Corrupt third parties undermine trust and prosocial behaviour between people

This preprint has not undergone peer review or any post-submission improvements or corrections. The Version of Record of this article in published in *Nature Human Behaviour* and is available online at https://doi.org/10.1038/s41562-022-01457-w.

Corrupt Third-Parties Undermine Trust and Cooperation Between People

Giuliana Spadaro[a,b], Catherine Molho[c,d], Jan-Willem Van Prooijen[a,b,e], Angelo Romano[f], Cristina O. Mosso[g], Paul A. M. Van Lange[a,b]

[a] Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, 1081BT Amsterdam, The Netherlands

[b] Institute for Brain and Behavior Amsterdam (IBBA), Vrije Universiteit Amsterdam, 1081BT Amsterdam, The Netherlands

[c] Institute for Advanced Study in Toulouse, Esplanade de l'Université 1, Toulouse, 31080, Cedex 06, France

[d] Center for Research in Experimental Economics and Political Decision Making (CREED), University of Amsterdam, 1001 NJ Amsterdam, The Netherlands

[e] The Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), 1081HV Amsterdam, The Netherlands

[f] Department of Social, Economic and Organisational Psychology, Leiden University, Wassenaarseweg 52, 2333AK Leiden, The Netherlands

[g] Department of Psychology, University of Turin, 10124 Turin, Italy

Running Head: Corrupt Third-Parties Undermine Trust and Cooperation

Corresponding Author: Giuliana Spadaro, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands, g.spadaro@vu.nl

Abstract

Corruption is a pervasive phenomenon that affects the quality of institutions, undermines economic growth, and exacerbates inequalities around the globe. It is often assumed that perceiving public institutions as corrupt is a cause of low interpersonal trust among citizens. However, to the best of our knowledge, this causal claim has never been directly tested. Here we ask whether perceiving representatives of institutions as corrupt undermines interpersonal trust and subsequent cooperation among strangers. We used a novel experimental game paradigm that modeled institutional representatives as third-party punishers (TPP). Five studies manipulated or assessed corruption of a TPP and then examined its effects on trust and cooperation. In a die-rolling task, participants were exposed to dishonest behavior of a target who would subsequently serve as a TPP in a trust game (Study 1a, $N = 540$), in a prisoner's dilemma game (Study 1b, $N = 503$), and in dictator games (Studies 2-4, $N = 785$, pre-registered). Overall, the studies consistently revealed that perceiving a third-party as corrupt decreases interpersonal trust and, in turn, cooperation towards an unrelated stranger. These findings contribute to our understanding of the critical role that representatives of institutions play in shaping cooperative relationships in modern societies – undermining trust when they prove to be corrupt.

Keywords: Corruption; Trust; Cooperation; Institutions; Dishonesty

Corrupt Third-Parties Undermine Trust and Cooperation Between People

In 2015, an anonymous source leaked 11.5 million documents from the fourth biggest offshore law firm in the world. This leak unveiled a system of rich people, politicians, public officials, or close associates who exploited their privileged position to engage in tax evasion, fraud, and evasion of international sanctions. Overall, a total of 600 people from 42 countries were involved in what is considered one of the biggest leaks ever reported[1]. This scandal, now known as the *Panama papers*, pointed at a long-standing problem of institutional representatives taking advantage of their entrusted position in order to gain private benefits[2,3]. Whilst the institutional challenges that such scandals pose have been extensively examined[4], less attention has been devoted to the question of whether being exposed to dishonesty and corruption by third-parties affects interpersonal trust and subsequent cooperative behavior toward fellow citizens.

Corruption is a critical societal and scientific issue that has attracted considerable research interest across many disciplines, such as economics, political science, sociology, law, and psychology [5–8]. Decades of research have aimed to understand this phenomenon and its dramatic consequences on societies. Many studies focused on cross-cultural differences in corruption levels, while other research investigated what makes people engage in corrupt behavior [7,9,10]. Importantly, it has been hypothesized that corruption may affect social interactions involving interpersonal trust and cooperation[11–13]. In daily life, people learn about corruption by directly or indirectly witnessing the behavior of their representatives, such as public officials that accept bribes, or politicians that evade taxes, which in turn affects their trust toward institutions[14]. Yet, to date, experimental evidence on whether perceiving corruption by representatives of institutions causally undermines trust and cooperation toward strangers is lacking.

In the interdisciplinary literature on corruption and trust, two major streams of research have advanced. One may be labeled the *bottom-up* perspective, which assumes that the effectiveness of institutions depends largely on informal social processes, such as individuals' ability to solve local and small-scale social dilemmas[15–17]. In this view, interpersonal trust is considered the basis for ensuring the effectiveness of institutions[18]. A second perspective, which may be referred to as *top-down,* assumes that institutions shape human interactions, and therefore influence interpersonal trust and cooperation. Here interpersonal trust is considered a result of the quality of institutions. Often, this perspective goes even further by suggesting that one of the main functions of institutions is to mitigate vulnerability in interactions with strangers[19,20]. If public institutions and their representatives are perceived as unable to provide security, then interpersonal trust can emerge only in narrow and tight networks. Yet, in modern globalized contexts, transactions with strangers are frequent and necessitate building *generalized* trust[21].

There is some empirical evidence that provides support for this top-down hypothesis, showing that interpersonal trust increases among individuals who migrate to countries with lower levels of corruption[22], and that institutional trust is one of the strongest predictors of interpersonal trust[23]. Notably, research suggests that experiencing corruption enacted by public officials or by other strangers is associated with individual behaviors, such as honesty or ingroup solidarity. In fact, individuals display less ethical values when they are exposed to institutions with more prevalent corrupt practices[24,25]. Moreover, the mere observation of corrupt behavior enacted by neighbors or ingroup members seems associated with individuals' propensity to act dishonestly[9,26]. Yet, the relation between exposure to corruption and generalized trust is still an unsolved issue. Importantly, if corrupt representatives of institutions have a negative effect on trust, this may also have crucial cascading implications for cooperation between strangers. Indeed, trust is one of the most influential factors that

determine cooperation in situations when a conflict between individual and collective interest occurs[27]. As the implementation of third-party sanctioning institutions is one of the most powerful strategies to promote cooperation in the absence of reputational information[28], it becomes vital to understand whether corruption of such third-parties may undermine the effectiveness of sanctioning[13].

Individuals witness norm violations from peers on a daily basis[29], but they are also exposed to violations from representatives of public institutions. Here we examine whether learning that institutional representatives are corrupt (i.e., act dishonestly to enhance their self-benefit and use their power to profit at the expense of the collective) undermines trust and cooperation toward strangers. In a set of five studies, we distinguish between two different sources of perceived corruption that may underlie beliefs about corrupt institutions in everyday life and negatively affect trust toward institutional representatives: second-hand learning (e.g., political scandals broadcasted in media) and first-hand experience (e.g., personal experience with corrupt authorities accepting bribes)[30]. Second-hand learning of corruption is very frequent in daily life and has been associated to sudden declines in trust toward political representatives[31,32]. First-hand experience of corruption may be less ubiquitous in some contexts or cultures[24], but elicits long lasting negative societal outcomes[13,33].

The Experimental Paradigm

We developed a novel experimental paradigm that is rooted in the tradition of research using economic games[34,35]. In this paradigm, individuals can make decisions to trust and cooperate with others under the scrutiny of a third-party observer that proved to be corrupt (or not) in a previous interaction. The game is divided into two phases (see Figure 1).

In Phase 1, participants observe a person cheating (or not) in a sequential die-rolling task, that is, a situation that allows to profit by acting dishonestly[7]. In this task, two players

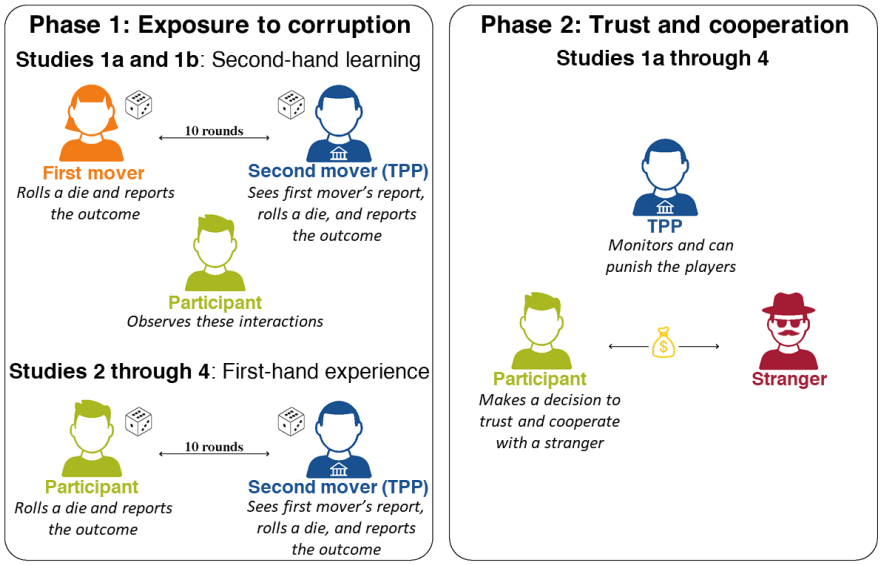
are instructed to roll a six-sided die privately and sequentially, and to report the given outcome. The first mover earns a monetary payoff regardless of the outcomes of the die-roll, while the second mover receives a payoff only when their outcome exactly matches the declared outcome of the first mover. Importantly, in this task, the second mover knows in advance the reported outcome of the first mover, while it is impossible for the experimenter to verify whether the second-mover's declared outcome corresponds with the actual die-rolling. Hence, for the second mover, this situation captures specific dishonest behaviors that are closely linked to corruption, as the second mover is tempted to misuse the information and declare corresponding outcomes for self-benefit. This paradigm has been demonstrated to elicit dishonest behavior from participants, whose reports deviate substantially from reports expected by chance[7,36,37]. Such "too good to be true" outcomes are unambiguously interpreted by others as dishonest behavior[38,39].

To resemble real-life situations where people can learn about corruption indirectly or directly, participants either observe an interaction between the two movers and learn that the second mover behaved honestly or dishonestly (i.e., second-hand learning; Studies 1a and 1b), or they personally engage in the die-rolling task as first movers and experience the second mover's honesty or dishonesty themselves (i.e., first-hand experience; Study 2 through 4). While Studies 1a through 2 specifically focus on dishonest behavior that enhances self-benefit (a specific feature of corruption), in Studies 3 and 4 we model corrupt behavior to also include its negative externalities for the collective. Specifically, the second mover's dishonesty in Phase 1 enhances self-benefit *and* directly harms the collective by taking possession of resources that would otherwise benefit all participants (Study 3) or the broader collective (Study 4).

Then, participants transition to Phase 2 of the game. Here, they learn that the person they just observed or interacted with as second mover will serve as a third-party punisher

(TPP) in an economic game (i.e., trust game, prisoner's dilemma game, dictator game) where they can decide whether or not to cooperate with a stranger. Participants are informed that the TPP (previously the second mover they learnt to be either honest or corrupt) can invest his or her own endowment to reduce players' outcomes in the game based on their behavior. The implementation of TPP has been used extensively in previous research using economic games to model behavior from institutional representatives[40]. In these settings, it is common to observe punishing behavior from third parties, even if it is costly and seemingly at odds with self-interest[41]. We measure interpersonal trust using a six-item scale, which asks participants how much they trust a new partner (i.e., a participant who was not part of the die-rolling in Phase 1) with whom they are matched in the one-shot TPP economic games played in Phase 2[42]. We measure cooperation as the amount given to this new partner in Phase 2.

Figure 1. The two phases of the experimental procedure of Studies 1a through 4.



Note. TPP = Third-party punisher

Overview of the Studies

Our main question here is whether knowing that a third-party punisher has behaved dishonestly or honesty in the past affects interpersonal trust and cooperative behavior toward

an unrelated stranger. Across five studies, we tested the following two hypotheses. The first hypothesis is that observing corrupt behavior by a person serving as a third-party who administers sanctions will undermine trust toward a stranger in a subsequent situation. The second hypothesis is that the influence of corruption on interpersonal trust should, in turn, undermine subsequent cooperation with the same unknown partner in an economic game.

Studies 1a and 1b provided a preliminary, internally valid, test for our hypotheses and focused on second-hand learning of dishonest behavior displayed by a third-party with punishment capacity. Participants observed an ostensible die-rolling task interaction between two movers in Phase 1. There, we manipulated corruption through varying the degree of cheating of the second mover by providing pre-programmed feedback about both players' behavior in the die-rolling task (1 out of 10 reported doubles vs 10 out of 10 reported doubles). The second mover subsequently served as a third-party punisher (TPP) in Phase 2. Additionally, we included a control condition in which participants simply observed a player reporting 10 out of 10 doubles but not acting as TPP in Phase 2. In this condition, the TPP was another stranger about whom participants had no reputational information. Then, we assessed interpersonal trust and cooperation with an unknown partner in a trust game involving third-party punishment[43]. In Study 1b, we replicated the design of Study 1a but we examined interpersonal trust and cooperation in a TPP prisoner's dilemma game[44].

Study 2 provided a pre-registered replication of our findings and tested the hypotheses in a setting in which participants could directly interact with a potentially corrupt (or honest) future third-party punisher. Contrary to Study 1a and 1b, participants could observe naturally emerging levels of corruption on behalf of the third-party – a feature that may better align with experiences from everyday life. Specifically, in Phase 1, participants were matched in dyads to take part in the sequential die-rolling paradigm[7]. Then, they transitioned to Phase 2 of the game, in which we measured first movers' interpersonal trust towards a stranger and

cooperative behavior in a TPP dictator game. The participants previously acting as second movers acted as the third-party punisher in Phase 2.

Studies 3 and 4 expanded on the previous studies by providing a pre-registered test of our hypotheses in a more ecologically valid setting, which focused on first-hand experience of a third-party misusing their power to profit at the expense of the collective. To do so, we adapted the procedure employed in Study 2 by introducing a different incentive structure that more closely resembles the definition of corruption as "*abuse of public means for private gain*" and "*power asymmetry over shared resources*"[2,3]. Accordingly, dishonest behavior from the second movers in the die-rolling task (and subsequent TPP in the cooperation tasks) directly resulted in the depletion of a common good relevant to the community of participants involved in the study (i.e., a fund to be equally allocated among all participants in the study; Study 3) or to the broader human collective (i.e., a fund to be allocated to pro-environmental charity; Study 4).

Results

Study 1a

To test whether the manipulation of corruption was successful, we asked participants to what extent they perceived the second mover as honest in reporting his/her score on a 7-point Likert scale (1 = completely dishonest, 7 = completely honest). We reversed-scored this item for easier interpretation, with high scores indicating greater perceived dishonesty. The manipulation resulted in greater perceptions of dishonesty when second movers reported 10/10 doubles (M = 5.60, SD = 1.92), compared to 1/10 doubles (M = 1.55, SD = 1.08), ($t(538) = 26.43$, $p < .001$, $d = 2.41$). A one-way ANOVA testing the effects of corruption of the TPP on interpersonal trust toward the stranger in the trust game (TG) revealed a main effect of the manipulation of corruption, $F(2, 537) = 4.67$, $p = .01$, $\eta^2_p = .017$ (see Table S1.1). We created two hypotheses-relevant orthogonal contrasts of our experimental conditions:

Contrast 1 (*corrupt TPP* vs. *honest TPP and control* conditions) and Contrast 2 (*honest TPP* vs. *control* conditions). Planned comparisons revealed a significant Contrast 1, $F(1, 539) = 5.63$, $p = .018$, $d = 0.22$, indicating less trust towards a stranger when being monitored by a corrupt TPP ($M = 4.50$, $SD = 1.36$), compared to an honest TPP and an unknown TPP ($M = 4.80$, $SD = 1.41$). Contrast 2 was not significant, $F(1, 358) = 3.62$, $p = .058$, indicating that having observed an honest behavior in Phase 1 did not affect interpersonal trust. Then, we tested whether being exposed to corrupt third-parties (corrupt TPP condition) affected interpersonal trust and, in turn, cooperation using the bootstrapping method for mediation analysis[45]. The results show evidence of a significant indirect effect of corruption on cooperation in the TG via interpersonal trust, $b = 0.21$, 95% CI [0.07, 0.37]. Hence, our first study provides initial evidence that perceiving institutional representatives as corrupt undermines trust toward strangers, and in turn cooperation.

Study 1b

The manipulation was again successful in affecting perceived dishonesty of the TPP, $t(501) = 21.92$, $p < .001$, $d = 1.96$, with TPP being perceived as more dishonest in the corrupt TPP condition ($M = 5.32$, $SD = 1.98$) compared to the honest TPP condition ($M = 1.93$, $SD = 1.46$). Results of Study 1b revealed that participants who faced a corrupt TPP trusted their partner less ($M = 4.91$, $SD = 1.42$) than participants who faced an honest TPP ($M = 5.22$, $SD = 1.25$), $F(1, 502) = 7.04$, $p = .008$, $d = 0.23$. Then, we tested whether perceiving the third-party as corrupt affected cooperation in the prisoner's dilemma (PD) indirectly through interpersonal trust, using the bootstrapping method for mediation analysis[45]. The results show a significant indirect effect of corruption on cooperation via interpersonal trust, $b = 1.82$, 95% CI [0.51, 3.50]. Altogether, these results replicate findings of Study 1a, showing a negative effect of corruption on trust and subsequent cooperative behavior.

Study 2

We conducted a simple linear regression in which interpersonal trust toward the stranger in the dictator game (DG) was regressed on the sender's perceptions of dishonesty of the TPP in reporting the outcomes in the die-rolling task. Consistent with our hypothesis, perceived dishonesty of the TPP significantly and negatively predicted the extent to which the senders trusted the receivers, $\beta = -.39$, $t(189) = 5.88$, $p < .001$, and explained a significant proportion of variance $R^2 = .16$, $F(1, 189) = 34.55$, $p < .001$. Then, we tested whether interacting with a TPP (perceived as honest or dishonest) in a previous die-rolling task would indirectly affect cooperation via interpersonal trust. Using the bootstrapping method for mediation[45], we replicated the findings of the previous studies, showing a significant indirect effect on cooperation, $b = -0.63$, 95% CI [-1.25, -0.05]. Overall, Study 2 presents compelling evidence within a real-interaction setting that the more the participants perceived the second movers (and subsequent TPP) as corrupt in the die-rolling task, the less they trusted an unrelated player in the subsequent DG with third-party sanctioning. Moreover, we found again that this decline in trust had a cascading negative effect on cooperative behavior.

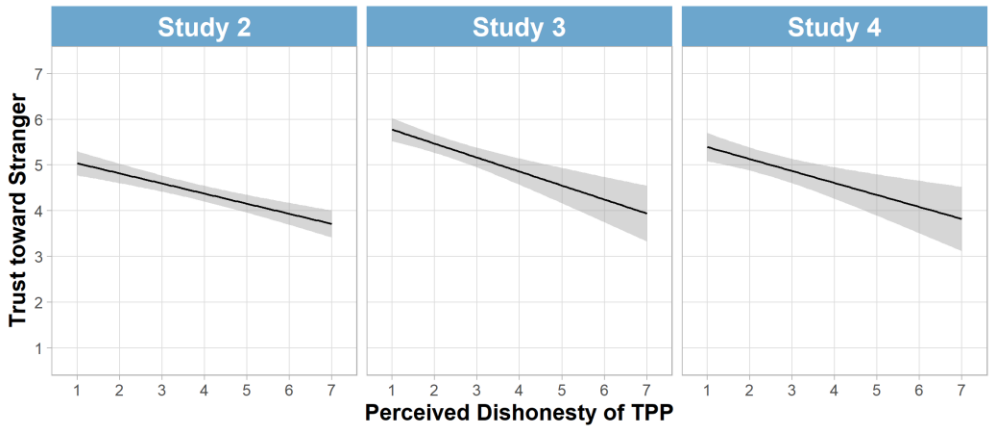
Studies 3-4

Studies 3 and 4 used an incentive structure in which corruption is operationalized in terms of self-benefit and detrimental consequences for the collective. Again, the findings of both studies supported both preregistered hypotheses, thereby fully replicating the patterns observed in Study 2. In both studies, perceived dishonesty of the third-party significantly and negatively predicted the extent to which the senders trusted the receivers in the DG ($R^2_{\text{Study 3}} = .19$, $R^2_{\text{Study 4}} = .12$) (Figure 2), and showed a significant indirect effect on cooperation via trust ($b = -2.42$, 95% CI $_{\text{Study 3}}$ [-5.00, -0.54], $b = -2.04$, 95% CI $_{\text{Study 4}}$ [-3.83, -0.61]). Notably, findings of Study 5 remained significant while controlling for the sender's subjective importance of the mission of the charity. Results are presented in detail in the SI, along with robustness checks. An internal random effects meta-analysis of the correlation between the

perception of dishonesty of the TPP in the die-rolling task and interpersonal trust toward the stranger in Studies 2-4 displayed a medium size negative meta-analytic correlation (k = 3, r = -.40, 95% CI [-.48, -.31]).

Expanding upon previous studies, Studies 3 and 4 included a measure of the senders' expectations about punishment enacted by the TPP in the DG to explore a potential underlying mechanism. Specifically, we aimed to explore whether the changes in cooperation can be attributable to the belief that the TPP would (not) punish senders. Given that punishment decisions are costly, dishonest third-parties could be expected to be more selfish and punish less, and this might explain why senders cooperate less with others in the presence of a dishonest TPP. Results of the two studies provided inconsistent findings, revealing an indirect effect via punishment expectations on cooperation in Study 3 (95% CI [-6.02, -2.22]), and a lack thereof in Study 4 (95% CI [-2.38, -1.78]) (see SI). Thus, our exploratory findings do not provide enough evidence to support (or rule out) this potential mechanism.

Figure 2. Effect of perceived dishonesty of the TPP in the die-rolling task on trust toward a stranger in the dictator games across Studies 2 through 4.



Discussion

Considerable research in various scientific disciplines has addressed the intricate association between the degree to which institutions are corrupt and the extent to which people trust one another and build cooperative relations. One perspective suggests that the success of institutions is rooted in interpersonal processes such as trust[16]. Yet another perspective assumes a top-down process, suggesting that the functioning of institutions serves as a basis to promote and sustain interpersonal trust[20,23]. However, as far as we know, no study has tested this latter causal claim in an experimental setting.

In the present research we provided an initial test of a top-down perspective, examining the effects of a corrupt versus honest institutional representative, here operationalized as a third-party observer with the power to regulate interaction through punishment. To do so, we revisited the sequential die-rolling paradigm where participants could learn whether the third-party was corrupt or not via second-hand learning or via first-hand experience. Across five studies (N = 1,828), we found support for the central hypothesis guiding this research, that perceiving third parties as corrupt causes a decline in interpersonal trust, and subsequent cooperation, toward strangers. This result was robust across a broad set of economic games and designs. Our findings contribute to the trust literature by suggesting

that institutional representatives exert substantial influence on interpersonal trust within societies. Hence, this can be interpreted as first evidence for a top-down causal route in the relation between institutions, trust, and cooperation. Importantly, this evidence does not rule out the influence of bottom-up processes, which could coexist and even have a reciprocal influence[46]. However, it is almost inevitable that eventually at least some dishonest behavior by institutions will reach the public eye, which in turn endangers interpersonal trust. The result of such a challenge to trust can be dramatic, because the repair of trust may not always last in the long term, and it is a process that requires considerable effort *and* time[47,48]. Because trust is essential to well-functioning groups, organizations, and societies[46], a lack of trust can at least temporarily undermine societal development, as it is related to important societal outcomes such as economic growth and political participation[49,50].

We also found that lower levels of interpersonal trust caused by the exposure to corrupt institutional representatives have cascading negative effects on cooperative behavior. Promoting cooperation is essential to solve important problems such as global warming, pollution, tax evasion, and other societal collective challenges[27,51]. If such a spillover effect of corruption on cooperation occurs (as suggested by our findings), future interventions should be implemented following a top-down approach that starts from institutional representatives, rather than horizontally between individuals[47]. If citizens tend to distrust each other as a result of exposure to corrupt institutional representatives, the implementation of punishment or reputational systems may not be effective, or may even backfire, crowding-out interpersonal and institutional trust, or giving rise to antisocial punishment[15,52].

Before closing, we briefly discuss some limitations and avenues for future research. First, in the present research institutional representatives were operationalized at the micro-level as a third-party sanctioning actor in cooperative exchanges. Although this operationalization is commonly used in the experimental literature[53,54], it does not fully

capture the complex and encompassing world of institutions most people experience in everyday life. Second, the online setting also differs from daily experiences with institutional representatives. One key difference is that these experiences are often repeated (rather than one-shot) and usually extend over substantial periods of time[13]. Therefore, in everyday life people may often come to internalize norms of corruption, and the detrimental effects of corruption may be even more dramatic and enduring. This seems especially true for some countries in which individuals regularly observe and need to interact with corrupt representatives with sanctioning power[11].

It is important to acknowledge that our findings do not provide consistent evidence to support (or rule out) specific mechanisms underlying the relationship between corruption and cooperation. For example, although it is possible that participants cooperate less because corrupt third-parties are less likely to enact costly punishment, this hypothesis was only supported in one of the two studies that included a measure of expectations of punishment. Testing underlying mechanisms constitutes an important topic for future investigation, a recommendation that we also make for enhancing ecological validity. For example, future research could complement this set of studies by investigating the effect of common –and subtle– cues of dishonest behavior that characterize real-world trust (e.g., facial expressions[55]). Moreover, in our design participants did not benefit from the corrupt behavior of the third-party, while in many real-life situations, individuals often directly benefit from corrupt transactions[7]. Thus, future research should address this point by examining the boundary conditions of the relationship between corruption, trust, and cooperation in situations where participants benefit from the corrupt transaction.

To conclude, five studies uncovered that perceiving institutional representatives as corrupt has detrimental effects on interpersonal trust and cooperation. These findings illuminate the vital functions that institutions have in shaping human psychology, as well as

the negative effects that they may have on our perception and behavior with strangers. Hence, corruption among institutional representatives may shape and facilitate a culture in which not only corrupt activities may come to be viewed as relatively common and normative[24], but also give rise to distrust among strangers. The fact that corruption and distrust are partially rooted in institutional representatives is also relevant for policy that focuses on reducing corruption in a sustainable manner. One broader implication is that groups and societies should do all they can to attract institutional leaders with integrity, and perhaps equally important, shape and nurture an environment in which such leaders are least likely to push or cross ethical boundaries.

Method

The research was approved by the Scientific and Ethical Review Board (VCWE) of the Faculty of Behavioural & Movement Sciences, VU Amsterdam Application VCWE-2017-085. Materials and pre-registrations are accessible on the OSF at https://tinyurl.com/yablrfam (materials Studies 1a-4), https://tinyurl.com/ya5bgbq5 (pre-registration Study 2) https://tinyurl.com/yb9or2sf (pre-registration Studies 3 and 4). Participants provided their informed consent in all studies.

Study 1a

Participants and Procedure

An a-priori power analysis (G*Power[56]) revealed a required sample size of 528 to achieve statistical power of .80 to detect an effect size of $d = 0.30$ of the corruption manipulation on interpersonal trust (two-sided). Participants ($N = 540$; 47% women: $M_{age} = 35.44$, $SD = 11.10$) were recruited from Amazon Mechanical Turk (MTurk) and completed the online study for $1. Moreover, they could earn up to 1.50$ based on their decisions, and five participants won a $10 prize. Samples from MTurk are heterogeneous in terms of socio-economic and ethnic diversity, and MTurk is a reliable platform to perform behavioral

tasks[57,58]. We used a between-participants design with three conditions: honest TPP, corrupt TPP, and control. All tests reported below are two-sided.

Die-Rolling Task (Phase 1). The manipulation of corruption occurred in Phase 1 of the game. In the *honest TPP* condition, participants observed a targeted prospective third-party punisher (TPP) behaving honestly across 10 rounds of the die-rolling task. Specifically, they learned that the second mover (the prospective TPP) mimicked the outcome of the first mover only in 1 out of 10 rounds (rounding down the expected number of doubles assuming honest reporting: 1.66). In the *corrupt TPP* condition, participants observed the future TPP reporting the same outcome in the die-rolling task on 10 out of 10 rounds. Importantly, we included a *control* condition where participants observed a dishonest player, but during the following game they faced a TPP they never encountered before. This condition allowed us to rule out the alternative explanation that it is the mere exposure to corrupt behavior that influences interpersonal trust, rather than the perception of an institutional representative (i.e., the TPP), in particular, as corrupt. Across conditions, participants were not aware that the second mover in Phase 1 would take part in Phase 2. To elicit and incentivize the attention of the participants when they observed behavior in the die-rolling task, they were informed that they would be eligible for a lottery prize of $10 (5 prizes in total) in case of all correct answers in the attention check questions. To this purpose, the observers received questions about the rules of the game, the role of each player, and the outcomes of the 10 rounds. Prior to receiving instructions, participants were asked to roll a computerized die on an online website in order to increase the belief that the game and the partners were actually interacting. In reality, their reports were pre-programmed feedback provided according to our experimental manipulations.

Interpersonal Trust and Cooperation (Phase 2). In Phase 2, participants were matched with a stranger and played a trust game (TG[43]) with TPP. In this game, participants

were endowed with 5 monetary units (MU, each worth $0.10) that they could decide to give to the unrelated stranger (the trustee). They knew that each MU they sent to the stranger would be tripled, and then the stranger could decide to return (or not) any amount. Importantly, they knew that their decisions would be observed by the TPP who could then decide to invest (or not) part of the endowment to deduct any MUs that the participant and the trustee earned during the TG. Our dependent measure of interpersonal trust was an adaptation of the general trust scale, a six-item, 7-points Likert scale (1 = strongly disagree, 7 = strongly agree) (e.g., "*I believe that Player 2 is trustworthy*", $\alpha = .96$)[42]. Higher scores on this scale mean that participants trusted their partner more.

Study 1b

Participants and Procedure

Participants ($N = 503$; 49% women, $M_{age} = 34.88$ years, $SD = 10.37$) were recruited from MTurk and completed the study for $1. In addition, participants had a chance to receive a lottery prize of $10 (five in total) in Phase 1, and a 0.01% chance to win a $2 prize depending on their decisions in Phase 2. We used a 2 (Corruption: honest vs. corrupt TPP) × 2 (Communication: present vs. absent) between-subjects design.

Die-rolling task (Phase 1). The die-rolling task was identical to Study 1a. Differently from Study 1a, we manipulated the possibility to receive a message from the partner prior to the decision in the PD to test whether the negative effects of corrupt institutions hold when a possibility for communication was present (vs. absent). Results from these treatments are presented in the SI.

Economic game with TPP (Phase 2). This phase was identical to Study 1a, except for the use of a prisoner's dilemma (PD) with TPP to assess cooperative behavior.

Study 2

Participants and Procedure

An a-priori power analysis (G*Power[56]) revealed a required sample size of 380 (190 dyads) to achieve statistical power of .95 to detect an effect size of $d = 0.24$ (two-sided). Participants ($N = 382$; 45.5% women, $M_{age} = 37.73$ years, $SD = 10.82$) were recruited through MTurk and completed the study for $2.50. Additionally, they could earn an extra bonus (up to $0.60) and could participate in a lottery to win a $2 prize (10 in total). We conducted the study through the platform SoPHIE[59], that enables real-time interactions among online participants.

Die-rolling task (Phase 1). Once logged into the platform, all participants were randomly matched in pairs and were assigned to either the role of first movers or second movers. Then, they were informed about their role in the game and received detailed instructions for the die-rolling task (see Study 1a for the general procedure of the game). To ensure that second movers might engage in dishonest behavior, we instructed them to either keep an actual die at hand while participating in the study or to open a suggested external web page that allowed to roll a fair six-sided die. The payoff scheme was disclosed to participants prior to the game, as in the previous studies. While first movers earned a fixed amount of $0.20 irrespective of scoring a double in each round, second movers could get triple that amount ($0.60) if they reported the same outcome of the die roll as the first mover. This removed any incentive for first movers to lie about their outcome, ruling out the possibility to engage in corrupt cooperation and to take advantage of the eventual dishonesty of the second movers. After each round, both players received real-time feedback on the reported outcomes of the die roll. At the end of the die-rolling task, we asked first movers to what extent they perceived the second mover as honest in reporting his/her score on a 7-point Likert scale (1 = completely dishonest, 7 = completely honest), and then reverse scored for easier interpretation of analyses. This constituted our independent variable.

Economic game with TPP (Phase 2). Then, participants engaged in a TPP DG. Participants who were previously playing as first movers (senders) were endowed with 100 lottery tickets (LT) and decided how much to give to an unknown receiver, while participants previously playing as second movers (TPP) decided how much to invest to reduce others' final earnings. Finally, we assessed interpersonal trust ($\alpha = .97$) as in the previous studies.

Studies 3-4

Participants and Procedure

An a priori sensitivity power analysis (G*Power[56]) revealed that a sample size of 100 dyads would give us statistical power of .80 to detect an effect size of $r = .24$ (one-sided). Participants in Study 3 ($N = 215$; 40.2% women, $M_{age} = 35.20$ years, $SD = 9.80$) and Study 4 ($N = 188$; 34% women, $M_{age} = 37.47$ years, $SD = 10.86$) were recruited through MTurk and completed a real-time interaction study for $2.50 in the platform SoPHIE[59]. Additionally, they could earn an extra bonus (up to $0.60) and could participate in a lottery to win a $2 prize (75 in total for each study). In a limited number of experimental sessions (7% in Study 3 and 8% in Study 4), participants were matched with the experimenter, who would then play as TPP. Such sessions are included in the current analyses. Results of analyses excluding such sessions were consistent both in terms of the main effect of perceived dishonesty on interpersonal trust and the indirect effect on cooperation via trust (see SI).

Die-rolling task (Phase 1). The procedure of the studies resembled the one adopted in Study 2 with one main difference in the incentive structure. As in Studies 1a-2, second movers would be rewarded ($0.60) only if their reported outcome matched with that of the first mover. However, participants were informed that at the end of each session, any money not awarded to the second mover in case he/she did not score a double in the dice rolling would be allocated to an experimental fund to benefit the collective. Thus, dishonest behavior of second movers directly resulted in the depletion of the common good. Specifically, Studies

3 and 4 involved two types of common goods to be exploited by an inflated report of doubles by the second mover. In Study 3, the money in the experimental fund was equally divided and allocated to all participants at the end of the data collection. In Study 4, it was donated to a pro-environmental charity that offsets CO_2 emissions (https://www.cooleffect.org). On average, participants playing the role of first mover in the die rolling task and subsequently the role of sender in the DG reported that the mission of the charity was moderately important for them (M = 5.16, SD = 1.49), as measured on a 7-point Likert scale (1 = not at all important, 7 = extremely important).

Economic game with TPP (Phase 2). Afterwards, participants engaged in a TPP DG as in previous studies to assess interpersonal trust (α = .93 - .95) and cooperation. In addition, we assessed expectations about punishment from the TPP asking senders to indicate how many LT they expected the third-party to invest to reduce the earnings of the other players in the DG (0-100).

Data Availability

The datasets generated and analyzed during the current studies will be available upon request for peer-review and will be made publicly available in the OSF repository upon publication.

Code Availability

The code used to analyze data will be available upon request for peer-review and will be made publicly available in the OSF repository upon publication.

Acknowledgements

We thank Daniel Balliet, Leonard Hoeft, and the members of the Amsterdam Cooperation Lab for helpful comments on the manuscript. Catherine Molho acknowledges IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program).

References

1. Harding, L. What are the Panama Papers? A guide to history's biggest data leak. *The Guardian* (2016).
2. Rose-Ackerman, S. Trust, honesty and corruption: Reflection on the state-building process. *Arch. Eur. Sociol.* **42**, 526–570 (2001).
3. Köbis, N. C., Van Prooijen, J. W., Righetti, F. & Van Lange, P. A. M. Prospection in individual and interpersonal corruption dilemmas. *Rev. Gen. Psychol.* **20**, 71–85 (2016).
4. Graycar, A. & Smith, R. G. *Handbook of global research and practice in corruption*. *Handbook of Global Research and Practice in Corruption* (Edward Elgar Publishing Ltd., 2011).
5. Mauro, P. Corruption and growth. *Q. J. Econ.* **110**, 681–712 (1995).
6. Rose-Ackerman, S. The economics of corruption. *J. Public Econ.* **4**, 187–203 (1975).
7. Weisel, O. & Shalvi, S. The collaborative roots of corruption. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10651–10656 (2015).
8. Gross, J., Leib, M., Offerman, T. & Shalvi, S. Ethical free riding: When honest people find dishonest partners. *Psychol. Sci.* **29**, 1956–1968 (2018).
9. Gino, F., Ayal, S. & Ariely, D. Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychol. Sci.* **20**, 393–398 (2009).
10. Köbis, N. C., van Prooijen, J. W., Righetti, F. & Van Lange, P. A. M. The Road to Bribery and Corruption: Slippery Slope or Steep Cliff? *Psychol. Sci.* **28**, 297–306 (2017).
11. Rothstein, B. & Eek, D. Political corruption and social trust: An experimental approach. *Ration. Soc.* **21**, 81–112 (2009).
12. Banerjee, R. Corruption, norm violation and decay in social capital. *J. Public Econ.*

137, 14–27 (2016).

13. Muthukrishna, M., Francois, P., Pourahmadi, S. & Henrich, J. Corrupting cooperation and how anti-corruption strategies may backfire. *Nat. Hum. Behav.* **1**, (2017).
14. Baumert, A., Halmburger, A., Rothmund, T. & Schemer, C. Everyday dynamics in generalized social and political trust. *J. Res. Pers.* **69**, 44–54 (2017).
15. Balliet, D. & van Lange, P. A. M. Trust, punishment, and cooperation across 18 societies: A meta-analysis. *Perspect. Psychol. Sci.* **8**, 363–379 (2013).
16. Ostrom, E. *Governing the commons: The evolution of institutions for collective action*. (Cambridge University Press, 1990).
17. Powers, S. T., van Schaik, C. P. & Lehmann, L. Cooperation in large-scale human societies—What, if anything, makes it unique, and how did it evolve? *Evolutionary Anthropology* (2021) doi:10.1002/evan.21909.
18. Yamagishi, T. The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51**, 110–116 (1986).
19. Hruschka, D. *et al.* Impartial Institutions, Pathogen Stress and the Expanding Social Network. *Hum. Nat.* **25**, 567–579 (2014).
20. Spadaro, G., Gangl, K., Van Prooijen, J.-W., Van Lange, P. A. M. & Mosso, C. O. Enhancing feelings of security: How institutional trust promotes interpersonal trust. *PLoS One* **15**, e0237934 (2020).
21. Macy, M. W. & Sato, Y. Trust, cooperation, and market formation in the U.S. and Japan. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 7214–7220 (2002).
22. Dinesen, P. T. Where you come from or where you live? Examining the cultural and institutional explanation of generalized trust using migration as a natural experiment. *Eur. Sociol. Rev.* **29**, 114–128 (2013).
23. Sønderskov, K. M. & Dinesen, P. T. Trusting the state, trusting each other? The effect

of institutional trust on social trust. *Polit. Behav.* **38**, 179–202 (2016).

24. Gächter, S. & Schulz, J. F. Intrinsic honesty and the prevalence of rule violations across societies. *Nature* **531**, 496–499 (2016).
25. Cohn, A., Maréchal, M. A., Tannenbaum, D. & Zünd, C. L. Civic honesty around the globe. *Science (80-.).* **365**, 70–73 (2019).
26. Keizer, K., Lindenberg, S. & Steg, L. The spreading of disorder. *Science (80-.).* **322**, 1681–1685 (2008).
27. Balliet, D. & Van Lange, P. A. M. Trust, conflict, and cooperation: A meta-analysis. *Psychol. Bull.* **139**, 1090–1112 (2013).
28. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994 (2000).
29. Molho, C., Tybur, J. M., Van Lange, P. A. M. & Balliet, D. Direct and indirect punishment of norm violations in daily life. *Nat. Commun.* **11**, 1–9 (2020).
30. Čábelková, I. & Hanousek, J. The power of negative thinking: Corruption, perception and willingness to bribe in Ukraine. *Appl. Econ.* **36**, 383–397 (2004).
31. Bowler, S. & Karp, J. A. Politicians, scandals, and trust in government. *Political Behavior* vol. 26 271–287 (2004).
32. Halmburger, A., Rothmund, R., Schulte, M. & Baumert, A. Psychological reactions to political scandals: Effects on emotions, trust, and the need for punishment. *undefined* (2012).
33. Gächter, S., Renner, E. & Sefton, M. The long-run benefits of punishment. *Science (80-.).* **322**, 1510 (2008).
34. Croson, R. & Gächter, S. The science of experimental economics. *J. Econ. Behav. Organ.* **73**, 122–131 (2010).
35. van Dijk, E. & De Dreu, C. K. W. Experimental games and social decision making.

Annu. Rev. Psychol. **72**, 415–438 (2021).

36. Soraperra, I. *et al.* The bad consequences of teamwork. *Econ. Lett.* **160**, 12–15 (2017).
37. Wouda, J., Bijlstra, G., Frankenhuis, W. E. & Wigboldus, D. H. J. The collaborative roots of corruption? A replication of weisel & shalvi (2015). *Collabra Psychol.* **3**, 1–3 (2017).
38. Choshen-Hillel, S., Shaw, A. & Caruso, E. M. Lying to appear honest. *J. Exp. Psychol. Gen.* **149**, 1719–1745 (2020).
39. Gerlach, P., Teodorescu, K. & Hertwig, R. The truth about lies: A meta-analysis on dishonest behavior. *Psychol. Bull.* **145**, 1–44 (2019).
40. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
41. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87 (2004).
42. Yamagishi, T. & Yamagishi, M. Trust and commitment in the United States and Japan. *Motiv. Emot.* **18**, 129–166 (1994).
43. Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity, and social history. *Games Econ. Behav.* **10**, 122–142 (1995).
44. Van Lange, P. A. M. & Kuhlman, D. M. Social Value Orientations and Impressions of Partner's Honesty and Intelligence: A Test of the Might Versus Morality Effect. *J. Pers. Soc. Psychol.* **67**, 126–141 (1994).
45. Preacher, K. J. & Hayes, A. F. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. in *Behavior Research Methods* vol. 40 879–891 (2008).
46. Knack, S. & Keefer, P. Does social capital have an economic payoff? *Q. J. Econ.* **112**, 1251–1288 (1997).
47. Lewicki, R. & Wiethoff, C. Trust, trust development, and trust repair. in *The Handbook*

of conflict resolution: Theory and practice (eds. Deutsch, M. & Coleman, P. T.) 104–136 (Jossey-Bass, 2000).

48. Van Lange, P. A. M. Generalized trust: Four lessons from genetics and culture. *Curr. Dir. Psychol. Sci.* **24**, 71–76 (2015).

49. Algan, Y. & Cahuc, P. Inherited trust and growth. *Am. Econ. Rev.* **100**, 2060–2092 (2010).

50. La Porta, R., Lopez-de-Silanes, F., Shleifer, A. & Vishny, R. W. Trust in large organizations. *Am. Econ. Rev.* **87**, 333–338 (1997).

51. Van Lange, P. A. M., Joireman, J. & Milinski, M. Climate Change: What Psychology Can Offer in Terms of Insights and Solutions. *Curr. Dir. Psychol. Sci.* **27**, 269–274 (2018).

52. van Prooijen, J. W. *The moral punishment instinct*. (Oxford University Press, 2017).

53. Stagnaro, M. N., Arechar, A. A. & Rand, D. G. From good institutions to generous citizens: Top-down incentives to cooperate promote subsequent prosociality but not norm enforcement. *Cognition* **167**, 212–254 (2017).

54. Marcin, I., Robalo, P. & Tausch, F. Institutional endogeneity and third-party punishment in social dilemmas. *J. Econ. Behav. Organ.* **161**, 243–264 (2019).

55. Stirrat, M. & Perrett, D. I. Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychol. Sci.* **21**, 349–354 (2010).

56. Erdfelder, E., FAul, F., Buchner, A. & Lang, A. G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* **41**, 1149–1160 (2009).

57. Casler, K., Bickel, L. & Hackett, E. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput. Human Behav.* **29**, 2156–2160 (2013).

58. Paolacci, G. & Chandler, J. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Curr. Dir. Psychol. Sci.* **23**, 184–188 (2014).

59. Hendriks, A. *SoPHIE-Software platform for human interaction experiments*. (2012).