negatively associated with the probability of undercutting one's partner (Models 3-5). Other individual characteristics, such as lying or trusting in the sender-receiver game in Stage 2 (respectively, Model 2 restricted to Senders and Model 3 restricted to Receivers), self-reported risk attitude and other personality traits do not predict the probability of undercutting one's partner.

Table 5. Undercutting instead of matching in the villain's dilemma

	Model 1	Model 2	Model 3	Model 4	Model 5
Period	0.001	0.001	0.002	0.001	0.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
Ref. Cat: Dyadic ID					
Public	-0.106***	-0.129***	-0.125***	-0.137***	-0.131***
	(0.031)	(0.045)	(0.039)	(0.034)	(0.033)
Dyadic no ID	0.038	0.035	0.039	0.035	0.041
	(0.031)	(0.045)	(0.042)	(0.032)	(0.035)
Experienced (Lab)	-0.053	-0.024	-0.160***	-0.093**	-0.085**
	(0.035)	(0.054)	(0.058)	(0.038)	(0.037)
Mean Dice Stage 1		0.069**	0.053**	0.057***	0.056***
		(0.030)	(0.025)	(0.019)	(0.019)
Lied in Stage 2		0.018			
		(0.037)			
Trusted in Stage 2			0.049		
			(0.037)		
Ref. Cat: risk seeking					
Risk neutral		0.007	0.034	0.028	0.028
		(0.060)	(0.043)	(0.038)	(0.040)
Risk averse		0.028	0.042	0.044	0.046
		(0.062)	(0.046)	(0.040)	(0.041)
Age				0.005	0.004
				(0.005)	(0.005)
Female				0.012	0.008
				(0.024)	(0.026)
Extraversion					-0.012*
					(0.006)
Agreeableness					-0.007
					(0.010)
Conscientiousness					-0.000
					(0.009)
Neuroticism					-0.011
					(0.007)
Openness					0.008
					(0.006)
Cognitive Reflection Score					-0.018
					(0.014)
N	3509	1803	1706	3509	3509

Notes: Average marginal effects from random effects probit models with random intercepts at the individual level and standard errors clustered at the group level (reported in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Subjects were classified as experienced if they had participated in more than 5 prior experiments.

We also checked for interactions between individual dishonesty and treatment that predict undercutting (Table A13). We find no significant interactions indicating that treatment effects on undercutting do not significantly vary by individual dishonesty.

In summary, we find that:

- Dyadic history (with or without id) both support some, and similar, levels of cooperation in the villain's dilemma. Yet, this level declines over round, and, is plagued by undercutting and low outcomes. Public history promotes the highest levels of collaborative corruption, in a substantively large way relative to the dyadic treatments, and, does so stably over time. This is because undercutting and poor collaborations, in the sense of matching on a low outcome, are substantially fewer than in the dyadic reputation treatments.
- Whether dishonesty is higher in the villain's dilemma than in the individual honesty settings depends upon the specific analysis. Choosing to collaborate—a plausible indicator of intention to be somewhat dishonest—are substantially higher than the proportion of somewhat dishonest in the individual tasks. Yet, comparing reported die-rolls shows that dishonesty is, with one exception, similar or lower in the individual die-roll task than in the villain's dilemma. Only for first-movers in the villain's dilemma, when there is public history, and across all rounds, is dishonesty higher.
- Individual factors are generally unrelated to opting in. The exceptions to this are agreeableness, which is positively associated with opting in, and openness, which is negatively associated with it. Moreover, the effect sizes are small (between 1-2%). Neither age nor gender are associated with opting to collaborate.
- When it comes to trustworthiness as a corrupt collaborator, we find that public history lowers the undercutting probability. Conversely, showing a higher lying tendency in the individual die-rolling task is positively associated with being an untrustworthy collaborator; this has a meaningful effect size with a 5.6% increase in betrayal for every one-unit increased in reported individual die-rolls.

5. Discussion and conclusions

Our villain's dilemma was designed to capture the tension between finding and collaborating amongst genuine cheaters and the possibility that genuine cheaters would betray any trust placed in them. All our treatments display results consistent with this tension but in the two treatments with dyadic history (Dyadic no ID and Dyadic ID) betrayal and distrust are particularly pronounced. A majority of subjects, declining over time, opt to collaborate. Yet, they frequently leave empty handed by being double-crossed, and rarely achieve the best outcome. This suggests that the levels of collaboration with only dyadic reputations would decrease even further over time, reducing collaboration to even lower levels. The situation is, instead, entirely different in the Public treatment. Entering collaborations is high, remains stable over time, and the outcomes from the realised interactions suggest that this stability should continue in the long term as a large proportion of the collaborations end in the maximum outcome for both parties. Indeed, the differences between treatments are substantial. Yet this does not imply a "criminal utopia" since there is also substantial back-stabbing. Rather a bifurcation happens: collaborators either work together to achieve the maximum outcome or one ends up cheating the other; alternative outcomes meanwhile (e.g. matching on 1) are rare.

Taken together, our results highlight the crucial role that reliable information plays in collaborative dishonesty. The importance of this information may be a key component preventing more collaborative crimes from happening. As Gambetta (2009a) highlights, "the conditions that make having a good reputation worthwhile and effective—easy diffusion of reliable information, easy reidentification of previous partners, stability, and long-lived firms—are not common in the underworld." (p. 40). If correct, this lack of information may be substantially constraining collaborative rule-breaking endeavours.

Yet, analysing die-rolling decisions in the collaborative setting paints a more complex picture. Dishonesty, in terms of reported die-rolls, in the villain's dilemma is similar or lower than in the individual die-roll task in almost every case. Only for first movers, in the Public treatment, is dishonest reporting larger. This is intriguing; it suggests that, on the one hand, intentions to be dishonest are promoted by the villain's dilemma, and on the other, it implies that whether these intentions are turned into actions depends upon the specific context and role. Indeed, further work should aim to tease apart the precise reasons for our findings. One promising approach is to study a more complete set of strategies that subjects adopt: from entering and deciding which die-roll to report conditional on prior experience to the strategies that second movers adopt based on their own interaction. Indeed, it is plausible that subjects enter the villain's dilemma, hoping to be a first mover, then report their die-rolls reasonably honestly with the hope that their second mover partner behaves dishonestly.

Finally, we find little support for the role of individual-level factors we collected. Both age and gender are unrelated to collaborating and to being a trustworthy in our experiment. More surprisingly, honesty in the solitary tasks does not predict entry into collaboration. However, and consistent with the villain's paradox, individual die-rolling dishonesty does substantively predict untrustworthiness as a collaborator. This is one of the key components of the villain's paradox, interacting with untrustworthy collaborators, and points to the difficulty in identifying collaborative cheaters before they undertake any dishonest behaviour. We do find some indication that agreeableness and openness are predictive of entering into collaborative dishonesty but further work needs to be undertaken to study the importance of these factors.

Declarations of interest: none

Funding: This work was partly supported by the Horizon 2020 Framework Programme Project PROTON "Modelling the Processes leading to Organised crime and Terrorist Networks" (No.: 699824) and the Knut and Wallenberg Grant "How do human norms form and change?" (2016.0167).

References

Abbink, K. (2004). Staff rotation as an anti-corruption policy: an experimental study. European Journal of Political Economy, 20(4), 887-906.

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

ACFE. (2020). *Report to the Nations: 2020 Global Study on Occupational Fraud and Abuse*. Association of Certified Fraud Examiners.

Barclay, P. (2016). Reputation. In D. Buss (Ed.), *The Handbook of Evolutionary Psychology* (2nd ed., Vol. 2). Wiley.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity and social history. *Games and Economic Behavior*, *10*, 122–142.

Childs, J. (2012). Gender differences in lying. *Economics Letters*, *114*(2), 147–149. https://doi.org/10.1016/j.econlet.2011.10.006

Cohn, A., Maréchal, M. A., Tannenbaum, D., & Zünd, C. L. (2019). Civic honesty around the globe. *Science*, *365*(6448), 70–73. https://doi.org/10.1126/science.aau8712

Conrads, J., Irlenbusch, B., Rilke, R. M., & Walkowitz, G. (2013). Lying and team incentives. *Journal of Economic Psychology*, *34*, 1–7. https://doi.org/10.1016/j.joep.2012.10.011

Dreber, A., & Johannesson, M. (2008). Gender differences in deception. *Economics Letters*, *99*(1), 197–199.

Erat, S., & Gneezy, U. (2012). White Lies. *Management Science*, *58*(4), 723–733. https://doi.org/10.1287/mnsc.1110.1449

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

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

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*(4), 25–42.

Fries, T., Gneezy, U., Kajackaite, A., & Parra, D. (2021). Observability and lying. *Journal of Economic Behavior & Organization*, *189*, 132–149. https://doi.org/10.1016/j.jebo.2021.06.038

Gambetta, D. (2009a). *Codes of the underworld: How criminals communicate*. Princeton University Press.

Gambetta, D. (2009b). Signaling. In P. Hedström & P. Bearman (Eds.), *The Oxford Handbook of Analytical Sociology* (pp. 169–194). Oxford University Press.

Gambetta, D., & Przepiorka, W. (2019). Sharing Compromising Information as a Cooperative Strategy. *Sociological Science*, *6*, 352–379. https://doi.org/10.15195/v6.a14

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

Giardini, F., Wittek, R., Giardini, F., & Wittek, R. (Eds.). (2019). *The Oxford Handbook of Gossip and Reputation*. Oxford University Press.

Gibson, R., Tanner, C., & Wagner, A. F. (2013). Preferences for Truthfulness: Heterogeneity among and within Individuals. *American Economic Review*, *103*(1), 532–548. https://doi.org/10.1257/aer.103.1.532

Gneezy, U. (2005). Deception: The Role of Consequences. *American Economic Review*, *95*(1), 384–394. https://doi.org/10.1257/0002828053828662

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

Gross, J., Leib, M., Offerman, T., & Shalvi, S. (2018). Ethical free riding: When honest people find dishonest partners. *Psychological Science*, 0956797618796480. https://doi.org/10.1177/0956797618796480

Gylfason, H. F., Arnardottir, A. A., & Kristinsson, K. (2013). More on gender differences in lying. *Economics Letters*, *119*(1), 94–96.

Hanna, R., & Wang, S.-Y. (2017). Dishonesty and Selection into Public Service: Evidence from India. *American Economic Journal: Economic Policy*, *9*(3), 262–290. https://doi.org/10.1257/pol.20150029

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

Irlenbusch, B., Mussweiler, T., Saxler, D. J., Shalvi, S., & Weiss, A. (2020). Similarity increases collaborative cheating. *Journal of Economic Behavior & Organization*, *178*, 148–173. https://doi.org/10.1016/j.jebo.2020.06.022

KPMG. (2016). *Global profiles of the fraudster*.

Leib, M., Köbis, N., Soraperra, I., Weisel, O., & Shalvi, S. (2021). Collaborative dishonesty: A meta-analytic review. *Psychological Bulletin*, *147*(12), 1241–1268. https://doi.org/10.1037/bul0000349

Milinski, M. (2016). Reputation, a universal currency for human social interactions. *Phil. Trans. R. Soc. B*, *371*(1687), 20150100. https://doi.org/10.1098/rstb.2015.0100

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

Muehlheusser, G., Roider, A., & Wallmeier, N. (2015). Gender differences in honesty: Groups versus individuals. *Economics Letters*, *128*, 25–29. https://doi.org/10.1016/j.econlet.2014.12.019

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), Article 7063. https://doi.org/10.1038/nature04131

Patt, A., & Zeckhauser, R. (2000). Action Bias and Environmental Decisions. *Journal of Risk and Uncertainty*, *21*(1), 45–72. https://doi.org/10.1023/A:1026517309871

Przepiorka, W., Norbutas, L., & Corten, R. (2017). Order without law: Reputation promotes cooperation in a cryptomarket for illegal drugs. *European Sociological Review*, *33*(6), 752–764. https://doi.org/10.1093/esr/jcx072

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

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

Rilke, R. M., Danilov, A., Weisel, O., Shalvi, S., & Irlenbusch, B. (2021). When leading by example leads to less corrupt collaboration. *Journal of Economic Behavior & Organization*, *188*, 288–306. https://doi.org/10.1016/j.jebo.2021.05.007

Roberts, G., Raihani, N., Bshary, R., Manrique, H. M., Farina, A., Samu, F., & Barclay, P. (2021). The benefits of being seen to help others: Indirect reciprocity and reputation-based partner choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), 20200290. https://doi.org/10.1098/rstb.2020.0290

Romano, A., Giardini, F., Columbus, S., de Kwaadsteniet, E. W., Kisfalusi, D., Triki, Z., Snijders, C., & Hagel, K. (2021). Reputation and socio-ecology in humans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), 20200295. https://doi.org/10.1098/rstb.2020.0295

Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, *45*, 181–196. https://doi.org/10.1016/j.joep.2014.10.002

Ścigała, K. A., Schild, C., Heck, D. W., & Zettler, I. (2019). Who Deals With the Devil? Interdependence, Personality, and Corrupted Collaboration. *Social Psychological and Personality Science*, *10*(8), 1019–1027. https://doi.org/10.1177/1948550618813419

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

Spence, M. A. (1973). Job market signaling. *Quarterly Journal of Economics*, *87*(3), 355–374.

Spence, M. A. (1974). *Market signaling: Informational transfer in hiring and related screening processes*. Harvard University Press.

Sutter, M. (2009). Deception Through Telling the Truth?! Experimental Evidence From Individuals and Teams*. *The Economic Journal*, *119*(534), 47–60. https://doi.org/10.1111/j.1468-0297.2008.02205.x

Számadó, S., Balliet, D., Giardini, F., Power, E. A., & Takács, K. (2021). The language of cooperation: Reputation and honest signalling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), 20200286. https://doi.org/10.1098/rstb.2020.0286

Thielmann, I., Hilbig, B. E., Klein, S. A., Seidl, A., & Heck, D. W. (2024). Cheating to benefit others? On the relation between Honesty-Humility and prosocial lies. *Journal of Personality*, *92*(3), 870–882. https://doi.org/10.1111/jopy.12835

Weisel, O., & Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences*, *112*(34), 10651–10656. https://doi.org/10.1073/pnas.1423035112

Weisel, O., & Shalvi, S. (2021). Moral currencies: Explaining corrupt collaboration. *Current Opinion in Psychology*. https://doi.org/10.1016/j.copsyc.2021.08.034

Wu, J., Balliet, D., & Van Lange, P. A. M. (2016). Reputation, gossip, and human cooperation. *Social and Personality Psychology Compass*, *10*, 350–364. https://doi.org/10.1111/spc3.12255

Zettler, I., Thielmann, I., Hilbig, B. E., & Moshagen, M. (2020). The Nomological Net of the HEXACO Model of Personality: A Large-Scale Meta-Analytic Investigation.

Perspectives on Psychological Science, *15*(3), 723–760. https://doi.org/10.1177/1745691619895036