

The Effect of Corruptible Institutions on Interpersonal Trust[★]

Alexander Dzionara^{★★a} and Mario Scharfbillig^a

^a *Johannes Gutenberg University of Mainz*

This version: February 23, 2023

[Click here for the latest version](#)

Trust and trustworthiness are malleable and depend on the institutional setting as well as prior experiences of individuals. A ubiquitous and effective feature established in many institutional and societal settings that increases trust in fair outcomes and norm-compliant behavior is allowing for the punishment of norm violators by third parties. However, an equally widespread phenomenon that might reduce the effectiveness of third-party punishment is corruption, as third parties with the power to punish may be influenced unduly. It is unclear, how these two factors interact. In this study, we investigate how the possibility of bribing a norm-enforcing third party affects interpersonal trust – which we define as beliefs about others’ trustworthiness – and trustworthiness itself. We conduct a laboratory experiment and compare behavior in three variations of the trust game; (i) without a punishment institution, (ii) with a punishment institution, and (iii) with a bribable punishment institution. We show that the possibility of bribing the punishment institution significantly reduces trustworthiness between individuals compared to a non-bribable punishment institution. While we find suggestive evidence for a similar reduction in trust, the results are not robust in all specifications. The effects on trust are strongest for individuals with low levels of authoritarianism and generally low trust in institutions. Overall, introducing options for bribery leads to a decrease in the payoff of trustors, unchanged payoffs for trustees, increased payoffs for punishers, and an overall reduction in welfare.

Keywords: trust, trustworthiness, bribe, trust game, punishment, corruption

JEL-Codes: D73, C91, D63, D02

*We thank Charles Bellemare, Sabine Kröger, Daniel Schunk, Valentin Wagner, and Niklas Witzig for helpful comments and discussions. Further, we thank participants from the research seminar of the IPP Mainz in 2019 for their helpful comments. We gratefully acknowledge funding by the interdisciplinary research unit IPP at the University of Mainz.

★★Corresponding author (dzionara@uni-mainz.de).

1 Introduction

Trust and trustworthiness are fundamental components of every interaction in society. Establishing and maintaining trust within a society is essential for economic success, civic engagement, and the stability of democracies. When “we trust someone or that someone is trustworthy, we implicitly mean that the probability that he will perform an action that is beneficial [...] is high enough for us to consider in engaging in some form of cooperation with him” (Gambetta, 1988, p. 217). Interpersonal trust in this sense is fragile and malleable, as experience (King-Casas et al., 2005), culture (Cronk, 2007), framing (Burnham et al., 2000), third-party intervention (Charness et al., 2008; Fiedler & Haruvy, 2017) and many other aspects have been shown to alter trusting behavior critically.

To foster trust, norm-compliant behavior, and societal cohesion, many societies implement third-party institutions and entrust them with the task of punishing violators of social norms (Henrich et al., 2006). The positive effects of such third-party punishment institutions for social capital have been extensively studied in the context of cooperation (e.g., Balafoutas et al., 2014; Balafoutas & Nikiforakis, 2012; Fehr & Fischbacher, 2004; Fehr, Fischbacher, & Gächter, 2002; Jordan et al., 2016; Rand & Nowak, 2013; Rockenbach & Milinski, 2006) and only sparsely in the context of trust (e.g., Bicchieri & Maras, 2022; Charness et al., 2008). This literature finds that cooperation rates between individuals are higher when institutional punishment by third parties is possible, compared to situations without third-party punishment institutions. Moreover, when given a choice, most people seem to anticipate the effects of these institutional frameworks and self-select into environments governed by punishment institutions over institution-free environments (Fehr & Williams, 2018; Gülerk et al., 2014; Nikiforakis & Mitchell, 2014).

However, a global phenomenon that might critically hamper the effectiveness of institutions in general and third-party punishment institutions in particular, is corruption. Recent estimates state that around 1\$ US Trillion per year is spent on corruption and bribes (Kaufmann, 2005). Besides the high direct economic costs, the behavioral consequences of corruption are less clear. While indirect spillover effects of previous experiences with corruption on trust have been shown in laboratory studies (Banerjee, 2016), the direct behavioral effects of interacting under a potentially corrupt institution have not been investigated. This, however, is a central problem, as corrupt elites often have extensive influence over institutions and can shape the rules under which punishment is meted out in a self-serving manner (Acemoglu & Robinson, 2008). In fact, in controlled laboratory experiments, bribes have been shown to lead to distortions in judgment by third parties (Gneezy et al., 2019). Individuals might anticipate this distortion of judgment, limiting the positive effect of third-party punishment institutions to increase trust. Although third-party punishment has been hailed in the literature as a fundamental driver

of societal progress due to its power to sustain cooperation and trust, the potential for corruptible punishers and their possibly detrimental consequences for social capital have received less attention. We hypothesize that bribes could decrease the positive effect of punishment and potentially even reduce trust and trustworthiness to levels below an institution-free environment.

In this paper, we address the following questions: How does a bribable third-party punisher affect interpersonal trust and trustworthiness compared to a non-bribable punisher? How do trust and trustworthiness compare between an environment with a bribable third-party punisher and an institution-free setting?

We conduct a laboratory experiment with three treatments based on the trust game paradigm (Berg et al., 1995) to measure how interpersonal trust is affected by different institutional frameworks. The first treatment is a standard *baseline* trust game to measure interpersonal trust and trustworthiness in the absence of a third-party punishment institution. In the baseline treatment, trustors decide how much of an initial endowment they send to a trustee. The amount sent is tripled by the experimenter, and the trustees subsequently decide if and how much of the received amount they return to the trustor. In the *punish* treatment, we introduce a third-party punisher who observes the trustor's sending and the trustee's sending back behavior. Subsequently, the third party can decide whether and how severely they want to punish the trustee. We implement a third treatment – called *bribe* – to study the effect of a bribable institution on trust and trustworthiness. In this treatment, the trustee can not only send money back to the trustor but to the punisher as well. The punishers can decide whether to accept or reject the amount sent to them and still have the option to punish the trustee. In the *punish* and the *bribe* treatments, the endowment of the punisher is as high as the maximally achievable payoff by either the trustor or the trustee. Therefore, the punisher should have no reason to punish out of envy, and the trustee has no motivation to send money to the punisher due to inequality aversion.

To measure trust and trustworthiness, we follow Sapienza et al. (2013) and define a trustor's expectations about the amount returned by a trustee as our measure of trust. This definition is becoming more widely used in the recent literature (e.g., Bartling et al., 2021), and has been shown to correspond best to other measures of generalized trust, such as the survey question from the World Values Survey (see earlier discussions in e.g., Glaeser et al., 2000). Furthermore, it has the advantage that it is less confounded by other motivations like altruism and inequality aversion when compared to the amount sent by the trustor – the measure traditionally used to quantify trust – (Sapienza et al., 2013). We define the amount returned by the trustee as trustworthiness.

The main finding of this paper relates to the effect of the introduction of the bribe channel on trust and trustworthiness. We find suggestive evidence that introducing the possibility to bribe a punisher reduces trust – measured as the return expectation – by

about 3.2 percentage points (pp.) of the share expected back. When controlling for individual covariates, the effect becomes larger (6.9 pp.) and increases in significance. However, we do not find an effect when defining trust as the amount sent by the trustors. Thus, the results for trust are not entirely conclusive. However, we find, that trustworthiness is significantly lower (by 6.0 pp.) in the treatment with a bribable punisher compared to a non-bribable one. Finally, comparing the baseline treatment with the bribe treatment, we find significantly lower trustworthiness in the bribe treatment – participants send back 6.9 pp. less of the received amount – but no effect on trust. Overall, we conclude that a bribable punisher seems to reduce trustees’ trustworthiness but that the effects on the trustors seem less robust.

A second finding relates to the difference between trust and trustworthiness when comparing the baseline to the punishment treatment. A previous study by Charness et al. (2008) finds that introducing a third-party punisher increases trust and trustworthiness compared to a baseline game. We can not replicate this result in our experiment, as trust and trustworthiness between the baseline and the punishment treatment are not significantly different. Therefore, we show that the effects of third-party punishment in the trust game might not be as robust as previously thought. We discuss a few possible explanations for the diverging results.

Finally, we analyze heterogeneous treatment effects of the introduction of the bribery channel. We conjecture that institutional preferences and experiences shape how participants in our experiment react to the different institutional frameworks in our experiment. Therefore, we elicit two measures related to institutional preferences at the end of the experiment: the Right-Wing Authoritarianism Scale (Beierlein et al., 2014), which measures the preference for authoritarian institutions and punishment, as well as questions on general institutional trust (e.g., trust toward the government, police, judges, etc.). We find that individuals with low levels of authoritarianism and low general institutional trust drive the treatment effects of bribing on trust. We interpret these findings as evidence that trust in institutions is essential in allowing people to also trust individuals. Conversely, individuals with low levels of trust in institutions are vulnerable to the possibility of corruption. Therefore, we emphasize that the effectiveness of institutions – put in place to foster norm-conforming behavior – depends not only on the institutions themselves but also on the trust individuals place in them. When corruption is a reality, it might be vital to invest in the credibility and accountability of these institutions.

Overall this paper is related to two distinct strands of literature. The first literature investigates the effects of third-party punishment in the context of social dilemmas. While the majority of the studies in the context of public goods games find cooperation-enhancing effects (e.g., Balafoutas et al., 2014; Fehr & Fischbacher, 2004; Jordan et al., 2016; Rockenbach & Milinski, 2006), punishment has also been shown to have potential negative side-effects. One issue closely related to bribery is that third-party punishers

can act norm-enforcingly (to foster pro-social behavior), but – without detailed oversight – they can also act anti-socially and punish pro-social behavior (Herrmann et al., 2008). This might potentially hamper their positive effect. Despite this, most of the laboratory evidence in the public goods game points to pro-social punisher behavior (in the absence of bribery). There is less evidence on the effect of third-party punishers in the trust game. In the standard Berg et al. (1995) trust game, trustworthiness, i.e., reciprocating a trusting action, is a social norm, while trust is not (Bicchieri et al., 2011). Third-party punishers motivated by their altruistic motivation to punish norm-violating behavior should therefore punish non-trustworthy behavior. Subsequently, this should increase the trust a trustor places in a potential trustee in an institutional framework with third-party punishment. Two previous studies investigate the effect of third-party intervention in the trust game. Charness et al. (2008) find an increase in trust and trustworthiness when introducing a voted-for third party with the power to punish or reward in the trust game. Fiedler and Haruvy (2017) analyze the mechanism by which third-party intervention affects trusting behavior and largely attribute behavioral changes to the monitoring property of the third party. Conversely, Fehr and Schneider (2010) do not find an “observer” effect in the trust game. Overall, while the effect of third-party punishment in the public goods game seems well established, there is less evidence in the trust game with mixed results.

The second strand of literature investigates the possible effects of corruption on social capital. Two papers are closely related to ours. The first paper by Muthukrishna et al. (2017) shows that in an institutional punishment public goods game, where one of the players in the game is randomly chosen to be a punisher, cooperation is significantly reduced when the possibility of bribing the punisher exists. Our study is distinctly different in that it measures trust rather than cooperation. While it has been argued that trust and cooperation are strongly related, the evidence is mixed.¹ Secondly, the punisher in Muthukrishna et al. (2017) is not an “outside” third party but a randomly selected member of the interacting group. Thus, the punisher in their setting benefits from mutual cooperation. Conversely, the punishers do not gain anything from fostering trust and trustworthiness in our setting. Therefore, we believe that we capture many real-life situations in which institutional punishers (such as the police or judges) are not affected directly by non-norm-compliant behavior of the parties they oversee. Thirdly, we run a one-shot game that does not allow for reputation-building and potential reciprocity that might affect cooperation or trust. A second paper by Banerjee (2016) investigates spillover effects of previously experienced corruption in a subsequent ultimatum game. They show that altruistic sharing is reduced after participants experience bribery. Schw-

1 Fehr, Fischbacher, von Rosenbladt, et al. (2002); Bellemare and Kröger (2007); Gächter et al. (2004) find mixed but mostly positive results, and Bauer et al. (2019) find no relationship between trust and cooperation.

erter and Zimmermann (2020), in a similar design, show that previous negative social experiences affect trust. Compared to these two studies, we are interested in the *direct* effect of knowledge about the possibility of corruption on trust and trustworthiness. We, therefore, contribute to this literature by showing how trust and trustworthiness – two integral components of social capital – are affected by the mere existence of the option to bribe an outside third-party punisher.

2 Experimental Design

2.1 Experimental Setup and Treatments

We implement three different versions of a trust game: a *baseline* specification of the standard setup (Berg et al., 1995), a *punish* treatment adding punishment by a third party, and a *bribe* treatment where the trustee can send side payments to the third party. We include the baseline treatment to retest the validity of the result that a third party with the possibility to punish non-trustworthy behavior increases interpersonal trust. By comparing the bribe treatment to the punishment treatment, we can estimate the effect of a bribe channel in a punishment institution on interpersonal trust.

At the beginning of each session, participants are randomly assigned to a treatment and a two- or three-player group according to their treatment. After reading the instructions, they make decisions in all roles of the game: First as a trustor, then as a trustee, and – if possible in their treatment – as a punisher. They make their decisions in the same order and under role uncertainty, i.e., they make decisions in every role but are informed which role will be payoff relevant for them at the end of the experiment.

Payoffs are calculated according to the assignment of roles, and the participants are shown a final screen informing them about their choices, the relevant choices of the other members in their group, and their own final payoff. In the instructions, we use neutral framing to describe all roles (A, B, C) and actions (i.e., “send money” rather than “bribe”).² Figures A7 and A8 in Appendix A.2.2 show schematic representations of the punishment and the bribe treatment.

2.1.1 Experimental Treatments and Payoff. The baseline treatment is a standard Berg et al. (1995) one-shot two-player trust game. The *trustor* starts with a fixed endowment of $E_i^{\text{Trustor}} = 20$ monetary units and may send $s_i^{\text{Trustor}} \in \{0, 4, 8, 12, 16, 20\}$ to the *trustee*. The amount sent is subsequently multiplied by $m = 3$. The *trustee* then decides how much of the received amount $m * s_i^{\text{Trustor}}$ they want to return to the *trustor*. We collect the return

2 See Appendix A.2.1 for a translated version of our instructions of the bribe treatment. The original instructions for all treatments in German are available upon request from the authors.

behavior using the strategy method and thus ask the *trustee* to specify a response strategy for each possibly received amount $r_i^{\text{Trustee}} = (r_{s=4}, r_{s=8}, \dots, r_{s=20})$. The payoff functions for the trustor and the trustee in the baseline treatment are:

$$\begin{aligned}\Pi_{\text{Baseline}}^{\text{Trustor}} &= E^{\text{Trustor}} - s^{\text{Trustor}} + r_{s=s^{\text{Trustor}}}^{\text{Trustee}} \\ \Pi_{\text{Baseline}}^{\text{Trustee}} &= m * s^{\text{Trustor}} - r_{s=s^{\text{Trustor}}}^{\text{Trustee}}\end{aligned}$$

A third party – the punisher – is introduced in the punishment treatment. The punisher is endowed with $E^{\text{Punisher}} = 60$ points and can make a costly choice to punish a *trustee* with p punishment points after seeing a partnered trustee’s return strategy r_i^{Trustee} . To not overload participants with an overwhelming strategy elicitation, we elicit the punishment strategy of individual j conditional on one randomly matched individual i ’s strategy in the role of *trustee*.³ That is, punishers have to specify their strategy $p_j^{\text{Punisher}} = (p(r_{i,s=4}), p(r_{i,s=8}), \dots, p(r_{i,s=20}))$. These punishment points are multiplied by a punishment parameter $\theta = 2$ and deducted from the *trustee*’s final payoff. Thus, punishment is socially inefficient and costly to the *punisher*. We implement a floor of 0 in the payoff function of the trustee to prohibit negative results. The payoff functions in the punishment treatment are:

$$\begin{aligned}\Pi_{\text{Punishment}}^{\text{Trustor}} &= E^{\text{Trustor}} - s^{\text{Trustor}} + r_{s=s^{\text{Trustor}}}^{\text{Trustee}} \\ \Pi_{\text{Punishment}}^{\text{Trustee}} &= \max(0, m * s^{\text{Trustor}} - r_{i,s=s^{\text{Trustor}}}^{\text{Trustee}} - \theta * p_{j,s=s^{\text{Trustor}}}^{\text{Trustee}}) \\ \Pi_{\text{Punishment}}^{\text{Punisher}} &= E^{\text{Punisher}} - p_{j,s=s^{\text{Trustor}}}^{\text{Punisher}}\end{aligned}$$

A vital component of the punishment treatment design is to ensure that the punisher’s only motivation to deduct points is to punish norm-violating behavior. Thus, we set the punisher’s endowment to the same amount of points the trustee would have if the trustor sent everything to them. Therefore, no motivation for the punisher to punish out of envy or inequality aversion exists, e.g., due to lower endowments or final payoffs relative to the other players. Moreover, the game is an anonymous one-shot game, such that there is no motivation to prevent higher-order punishment or to use punishment as a signal for trustworthiness, as there is no follow-up interaction. In this experiment, punishment is costly and levered, i.e., for every point the punisher spends on punishment, the trustee is deducted twice the points. Finally, the possibility of counter-punishment is excluded, which has been shown to reduce the efficiency of punishment as a driver of positive norm enforcement in public good games (Nikiforakis, 2008).

In designing the bribe treatment, we rely on the insights from Gneezy et al. (2019), who study when bribes influence behavior. They apply the literature on moral wiggles (Bénabou & Tirole, 2016; Dana et al., 2007; Haisley & Weber, 2010; Kunda, 1990)

3 Otherwise, even in a discrete choice setting and allowing for sending back amounts of full integers we would need to elicit 185 potential responses. As we are not interested in analyzing punishment behavior, we decided to follow this more feasible approach.

to bribery and find that bribed individuals exploit existing moral wiggle room to distort their judgment about norm violations in favor of a person who sent them a bribe. We relate this finding to the trust game, where sending back *something* is a clear social norm (Bicchieri et al., 2011). However, the social norm on the *amount* sent back is not entirely clear,⁴ creating a potential to utilize one’s moral wiggle room when assessing others’ actions. This ambiguity generates a norm space and allows punishers to utilize this moral wiggle room when deciding on their action after receiving a bribe

In real-life situations, bribed individuals are often able to reject unwanted bribes. Therefore, we allow punishers to reject bribes, consequently removing the channel of punishment motivated by reciprocity if the punisher does not want to be bribed but cannot reject the bribe. This approach follows Abbink et al. (2002) and Abbink et al. (2014), who study bribery in different institutional settings. To eliminate any behindness-averse inequality-driven motivation for punishment and, therefore, for bribing, we endow the punisher with an amount equal to the amount the trustee receives when the trustor sends everything. This strengthens a punisher’s credibility as a norm enforcer, and sending a bribe in our setting can only be motivated by fear of punishment due to non-norm conforming sending-back behavior.⁵ Gneezy et al. (2019) find that bribes distort behavior the most when bribed parties can keep the bribes only when their behavior favors the briber. In our setting, the punishers can decide to keep the bribe independent of their subsequent decisions; thus, bribes do not guarantee a favorable judgment for the briber. This might reduce a briber’s belief about the effectiveness of bribes, as punishers’ financial incentives are not directly connected to their judgment. We thus identify the lower bound of the effect that the possibility to bribe has on trust enforcement through punishment.

In the bribe treatment, the *trustee* has the additional option to send money b_i^{Trustee} to the *punisher*. Symmetric to the elicitation of the return strategy in the other treatments, we elicit bribing by the strategy method and thus ask a *trustee* to specify a return and bribe strategy jointly on one screen for each possibly received amount. The overall strategy is given by $b_i^{\text{Trustee}} = ([b, r]_{s=4}, [b, r]_{s=8}, \dots, [b, r]_{s=20})$. The amount $b[b]_i^{\text{Trustee}}$ is sent in full to the punisher.⁶ The *punisher* can decide to accept or to reject $b[b]_i^{\text{Trustee}}$, which re-

4 Potential candidates for a norm of the amount sent back include the initial amount sent by the trustor, a half-half split of the surplus, the equality-inducing amount, or the total surplus generated by the amount sent.

5 Gneezy et al. (2019) show that in their treatment “HighWage”, where punishers receive a large endowment, average bribes are significantly lower than in the other settings where the endowment of the punisher is below the other participants involved.

6 To make this transfer socially inefficient, we could introduce a factor $\zeta < 1$, however, we refrain from this in this study.

sults in $b[b]_i^{\text{Trustee}}$ being sent back to the *trustee* multiplied by a factor $\gamma = 0.8$.⁷ A *punisher* “*j*” has to define a response strategy to one matched *trustee*’s strategy b_i^{Trustee} . Specifically, the punisher has to specify $p_j^{\text{Punisher}} = (p([b, r]_{i,s=4}), p([b, r]_{i,s=8}), \dots, p([b, r]_{i,s=20}))$, where $p = \langle \text{accept bribe}(\cdot), \text{punish points}(\cdot) \rangle$.

The payoff functions in the bribe treatment are:

$$\begin{aligned} \Pi_{\text{Bribe}}^{\text{Trustor}} &= E^{\text{Trustor}} - s^{\text{Trustor}} + r_{s=s^{\text{Trustor}}}^{\text{Trustee}} \\ \Pi_{\text{Bribe}}^{\text{Trustee}} &= \begin{cases} \max(0, m * s^{\text{Trustor}} - b[r]_{i,s=s^{\text{Trustor}}}^{\text{Trustee}} - \theta * p_{j,s=s^{\text{Trustor}}}^{\text{Trustee}} - b[b]_{i,s=s^{\text{Trustor}}}^{\text{Trustee}}), & \text{if punisher accepts } b \\ \max(0, m * s^{\text{Trustor}} - b[r]_{i,s=s^{\text{Trustor}}}^{\text{Trustee}} - \theta * p_{j,s=s^{\text{Trustor}}}^{\text{Trustee}} + (1 - \gamma)b[b]_{i,s=s^{\text{Trustor}}}^{\text{Trustee}}), & \text{if punisher rejects } b \end{cases} \\ \Pi_{\text{Bribe}}^{\text{Punisher}} &= \begin{cases} E^{\text{Punisher}} - p_{j,s=s^{\text{Trustor}}}^{\text{Punisher}} + b[b]_{i,s=s^{\text{Trustor}}}^{\text{Trustee}} & \text{if punisher accepts } b \\ E^{\text{Punisher}} - p_{j,s=s^{\text{Trustor}}}^{\text{Punisher}} & \text{if punisher rejects } b \end{cases} \end{aligned}$$

To measure trust, we elicit the amount expected back by the trustee, as proposed by Sapienza et al. (2013). They show that a trustor’s beliefs about the amount returned by the trustee are a robust and reliable measure of interpersonal trust. We agree with the assessment by Sapienza et al. (2013) – echoing the definition by Gambetta (1988) – that the return expectations of the trustor best capture the nature of trust in experimental settings. Those beliefs are not affected by other factors that previous studies have found to influence the sending decision, such as inequality aversion, reciprocity (Cox, 2004) or ambiguity attitudes (Li et al., 2019). Therefore, a belief-based definition of trust seems more natural than a choice-based definition. We define our measure of trustworthiness as the fraction returned conditional on the amount sent, as done in the previous literature (Ben-Ner & Halldorsson, 2010).

To measure the belief-based component of trust, Sapienza et al. (2013) suggest an incentive-compatible⁸ mechanism to elicit beliefs. Specifically, participants are asked to provide their expectations about the amount returned by a trustee for each potential sending decision as a trustor. They receive an additional payoff of 5 points for each expectation that is within a 10 percent error bound around the true value a trustee matched with them specified in their return strategy. In the bribe treatment, we additionally elicit beliefs about the amount sent to the punisher, which we incentivize analogously to the beliefs about the trustee’s sending-back decisions.

7 We thus follow the approach in the literature on bribery in laboratory experiments to signify the potential loss due to a rejected bribe (cf. Abbink et al., 2002; Abbink et al., 2014).

8 According to Li et al. (2019) the mechanism might not be fully incentive compatible in the case of extreme beliefs.

2.2 Experimental Procedures

We conducted eight sessions with 150 participants in October 2018. All sessions took place in the MABELLA experimental laboratory at the University of Mainz. Participants were recruited via ORSEE, (Greiner, 2015) and were students of the University of Mainz. The experiment was programmed in oTree (Chen et al., 2016). During the experiment, the experimental currency was called “points”, and 10 points were equal to 1 Euro. At the end of each session, participants were individually paid in cash. The average session lasted 40 minutes (min = 31; max = 47), and participants earned 8,30 Euro on average.⁹ At the start of each session, participants randomly drew a card, allocating them to a working cubicle in the laboratory. There they received printed instructions about the study and were directed to read them before starting the experiment. At the beginning of the experiment, participants answered control questions testing their understanding of the experimental protocol, which they had to answer correctly to proceed to the main experiment. If they submitted a wrong answer, they were instructed to recheck their answers and raise their hand if they had problems understanding the instructions. At the end of the experiment, participants answered survey questions on altruism, positive and negative reciprocity from the Global Preference Survey module (Falk et al., 2022), institutional trust, Right-Wing Authoritarianism, prestige-dominance, gender, age, study subject, and study duration. We measure Right-Wing Authoritarianism (RWA) using the German RWA questionnaire (Beierlein et al., 2014).¹⁰ Institutional trust is measured by asking participants to state their level of trust towards an array of different institutions of the judicial, executive, and legislative system on a five-point Likert scale, a measure taken from the Life in Transition Survey (European Bank for Reconstruction and Development & World Bank, 2011).¹¹ We apply this approach, as these or similar measures are frequently used in political science to measure institutional trust (e.g., Chang & Chu, 2006; Mishler & Rose, 2001).

3 Predictions

According to the previous literature on the effects of punishment on cooperation and specifically the results by Charness et al. (2008) in a trust game, punishment increases norm-adherent behavior in general and trustworthiness in particular. Hence, in our punishment treatment, where non-norm-adherent behavior can be penalized by a third party, the trustworthiness of the trustee should be higher compared to a setting without the

⁹ This is equal to roughly 1.5 times the legal minimum wage at the time in Germany.

¹⁰ For a translated version of the questionnaire, see Appendix A.2.3.

¹¹ For translated version of the questionnaire see Appendix A.2.3.

possibility of punishment. We postulated that trust is measured by trustors' beliefs about trustees' trustworthiness. Eliciting the trustworthiness behavior of the trustees via the strategy method, we can make predictions about the fraction returned conditional on each potential sending amount. Given that trustors in our game anticipate the increased trustworthiness of the trustees, their trust should be higher in the treatment that allows for third-party punishment. We thus expect trustors in the baseline scenario to expect back less money compared to the punishment treatment. We thus formulate the following hypotheses:

Hypothesis 1a. *Introducing third-party punishment increases trust compared to the baseline. The fraction of money a trustor expects back given a certain amount sent is higher in the treatment with punishment than in the baseline.*

Hypothesis 1b. *Introducing third-party punishment increases trustworthiness compared to the baseline. The fraction of money sent back by a trustee given a certain amount received is higher in the treatment with punishment than in the baseline*

We now compare the bribe treatment to the punishment treatment. In our bribe treatment, we expect that trustees anticipate the potential to distort the judgment by the punishers in their favor. Hence, trustees know that sending a bribe might tilt a punisher's judgment in their favor. This would result in a lower (or even absent) punishment when trustees act in a non-norm-abiding and more self-serving manner by keeping more money to themselves. Anticipating this behavior by the punisher, a trustee would act accordingly, send a bribe, and send back less to the trustor. As trustees become potentially less trustworthy, we expect trustors to anticipate this behavior and become less trusting. Thus, we expect the trustors in the bribe scenario to expect back less for each amount sent compared to the punishment treatment.

Hypothesis 2a. *Introducing the option for side payments by the trustee to the punisher decreases the trust a trustor places in a trustee, compared to the pure punishment treatment. The fraction the trustor expects back given a certain amount sent is lower in the bribe treatment than in the punishment treatment.*

Hypothesis 2b. *Introducing the option for side payments by the trustee to the punisher decreases the trustworthiness of trustees compared to the pure punishment treatment. The fraction a trustee sends back when receiving a certain amount is lower in the bribe treatment compared to the punishment treatment.*

Beyond the main average expected treatment effects, individual preferences and attitudes might affect individuals' trust differently under different institutions (see e.g., the evidence in Murtin et al., 2018). Specifically, individuals that prefer strong institutions

and believe in their effectiveness and necessity should react positively to the introduction of a punishment institution. Introducing a bribing channel should not affect their trusting behavior significantly, as their positive attitudes towards strong institutions do not change. However, the effects on individuals with a strong aversion to such institutions might be different; in an environment with a bribable punisher, it is clear that the trustees can potentially avoid punishment for non-norm-abiding behavior. They can send a bribe, hope for a more favorable judgment by the punisher, and send back less to the trustor. Trustors with an aversion to authoritarian institutions will believe that these institutions will accept the bribe, turn a blind eye to the trustees' behavior and let them get away with their non-norm-abiding behavior. Hence, they will react more strongly to the introduction of the bribe channel and trust less compared to an institution without one. We capture this notion in the following hypothesis:

Hypothesis 3. *In the bribe treatment, the stronger the aversion against authorities, the larger the reduction in trust compared to an environment without the option to bribe.*

A second channel via which individual characteristics could affect the reaction to our treatment is an individual's general trust in political institutions. As Schwerter and Zimmermann (2020) point out, social experiences seem to have a significant effect on trust. Furthermore, Engl et al. (2021) show that institutions can have spillover effects on cooperative behavior and beliefs. Building on their results obtained in the lab, we extend this argumentation by measuring trust in institutions outside the lab. In case experiences in the lab shape decisions, it seems reasonable that experiences outside the lab and their effect on institutional trust might shape how participants' trust is affected by different institutional frameworks inside the lab. While we can not disentangle whether our institutional trust questions capture experiences or are shaped by other socio-political factors and beliefs, this analysis allows us to peek into the effect of attitudes outside the lab on behavior in the lab. We conjecture that individuals with low trust in political institutions might think a person in power is more easily corruptible. Hence, we predict that trustors with low institutional trust will react to the introduction of the bribe channel by lowering their trust more than in the punishment setting. This leads to the following hypothesis.

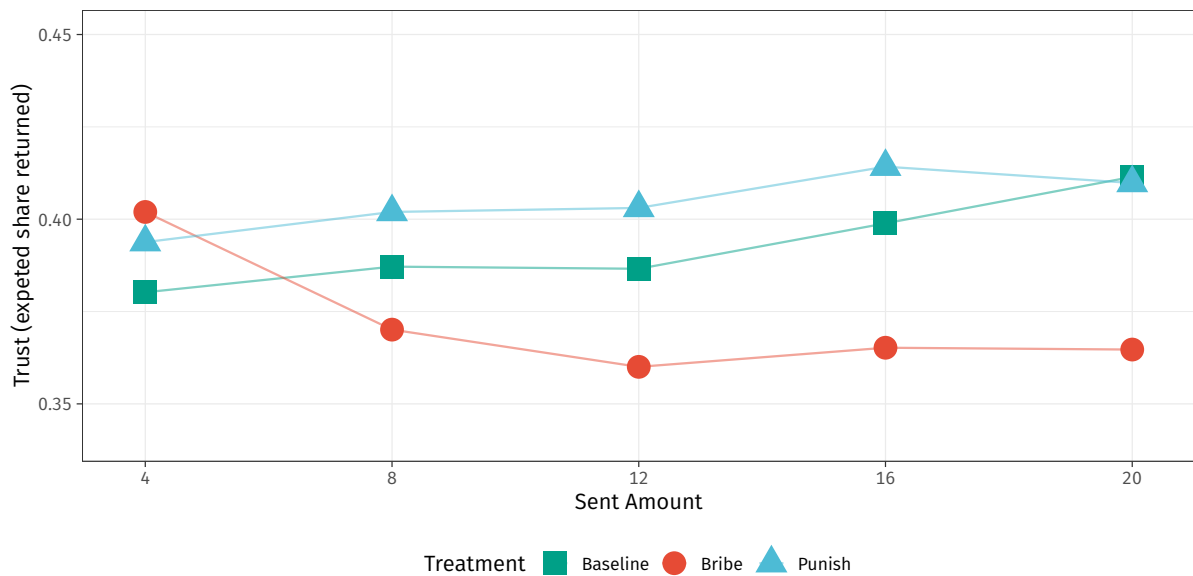
Hypothesis 4. *In the bribe treatment, the lower the trust in political institutions the larger the reduction in trust compared to an environment without the option to bribe.*

4 Results

The results section is organized as follows: we first present visual evidence for our main hypotheses before turning to a more detailed regression analysis. Afterward, we inves-

tigate potential mechanisms that might affect our results by exploring participant heterogeneity concerning attitudes towards authorities and their trust in them. Afterward, we briefly investigate punishment behavior and bribes. Finally, we conclude the section by comparing the average payoff in each role and the overall generated surplus across treatments to capture the distributional and welfare effects of corruption on trust.

The key hypotheses of our paper concern the effects of different institutional settings on trust and trustworthiness. Figure 1 displays our measure of trust in all treatments, and Figure 2 the trustworthiness in all treatments. In each graph, the green line and squares represent the baseline treatment, the blue line and triangles the punishment treatment, and the red line and circles the bribe treatment. Figure 1 displays the mean expectations about the fraction returned for all possible sending amounts in each treatment. To facilitate interpretation, we represent our result as fractions of the trustee's received amount (i.e., the tripled sent amount).



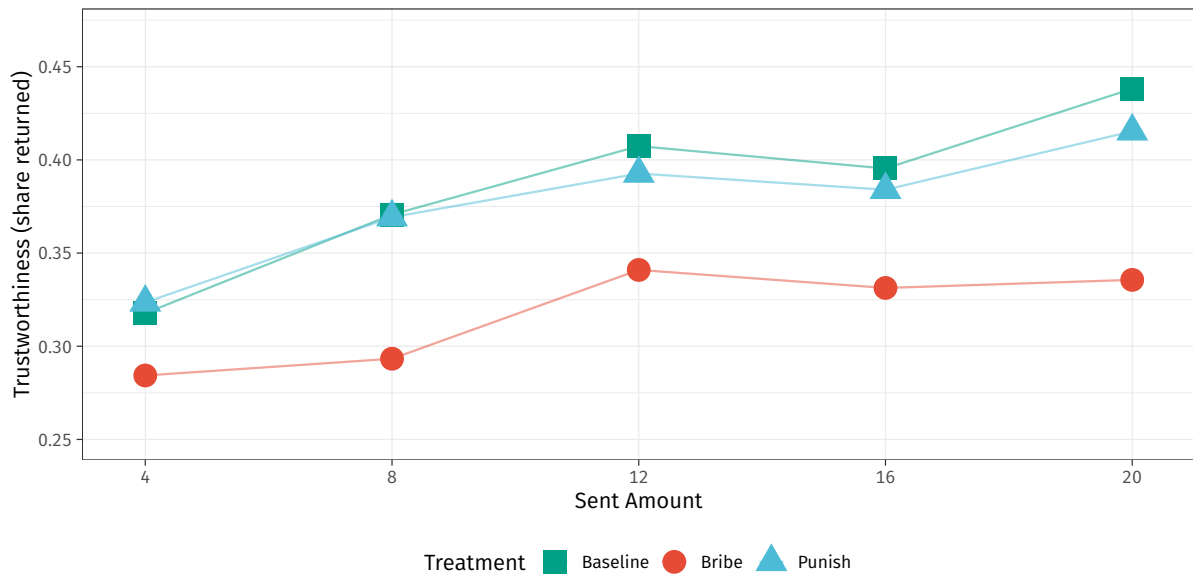
Notes: The graph depicts the mean expected share returned based on the strategy elicitation method in each treatment.

Figure 1. Trust across treatments

Comparing trust in our treatments (Figure 1), we find that the average fraction expected back in the punishment treatment is the highest overall, with an average expectation of 40.5% of the amount received by the trustee to be returned. In the baseline treatment, we observe slightly lower levels of trust, namely 39.3% average expected returns. In the bribe treatment, we see the lowest levels of trust with 37.2%. Additionally, we observe that the level of trust in the bribe treatment is consistently lower than in the punishment and the baseline treatment above the sent amount of 4. While trust in the baseline treatment seems to be lower than in the punishment treatment for low to medium amounts, we observe that they are almost identical for the highest possible

sending amount. Overall, we find a 3.21 pp. decrease in trust in a punishment institution with an additional bribe channel compared to the punishment-only treatment. This corresponds to a change of 0.24 standard deviations. While we find a slight increase in trust in a punishment institution compared to the baseline treatment, they seem almost identical, with a difference of only 1.2 pp. or 0.09 standard deviations.

Figure 1 shows another interesting pattern: the return expectation of trustors for the sending amount of 4 are almost identical across all treatments. This finding could be related to a feature about the beliefs that Sapienza et al. (2013) point out when proposing this measure. They argue that sending low amounts could rather be interpreted as “an act of charity, more than an act of trust” (Sapienza et al., 2013, p. 1325). Thus, the return expectations to sending a small fraction of the endowment (which leads to a payoff for the trustee that is still lower than the payoff of the trustor) might capture different beliefs not related to trust. To corroborate this interpretation, we compare the significance of robust percentage bend correlation coefficients adjusted for multiple testing and find no significant (at the 1%-level) correlations between the return expectation when sending an amount of 4 with the return expectation for 16 and 20 (Baseline), 8, 12, 16, and 20 (Bribe), 12, 16, and 20 (Punish)¹².



Notes: The graph depicts the mean share returned based on the strategy elicitation method in each treatment.

Figure 2. Trustworthiness

For trustworthiness, we observe similar overall results compared to our measure of trust (Figure 2). The average fraction sent back over all decisions made by the trustees is the highest in the baseline treatment with 38.6%. In the punishment treatment, trustworthiness is almost identical, with 37.7% returned on average. The patterns of the two

¹² See Figure A6 in Appendix A.1 for a correlation heatmap

treatments are similar and seem almost indistinguishable. However, in the bribe treatment, only 31.7% of the received amount is returned on average. Overall, trustworthiness in the bribe treatment is 6 pp. lower compared to the punishment treatment. This corresponds to a change of 0.43 standard deviations. Hence, trustworthiness seems to be reduced in the bribe treatment compared to the punishment-only treatment. Contrary to our expectations, we do not find any apparent difference in trustworthiness between the punishment and the baseline treatment.

4.1 Regression Analysis

4.1.1 Trust. Table 1 presents the regression results of the main treatment effects on trust. In all columns, we estimate regression models with individual random effects and heteroscedasticity robust standard errors clustered at the individual level to account for multiple observations per individual due to the strategy method. We additionally include dummies for each sending amount to control for potential scale effects. We are interested in each treatment effect separately, so we estimate pairwise regressions and exclude one treatment sample for each estimation. Hence, in columns 1, 3, and 5, we estimate a regression of the following functional form:

$$Trust_{id} = \beta_0 + \beta_1 Treatment_i + \gamma d_t + c_i + u_{id}$$

$Trust_{id}$ is the expected fraction returned of individual i for each possible sending amount d and $Treatment_i$ a dummy indicating treatment status. In columns 2, 4, and 6, we control for individual characteristics¹³ of the participants.

Does a punishment institution increase trust? While we observe a positive coefficient for the treatment dummy of punishment on trust (Table 1, column 1), this effect is insignificant and small (0.011; $p = 0.66$) and robust to the inclusion of individual characteristics (Table 1, column 2). Overall, this corroborates the results observed in the graphical analysis and leads us to conclude that we do not find sufficient evidence to support **Hypothesis 1a**, namely that punishment can meaningfully increase trust. This result is not in line with the results obtained by Charness et al. (2008), who find a trust-increasing effect of punishment. In contrast to our study, they use the amount sent by the trustor as a measure of trust. However using their measure does not change the results.¹⁴

13 The individual controls are: gender, age, dummies for the study subject, number of semesters studied, and understanding of the game operationalized by the number of errors per control question a participant made. For summary statistics on these measures and pairwise test of differences between the treatments see Table A1 and Table A2 in the appendix.

14 Using their methodology and measurement to analyze our experimental data, we find no statistically significant effects. A Mann-Whitney-U test finds no significant difference between punishment and baseline treatments ($W = 1122$, $p = 0.77$, testing one-sided for a difference larger than 0 in the average amount sent).

Table 1. Treatment effects on trust

	Baseline vs. Punish		Bribe vs. Punish		Bribe vs. Baseline	
	(1)	(2)	(3)	(4)	(5)	(6)
Punishment	0.012 (0.027)	0.022 (0.028)				
Bribe			-0.032 (0.025)	-0.069** (0.028)	-0.020 (0.025)	-0.010 (0.033)
Amount Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Random Effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls		Yes		Yes		Yes
N	495	495	510	510	495	495
Unique N	99	99	102	102	99	99

Notes: Estimates are from a random effects model. The dependent variable is the trustee's expectation about the fraction of money returned. Punishment and Bribe are dummy variables indicating treatment status. In column 1 and 2 the omitted sample is Bribe, in 3 and 4, Baseline and in columns 5 and 6, Punishment. Individual controls include gender, age, number of semesters studied, one dummy per field of study, and number of errors per control question answered. Heteroscedasticity robust standard errors clustered on the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

We discuss this finding in light of two different possible explanations – the endogenous selection mechanism for the third party and the average amount of money sent – in the discussion section of this paper.

Do bribable institutions decrease trust? We first compare the bribe treatment with the punishment treatment. We find a negative but insignificant effect (-0.032 ; $p = 0.2$) of the treatment dummy for the bribe treatment in a regression without controlling for participant characteristics (Table 1, column 3). Including individual controls (Table 1, column 4), the effect increases in magnitude and is significant at the 5%-level (-0.069 ; $p = 0.013$). This provides a mixed result and could indicate a lack of power for our study, or a large degree of noise, that we capture with our individual controls. Our results indicate that in the bribe treatment, average trust decreases by 3.2 pp. (7.5 including controls) compared to the punishment treatment, consistent with the graphical analysis conducted above. Overall, we only find suggestive evidence for our **Hypothesis 2a** as trust in the bribe treatment is lower than in the punishment treatment. However, this difference is only significant when controlling for individual characteristics. Finally, we examine the difference between the institutional setup in the baseline and bribe treatments. Trust is slightly higher in the baseline treatment compared to the bribe scenario (-0.02 ; $p = 0.42$) but the difference is insignificant. Including individual con-

trols, the treatment difference slightly decreases while remaining statistically insignificant (-0.01 ; $p = 0.75$).

Alternative measures of trust. As seen in the graphical analysis, the expected return for sending a small share of the endowment (4 points) is very similar across all treatments. Following Sapienza et al. (2013), we conjecture that these return expectations might not capture the core of trust. We, therefore, repeat our analysis and exclude the return expectations for the lowest sending amount from the analysis. This slightly changes our results, as signs and magnitudes remain largely similar (see Tables A3 in Appendix A.1), but standard errors decrease. This increases the statistical significance of the estimated treatment difference between the punishment and the bribe treatment. The effect is now slightly larger and marginally significant (-0.042 ; $p = 0.082$). If we instead exclude the two (4,8) or three (4,8,12) lowest expectations, the results are similar in magnitude and significance to only excluding the lowest amount of 4.

A second measure for trust, used in many experimental investigations, is the amount sent by the trustor. While the measure proposed by Sapienza et al. (2013) captures a belief-based notion of trust – that, in our opinion, is a less confounded and hence cleaner measure of trust – we briefly discuss the results for the amount sent by trustors. Table A1 and A2 in Appendix A.1 report summary statistics of the sent amount and pairwise t-test between treatments. Importantly, we find no significant difference in the amount sent between all pairwise comparisons of our treatments using t-test (all $p > 0.25$) or Mann-Whitney-U tests (all $p > 0.024$). This indicates that the bribe channel affects trustors' *beliefs* about trustees' trustworthiness but not their *actions*. We discuss potential explanations for this finding in the discussion section.

Overall, we find suggestive evidence that trust – as measured by the return expectations – in the bribe treatment is lower compared to the punishment treatment. We do not find any evidence for a difference in trust between the baseline and the punishment treatment.

4.1.2 Trustworthiness. Analogous to the analysis of trust, Table 2 presents the results of the regression analysis for treatment effects on trustworthiness. Changing the dependent variable to the fraction of money returned for each possible sending amount, we utilize the same estimation strategy described above.

Does a punishment institution increase trustworthiness? The effect of the punishment treatment on trustworthiness compared to the baseline treatment (Table 2, column 1) is small and insignificant (-0.009 ; $p = 0.74$). We thus conclude that introducing punishment into a trust game did not significantly affect the trustees' trustworthiness. We, therefore, do not find support for **Hypothesis 1b**. In the discussion section, we discuss this result analogous to the arguments for the effects on trust.

Table 2. Treatment effects on trustworthiness

	Baseline vs. Punish		Bribe vs. Punish		Bribe vs. Baseline	
	(1)	(2)	(3)	(4)	(5)	(6)
Punishment	-0.009 (0.027)	-0.008 (0.031)				
Bribe			-0.060** (0.028)	-0.066** (0.030)	-0.069** (0.027)	-0.069** (0.030)
Amount Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Random Effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls		Yes		Yes		Yes
N	495	495	510	510	495	495
Unique N	99	99	102	102	99	99

Notes: Estimates are from a random effects model. The dependent variable is the fraction of money returned by the trustee. Punishment and Bribe are dummy variables indicating treatment status. In column 1 and 2 the omitted sample is Bribe, in 3 and 4, Baseline and in columns 5 and 6, Punish. Individual controls include gender, age, number of semesters studied, field of study, and understanding of the game. Heteroscedasticity robust standard errors clustered on the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Do bribable institutions decrease trustworthiness? Comparing the punishment treatment to the bribe treatment (Table 2, column 3), trustworthiness is significantly lower in the bribe treatment (-0.060 ; $p = 0.034$). Including individual controls, the effect is of similar magnitude and significance (-0.066 ; $p = 0.026$). Thus, participants in the bribe treatment are significantly less trustworthy than those in a punishment-only institution. This result is in line with **Hypothesis 2b** and the suggestive evidence on trust. When being able to send money to the punisher, participants reduce their trustworthiness and return on average 6 pp. less to the trustor.

Finally, trustworthiness in the bribe treatment is significantly lower than in the baseline treatment (-0.069 ; $p = 0.012$). Controlling for individual participant characteristics does not change the estimated effect or its statistical significance in any meaningful way.

Similar to the main analysis on trust, excluding responses to the sent amount of 4 does not significantly affect our measure of trustworthiness (see Table A4 in Appendix A.1). We thus conclude that trustworthiness seems to be reduced in a setting with a bribe channel but not in a setting with a punishment channel.

4.2 Bribes and Punishments

We now briefly discuss the use of punishment. In the punishment treatment, 30% of the strategy responses by the punishers entail a punishment to the trustee. The share of punishment points sent does not differ significantly between all pairwise comparisons of the

strategy responses (pairwise Mann-Whitney-U tests: all $p > 0.12$). In the bribe treatment, the share of strategy responses that contain a punishment to the trustee is similarly high (29%) and does not differ significantly between all pairwise comparisons of the strategy responses. Regarding the use of bribes, 71% of all trustee response strategies contain a bribe. The average share sent as a bribe is 12% of the received amount by the trustees across all strategy responses. This share does not differ significantly between all pairwise comparisons of the strategy responses to the received amounts (pairwise Mann-Whitney-U tests: all $p > 0.17$).

To investigate how punishments, bribes, and trustworthiness relate to each other, Table 3 presents regression estimates, where the dependent variable is the number of punishment points sent by a punisher and the independent variable is the return strategy of the trustee. In the punishment treatment (column 1), punishers send back slightly (0.036; $p = 0.23$) more punishment points for each point sent back by the trustee. Controlling for the punisher's individual characteristics (column 2) does not significantly affect the results. In the Bribe treatment, we additionally include the amount received as a bribe as an independent variable. The results show that lower punishments are associated with larger bribes (-0.182 ; $p = 0.033$) and a larger return to the trustee (-0.163 ; $p = 0.004$). We, therefore, conclude that punishment seems to be affected mainly by the size of bribes sent to the punisher and the amount sent back to the trustor. The results are robust to the inclusion of individual controls (columns 2 and 4).

4.3 Heterogeneity of Treatment Effects

Our results reveal that the possibility of side payments seems to decrease trust. To better understand how this effect varies with individual characteristics, we look at the interaction of individual characteristics with the treatment. The following section reports the results of an analysis of treatment heterogeneity concerning individual attitudes toward strong institutions and general institutional trust. The two measures discussed in this section are Right-Wing Authoritarianism (RWA) and general institutional trust. A factor analysis on all items contained in the two scales confirms that they describe distinct concepts, as proposed by the previous literature.¹⁵ We thus analyze treatment heterogeneity concerning the two scales separately.

15 Velicer's MAP test suggests retaining two factors, and two factors have eigenvalues > 1 . The results with orthogonal varimax and oblique promax rotated factor loadings provide similar results to the ones based on the standardized mean scores presented in the following section. While the orthogonal varimax results are slightly more conservative than the mean results due to the assumed orthogonality of the factors, the results based on the factor loadings after promax rotation provide smaller standard errors and stronger significance.

Table 3. Determinants of punishment

	Punishment		Bribe	
	(1)	(2)	(3)	(4)
Amount sent back	-0.036 (0.030)	-0.042 (0.031)	-0.163*** (0.056)	-0.200*** (0.060)
Received as bribe			-0.182** (0.085)	-0.191** (0.075)
Amount Dummies	Yes	Yes	Yes	Yes
Random Effects	Yes	Yes	Yes	Yes
Individual Controls		Yes		Yes
Num.Obs.	255	255	255	255

Notes: Estimates are from a random effects model. The dependent variable is the punishment points sent by the punisher in reaction to the return strategy of the trustee. "Amount sent back" captures the return strategy of the trustee, and "Received as bribe" how many points the trustee has sent to the punisher. The sample in column 1 and 2 is the Punishment treatment, and in 3 and 4 the Bribe treatment. Individual controls include gender, age, number of semesters studied, one dummy per field of study, and number of errors per control question answered. Heteroscedasticity robust standard errors clustered on the individual level.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.3.1 Right-Wing Authoritarianism (RWA). People’s attitudes towards punishment institutions relate to how they perceive authority within society (c.f. Haidt, 2012). A well-established measure for people’s tendencies toward strong authorities is the concept of Right-Wing Authoritarianism (RWA) (Altemeyer, 1981, 1996). The measure has been analyzed in many different contexts and usually finds that people scoring high on the RWA scale endorse following authorities and want society to be regulated by them (i.e., a strong punisher). However, contextual factors, such as the authority’s perceived legitimacy, mediate this effect’s strength (Stenner & Haidt, 2018). Relating the concept of authoritarianism to our treatment, we hypothesized that individuals with stronger authoritarian attitudes prefer stronger institutions and trust more in the presence of such institutions.¹⁶

We present the regression results of our main specification with an additional interaction term between the treatment dummies with the standardized RWA scores in Table 4.

We find that the treatment effect of introducing punishment on trust does not vary significantly with the RWA score (-0.035 ; $p = 0.25$) (Table 4, column 1) against our expectation. Controlling for individual characteristics, the parameter increases slightly but remains insignificant (-0.056 ; $p = 0.13$). Comparing the effects between the bribe and the punishment treatment (Table 4, column 3), we find a positive interaction between the treatment effect and the level of authoritarianism (0.050 ; $p = 0.06$), where the main effect of the bribe channel at the mean level of authoritarianism remains similar in magnitude and significance (-0.036 ; $p = 0.14$). Controlling for individual characteristics, both the interaction effect (0.058 ; $p = 0.028$) and the main treatment effect (-0.074 ; $p < 0.001$) increase in size and significance. Looking at predictive marginal effects across the RWA scale, we find that the interaction effect is driven mainly by individuals with low values of authoritarianism, while we can not distinguish between treatment responses for individuals with RWA values at the mean or higher.¹⁷

We interpret this finding as descriptive evidence for **Hypothesis 3**, namely that individuals with a strong aversion to authorities reduce their trust towards others when observed and potentially punished by corruptible institutions. In contrast, individuals scoring high on the authoritarianism scale seem to not be concerned about the possibility of corruption. This might be caused by the fact that individuals with moderate to high levels of authoritarianism do not perceive an institutional effect on trustee behavior. However, individuals with low levels of authoritarianism doubt the punishing institution in light of possible bribery.

16 While RWA can further be divided into three distinct sub-scales (authoritarian aggression, authoritarian submission, and conventionalism), we present results to the overall individual RWA score.

17 See Figure A1 and A2 in Appendix A.1 for the marginal effect graphs.

Table 4. Heterogenous treatment effects: RWA

	Baseline vs. Punish		Bribe vs. Punish		Bribe vs. Baseline	
	(1)	(2)	(3)	(4)	(5)	(6)
RWA	0.005 (0.022)	0.039 (0.027)	-0.030 (0.021)	-0.029 (0.023)	0.005 (0.022)	0.023 (0.024)
Punishment	0.016 (0.027)	0.024 (0.026)				
Punishment \times RWA	-0.035 (0.030)	-0.056 (0.037)				
Bribe			-0.036 (0.025)	-0.074*** (0.026)	-0.021 (0.026)	-0.009 (0.032)
Bribe \times RWA			0.050* (0.027)	0.058** (0.026)	0.015 (0.027)	-0.008 (0.029)
Amount Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Random Effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls		Yes		Yes		Yes
N	495	495	510	510	495	495
Unique N	99	99	102	102	99	99

Notes: Estimates are from a random effects model. The dependent variable is the trustee's expectation about the fraction of money returned. Punishment and Bribe are dummy variables indicating treatment status. In column 1 and 2 the omitted sample is Bribe, in 3 and 4 Baseline and in column 5 and 6 Punish. Columns 2, 4 and 6 additionally control for the same individual characteristics as the main regressions: gender, age, number of semesters studied, one dummy per field of study, and number of errors per control question answered. RWA is the standardized value of the response to the Right Wing Authoritarian Value scale. Heteroscedasticity robust standard errors clustered on the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

A potential caveat might be that the students in our sample are not authoritarian compared to the general population, thus limiting the external validity of our analysis. To investigate this, we compare the results of our experimental student sample to two representative German samples collected in 2011 (Beierlein et al., 2014). Our sample's mean RWA score is lower than in the representative German samples (2.14 vs. 2.52/2.53). However, the range (min=1; max=4.4 vs. min=1/1; max= 4.56/4.11) and the standard deviation are of similar magnitude (0.63 vs. 0.63/0.69).

4.3.2 Institutional Trust. Similar to views on authoritarianism, individuals' general trust in institutions might affect how participants react to being subjected to different institutional settings. In the previous analysis, we investigated individuals' preference for a strong punishing authority in general. We now turn to the relationship between the treatment effects and trust in institutions. While a person may not score high on the authoritarianism scale endorsing a powerful institution or leader, they may nevertheless trust institutions in society to run well, e.g., due to trust in institutions and division of powers. We thus compare the individual treatment effects across the spectrum of institutional trust.

Table 5 reports the results of treatment interactions with institutional trust. Comparing the baseline to the punishment treatment (Table 5, column 1), we do not find any interaction effect between the treatment and the level of institutional trust without (-0.006 ; $p = 0.84$) or with individual controls (-0.035 ; $p = 0.28$). Similarly, we do detect a positive but insignificant interaction effect (0.037 ; $p = 0.063$) when comparing the bribe to the punishment setting (Table 5, column 3). The average treatment effect at the mean level of institutional trust is negative, slightly smaller compared to the main specification, and insignificant (-0.019 ; $p = 0.51$). When we include individual controls, we again see an increase in significance and size for both the main treatment effect (-0.059 ; $p = 0.068$) and the interaction effect (0.064 ; $p = 0.045$). Comparing the predictive margins, we see that the interaction effect is driven by individuals with below-average levels of institutional trust, while we cannot distinguish treatment response at mean levels or above.¹⁸ The positive coefficient on the interaction term indicates that a person low in institutional trust decreases their trust in the bribe setting relative to the punishment setting. We interpret this as suggestive evidence for **Hypothesis 4**.

Overall, we conclude that the knowledge about the existence of the potential to bribe a punishment institution might invoke heterogeneous reactions in individuals. Specifically, individuals with strong aversions against authoritarian institutions and low trust in political institutions seem to react with a substantial reduction in interpersonal trust. General trust in institutions and attitudes towards them thus might be key predictors of

¹⁸ See Figures A3 and A4 in Appendix A.1 for the marginal effects graph.

Table 5. Heterogenous treatment effects: institutional trust

	Baseline vs. Punish		Bribe vs. Punish		Bribe vs. Baseline	
	(1)	(2)	(3)	(4)	(5)	(6)
Inst. Trust	0.001 (0.026)	0.040 (0.027)	-0.005 (0.015)	-0.004 (0.018)	0.001 (0.026)	0.004 (0.026)
Punishment	0.008 (0.030)	0.030 (0.030)				
Punishment × Inst. Trust	-0.006 (0.030)	-0.035 (0.033)				
Bribe			-0.019 (0.028)	-0.059* (0.032)	-0.011 (0.027)	-0.005 (0.031)
Bribe × Inst. Trust			0.030 (0.022)	0.064** (0.032)	0.024 (0.030)	0.055* (0.031)
Amount Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Random Effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls		Yes		Yes		Yes
N	435	435	420	420	465	465
Unique N	87	87	84	84	93	93

Note: Estimates are from a random effects model. The dependent variable is the trustee's expectation about the fraction of money returned. Punishment and Bribe are dummy variables indicating treatment status. In column 1 and 2 the omitted sample is Bribe, in 3 and 4 Baseline and in column 5 and 6 Punish. Columns 2, 4, and 6 additionally control for the same individual characteristics as the main regressions: gender, age, number of semesters studied, one dummy per field of study, and number of errors per control question answered. Inst. Trust is the standardized value of the response to the Institutional Trust questions. Heteroscedasticity robust standard errors clustered on the individual level.

Due to a technical incident at the end of one session, 18 participants did not fill out the institutional trust questionnaire. We do not find any systematic difference with respect to other characteristics comparing the participants with the missing data to the rest of the sample.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

how interpersonal trust is affected under corrupt institutions. This sheds further light on the detrimental effect of corruption and its interaction with attitudes toward authorities.

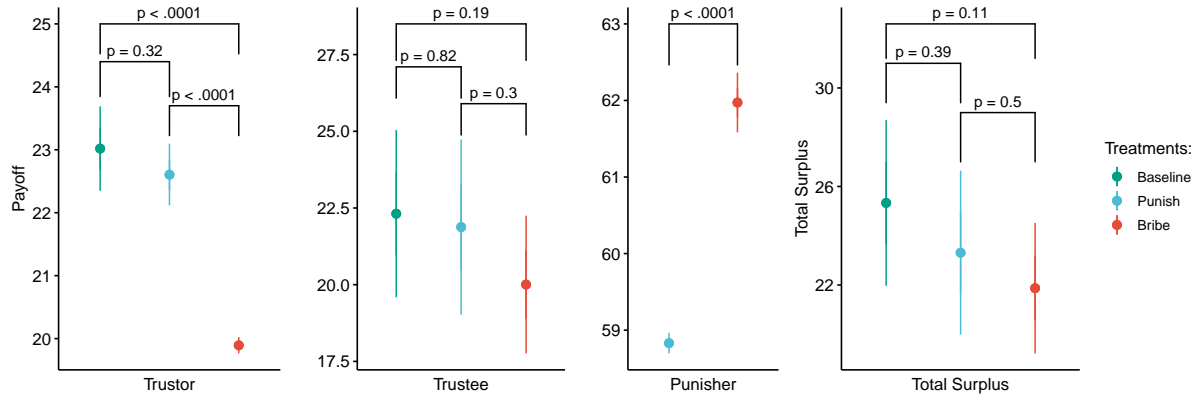
4.4 Payoffs and Surplus

An important aspect missing from the analysis so far are the material consequences of the treatments on both individual payoffs and overall welfare. Thus, we present an analysis of individual payoffs and the surplus generated by trusting in the different treatments to determine their monetary consequences. Having collected data from all responders (trustee and punisher) via the strategy method and from the trustor via single choice, we match each sending decision by one player with all response strategies of all other players in the same treatment. This procedure generates a robust and interesting data set, as it matches the empirical distribution of trustors' actions to the distribution of response strategies and uses all available data. We believe this to be an improvement over analyzing the payoffs of the (randomly) generated matched player pairs in the experiment.

We define the single sending decision of a trustor in Treatment t to be $s_{t,i}^{\text{Trustor}}$ where $s \in \{0, 4, \dots, 20\}$ and $t \in \{\text{baseline}, \text{punish}, \text{bribe}\}$. The (pure) response strategy by one trustee is given by $r_{t,i}^{\text{Trustee}} = (r_{s=4}, r_{s=8}, \dots, r_{s=20})$. We define the set of all n trustees response strategies in the same treatment as $\mathbf{r}_t^{\text{Trustee}} = (r_{t,i=1}, \dots, r_{t,i=n})$. Thus, the set of all response strategies for all players in the role of trustee except i in treatment t is defined by $\mathbf{r}_{t,-i}^{\text{Trustee}}$. Since the punishment (and bribe) strategy of the punishers in the punishment and the bribe treatments is always a response to the strategy by one specific trustee, the response strategy $r_{t,i}^{\text{Trustee \& Punisher}}$ in these two treatments is defined to contain the strategy by both the trustee and the partnered punisher. Finally, each sending decision $s_{t,i}^{\text{Trustor}}$ is matched with each response strategy in $\mathbf{r}_{t,-i}^{\text{Trustee (\& Punisher)}}$ and the payoffs for each player are calculated. The resulting mean payoff and the mean generated surplus (i.e., the sum of all participants' payoffs minus the initial endowment(s)) with their 95 percent CI are presented in Figure 3. Brackets signify the pairwise comparisons and the associated p-values of two-sided t-tests of a difference in means.¹⁹

The payoff of the trustor is the highest in the baseline treatment (23.0 points), only slightly lower in the punishment treatment (22.6 points), and the lowest in the bribe treatment (19.9 points). Similarly, the trustee's payoff is the highest in the baseline treatment (22.3 points), followed by the punishment treatment (21.88 points) and the bribe treatment (20.01 points). The difference between the trustor payoff in the baseline and the punishment treatment is small (0.41 points) and insignificant, when using a variety

¹⁹ For a graph showing the resulting p-values for the same comparisons using a two-sided Mann-Whitney-U Test see Figure A5 in Appendix A.1.



Notes: The graph depicts the mean end-of-game payoff for each participant in each role in each treatment calculated by matching strategy responses to actual trustor decisions for all participants in the three left panels. The right panel depicts the total surplus generated. Whiskers represent the 95-percent confidence interval around the mean. Brackets and p-values represent the pairwise comparisons of means based on a two-sided t-test.

Figure 3. Payoffs and surplus

of parametric and non-parametric tests.²⁰ The payoff difference between the trustee in the two treatments is small in magnitude and insignificant,²¹ as trustees in the baseline treatment are on average only slightly better off (0.43 points). Thus, the introduction of the punisher has not increased the average payoff of the trustor or the trustee. This is consistent with our results on trust and trustworthiness, as we failed to detect effects on either in our principal analysis.

Comparing the differences between the punishment and the bribe treatment, we observe a significantly²² lower trustor payoff in the bribe setting (2.72 points). The trustee's payoff difference is more negligible (1.87 points) and insignificant.²³ Finally, we can compare the payoff of the punishers between the two treatments. The punishers in the bribe treatment have a significantly higher average payoff (by 3.14 points). Overall, the trustors are worse off with the existence of a bribe channel compared to the punishment-only institution, the punishers have gained, and the payoff of the trustees has remained untouched. Thus, the burden of introducing the bribe channel is carried mainly by the

20 ($W = 1464.5$; $p = 0.093$) using a Mann-Whitney-U Test; ($Z = 1.67$; $p = 0.095$) using an approximative (100000 draws) van der Waerden Test; ($t = 1.00$; $p = 0.32$) using a Welch t-test; ($Z = 1.01$; $p = 0.31$) using an approximative (100000 draws) Two-Sample Fisher-Pitman Permutation Test

21 ($W = 1385$; $p = 0.26$) using a Mann-Whitney-U Test; ($Z = 1.5$; $p = 0.13$) using an approximative (100000 draws) van der Waerden Test; ($t = 0.22$; $p = 0.82$) using a Welch t-test; ($Z = 0.22$; $p = 0.82$) using an approximative (100000 draws) Two-Sample Fisher-Pitman Permutation Test

22 ($W = 147$; $p \ll 0.001$) using a Mann-Whitney-U Test; ($Z = -7.35$, $p \ll 0.001$) using an approximative (100000 draws) van der Waerden Test; ($t = -10.75$; $p \ll 0.001$) using a Welch t-test; ($Z = -7.36$; $p \ll 0.001$) using an approximative (100000 draws) Two-Sample Fisher-Pitman Permutation Test

23 ($W = 1146$; $p = 0.30$) using a Mann-Whitney-U Test; ($Z = -1.23$; $p = 0.22$) using an approximative (100000 draws) van der Waerden Test; ($t = -1.04$; $p = 0.30$) using a Welch t-test; ($Z = -1.04$; $p = 0.30$) using an approximative (100000 draws) Two-Sample Fisher-Pitman Permutation Test

trustors, consistent with our analysis of the effects of the different institutional setups on both trust and trustworthiness.

The difference in trustor payoff between the bribe and the baseline treatment is relatively large (3.12 points) and statistically significant.²⁴ The difference in trustee payoff, however, is slightly smaller (2.31 points) and insignificant.²⁵ Contrasting the punishment and bribe free environment to the one where punishment and bribing are possible, the payoff of the trustor is significantly lower in the bribe setup, while the payoffs of the trustee seem to not change to a statistically measurable degree.

Finally, the surplus generated through player interaction can be calculated to compare the overall welfare effects of the different institutions. The generated surplus in the baseline treatment is the highest (25.33 points), followed by the punishment treatment (23.31 points) and the bribe treatment (21.87 points). Despite the seemingly large numerical effect, we do not find any statistically significant differences in the surplus generated between any pair of our treatments when testing two-sided.²⁶ We thus conclude that the existence of punishment has not changed the individual payoffs or aggregate surplus compared to a punishment-free environment. However, the existence of the bribe channel has decreased the payoff of trustees compared to the other two environments, while the punishers have gained compared to the punishment-only setting. While the estimates of differences in overall surplus are insignificant, they seem to indicate that the existence of the bribe channel might induce welfare losses due to reduced trust and trustworthiness.

5 Discussion

In this section, we first discuss a potential issue in the design of our study, namely that participants might not have interpreted the “bribe”-channel as a bribe channel. Secondly, we discuss potential reasons for the differences in participants’ trust when measured by their beliefs or when measured as the amount sent. Finally, we discuss the differences

24 ($W = 172$; $p \ll 0.001$) using a Mann-Whitney-U Test; ($Z = -6.86$, $p \ll 0.001$) using an approximative (100000 draws) van der Waerden Test; ($t = -9.21$; $p \ll 0.001$) using a Welch t-test and ($Z = -6.86$; $p \ll 0.001$) using an approximative (100000 draws) Two-Sample Fisher-Pitman Permutation Test

25 ($W = 1097$; $p = 0.38$) using a Mann-Whitney-U Test; ($Z = -1.32$, $p = 0.19$) using an approximative (100000 draws) van der Waerden Test; ($t = -1.31$; $p = 0.19$) using a Welch t-test; ($Z = -1.32$; $p = 0.19$) using an approximative (100000 draws) Two-Sample Fisher-Pitman Permutation Test

26 The largest difference in surplus is between the Baseline and the Bribe treatment. ($W = 1223$; $p = 0.48$) using a Mann-Whitney-U Test; ($Z = -1.62$, $p = 0.11$) using an approximative (100000 draws) van der Waerden Test; ($t = -1.63$; $p = 0.11$) using a Welch t-test and ($Z = -1.62$; $p = 0.10$) using an approximative (100000 draws) Two-Sample Fisher-Pitman Permutation Test

between our study and the study by Charness et al. (2008) and how these differences might explain the diverging results.

A first criticism of our design might be that participants did not interpret the transfer from a trustee to the punisher as a bribe but, for example, as a gift due to the neutral framing of the experiment and the relatively low prevalence of corruption in Germany. Our findings of lower trust and trustworthiness could therefore result from participants' altruistic preferences to share some of their endowment with the other players. We discuss three pieces of evidence that support our interpretation.

Firstly, note that we set the punisher's endowment to the maximum amount of points a trustee could have in case the trustor had sent everything. Therefore, trustees should not be motivated to send something to the punisher due to advantageous inequality aversion. Secondly, we test more directly for a relationship between altruism²⁷ and money sent to the third party. We estimate random effects panel regressions on the sub-sample of the participants in the bribe treatment (see Table A5 in Appendix A.1), controlling for scale effects of the strategy method by including amount dummies. Using the fraction sent back (i.e., trustworthiness) as the dependent variable and the altruism survey measure (Falk et al., 2022) as the independent variable, the coefficient for altruism is positive but small and insignificant in explaining trustworthiness (0.019; $p = 0.38$). Using the fraction bribed as a dependent variable, the coefficient for altruism is again small and insignificant (0.019; $p = 0.26$). Including the amount sent back by the trustee as an additional independent variable does not change the robustness of that result. We thus conclude that altruistic preferences are not associated with either the bribing or the sending back behavior in the bribe treatment.

Finally, we checked more explicitly for participants' interpretation regarding the amount sent to the punisher. At the end of the experiment, we asked all participants in the bribe treatment to rate how important the amount sent by the trustee to (a) the trustor and (b) the punisher was in determining their punishment decision as punishers. They answered on a scale of 1 ("not important at all") to 7 ("very important"). If interpreted as a bribe, sending something back to the trustor is fundamentally differently motivated (e.g., reciprocity) than something sent to the punisher (e.g., bribing). Therefore, we expect no or only a small correlation between the importance rating of sending back to the trustor and to the punisher. Indeed, we only find a small and insignificant correlation (Spearman's $\rho = 0.22$; $p = 0.12$). Thus sending back to the trustor and sending to the punisher seem to be perceived as different channels for determining punishment. Finally, we asked all participants in the bribe treatment about their motivation for sending some-

27 To test the existence of treatment effects on altruism, we conduct two-sample t-tests for each pair of sub-samples and find that in the bribe treatment, participants reported lower measures of altruism compared to the other treatments (c.f. Table A2 in Appendix A.1). We thus refrain from using the measure of altruism to compare heterogeneous treatment effects.

thing to the punisher in an open-ended question. Of all subjects who sent a bribe, 52% (24/46) answered that they sent a bribe to either send less to the trustor or to prevent punishment overall.²⁸ Since subjects might have moral costs associated with admitting to bribery directly, we believe this number to be underestimated and the true percentage of bribes motivated by sending back less to be higher. Finally, individuals who act out of altruism towards both the punisher and the trustor, or participants who fear the punisher more strongly in general, would send more to both the punisher and the trustor leading to a positive correlation between these amounts. However, when regressing the fraction sent to the punisher on the fraction sent to the trustor, we get a negative and highly significant coefficient (-0.17 ; $p = 0.02$).²⁹ Thus individuals who send more to the punisher send back significantly less to the trustor. This result supports the argument that people sent more to the punisher to get away with sending less to the trustor. Overall, we conclude that participants' altruistic preferences fail to predict bribing behavior and that participants seem to have interpreted the option for side payments as bribes.

A second – at first sight – puzzling finding is that we find an effect of the introduction of the bribe channel in the belief-based measure of trust but not the sending decisions of participants. We suggest two potential explanations for this finding: Firstly, the action of sending something as a trustor does not only capture trust but might also be motivated by inequality aversion (Cox, 2004), altruism (Ashraf et al., 2006), or can be affected by uncertainty attitudes (Li et al., 2019), or emotional factors (Dunning et al., 2014; Dunning et al., 2012). It might be that those factors somehow determine a significant share of the amount sent in our experimental design of the trust game. Therefore, we would be less likely to find a significant effect of the bribe channel in case its main effect is through the belief-based component of trust.

A second reason might be that participants do not always act according to the best strategy implied by their own beliefs in strategic settings (Costa-Gomes & Weizsäcker, 2008; Nyarko & Schotter, 2002). To further understand whether there is a causal effect of beliefs on actions in the trust game, Costa-Gomes et al. (2014) experimentally manipulate trustors' beliefs and can thus causally estimate their effect on actions. They show that sending decisions are causally influenced by beliefs but that this effect is not perfect. Relating their finding to our result, we do not find a strong correlation between the optimal action derived from participants' beliefs and their actual sending behavior (Pearson's $r = 0.05$, $p = 0.58$).³⁰

28 16/46 participants directly answered “Yes” to the question “I sent points to the punisher to send fewer points to the trustor, without getting points deducted by the punisher.” In addition, 8/30 answered in a free comment field that they sent points to the punisher to not get points deducted in general.

29 (Column 3 and 4 in Table A5 in Appendix A.1)

30 We calculate trustors' best action as the maximum end-of-game payoffs for each possible sending amount based on their beliefs about the returned amounts. For this analysis, we exclude 37 participants with multiple actions yielding the same maximum payoff.

We now briefly discuss the potential reasons why we might not detect an effect of punishment on trust in contrast to Charness et al. (2008) (further CCJ).

Firstly, the experimental protocol utilized by CCJ implements an *endogenous selection* mechanism of direct voting for the third-party punisher at the beginning of an experimental session, where participants see each other but cast their vote in private. In such a setup, participants might form (or already have) beliefs about a shared norm within certain group members and select a punisher they feel might represent their own norm. Additionally, the endogenous selection mechanism and non-anonymity might induce participants to build trust in the punishment institution, as a voted-for institution might resemble a more trustworthy institution than an externally imposed one. Finally, having a choice over the institutional setting in which one interacts gives rise to the so-called “democracy effect”.³¹ While participants in CCJ do not have a choice of the institutional framework per se, the process of voting for a punisher might invoke similar effects by relying on participants’ procedural preferences (e.g., Frey et al., 2004; Frey & Stutzer, 2005). In contrast, in our setting, we explicitly exclude this channel to impact our results. To answer our research question, we want to measure the impact of punishment by an exogenously imposed outside party and the effect of corruptibility of this party on trust. This resembles real-world settings where punishing institutions (like the police or judges) are often chosen exogenously, and anonymity might play a crucial role in people’s beliefs about other persons’ propensity to be corrupt.

Secondly, the average amount of money sent in our baseline setting is already high, as trustors sent 63.5% of their initial endowment on average, compared to 37.3% in CCJ. The share of endowment that is sent in our study in the punishment setting (67%) is close to the share in CCJ (60.2%). When testing for a treatment effect on the amount sent using the same methodology as CCJ, we find no significant effect. When looking at trustworthiness, participants in the baseline treatment specified a return strategy that returned 38.6% of the received (tripled) amount on average compared to 37% in CCJ. In the punishment treatment, participants specified a return strategy of 37.7% in our setting and 46% in CCJ. Conditioning these numbers on the empirical distribution of trustor behavior does not change the results.³² A potential explanation for the higher baseline trust in our setting could be the different subject pool (German vs. American students). An alternative explanation might be due to our use of role uncertainty, while CCJ use a fixed role assignment. While (to our knowledge) no systematic analysis of the effect of role uncertainty in the trust game exists, mixed evidence in the context

31 See the recent overview paper by Dannenberg and Gallier (2020) for a detailed analysis of the impact of exogenously and endogenously imposed institutions on behavior.

32 See Section 4.4 for the methodology to calculate player payoff and surplus in the presence of the double strategy method elicitation. Using this methodology, we calculate mean return rates of 40% in the punishment treatment and 41% in the baseline treatment.

of dictator games suggests that role uncertainty might increase generosity (Iriberry and Rey-Biel (2011); Walkowitz (2021) find positive effects while Engelmann and Strobel (2004) find no effects). Similarly, it is thus conceivable that role uncertainty might have increased trust in our baseline setting.

We thus conclude that in addition to the endogenous selection mechanism employed by CCJ, baseline trust is higher in our setting than in CCJ's study. Both of these factors may explain the different findings.

6 Conclusion

Trust and trustworthiness are major preconditions for the successful functioning of markets and society in general. Because trust is malleable, depends on the context, and can be influenced by institutions, many institutions in society are constructed to take action against norm-violators through punishment, such as law enforcement (Herreros, 2023). Previous studies have shown that third-party punishment may increase interpersonal trust (Charness et al., 2008; Fiedler & Haruvy, 2017). However, at the same time, institutions meant to increase trust in society may also be targeted by individuals trying to circumvent law enforcement, thereby destroying their potential for enhancing trust. One of those mechanisms is bribing.

In this study, we analyze the causal effect of the corruptibility of a punishing institution in a trust game setting. In contrast to previous studies, we do not find that adding a punisher to the trust game increases trust or trustworthiness. Nevertheless, when the punisher can receive a transfer from the trustee, we find suggestive evidence that trust is decreased and strong evidence for a decrease in trustworthiness. The trustor bears the cost of this corruptibility while the punisher benefits. Thus, the mere opportunity to bribe a third-party punisher affects interpersonal relations that require trust between individuals interacting in a setting governed by a bribable institution. An alternative explanation for our findings is that trustees interpreted the “bribe” channel as an opportunity to altruistically share some points with another player. In this view, they send something to the punisher not to avoid punishment but simply out of altruism. While acknowledging this possibility and interpretation, we designed our experiment to exclude the motivation to share something with the punisher and provide descriptive evidence supporting our interpretation.

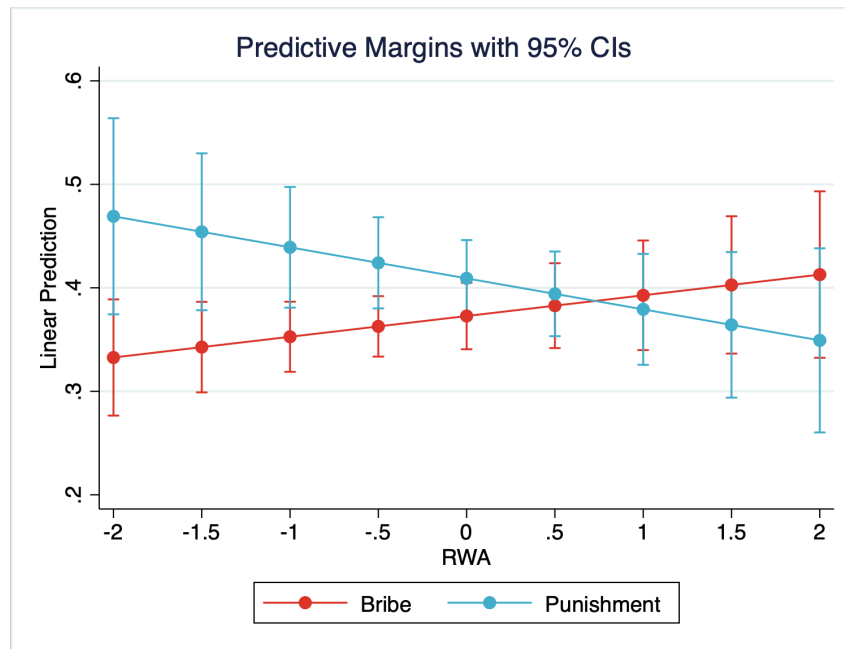
Our findings are related to the individual characteristics of participants. Specifically, we find that individuals scoring low on a measure of authoritarianism and institutional trust have significantly decreased their trust in the bribe setting compared to individuals with similar attributes in the punishment treatment. This means that the trust of people with a strong preference for authorities or a high trust in institutions is not as strongly

affected when confronted with institutional imperfections. However, as trustworthiness is reduced on average, this unresponsiveness might not be well placed.

This is an important finding, as third-party punishment has been shown in many studies to solve dilemmas of cooperation and trust (Charness et al., 2008; Charness et al., 2011; Fehr, Fischbacher, & Gächter, 2002; Fehr & Gächter, 2000; Rockenbach & Milinski, 2006). However, in these studies, third parties are often exogenous to the interaction they govern. In reality, this is not true, as the institutions meting out punishment are often addressable by the parties captured in the cooperative dilemma. Thus, the effectiveness of these institutions in building and enhancing trust might only be as good as a person's belief in the integrity of the punishing institution. An important influence on this belief might be how well the punishers themselves are controlled and can be held accountable for their actions. We show that in a setting without any oversight over or accountability of bribable punishers, their trust-inducing power might be significantly affected. In fact, we find that bribable punishers without any accountability might lead to reduced levels of trust and trustworthiness. Taking these behavioral consequences of corruption into account might be especially important when trying to estimate its costs, as a strong positive relationship between trust and a country's economic success exists (e.g., Algan & Cahuc, 2013; Knack & Keefer, 1997). This result sheds light on the importance of the effort to not only build up law enforcement institutions in developing countries but also to invest in the institutions' credibility.

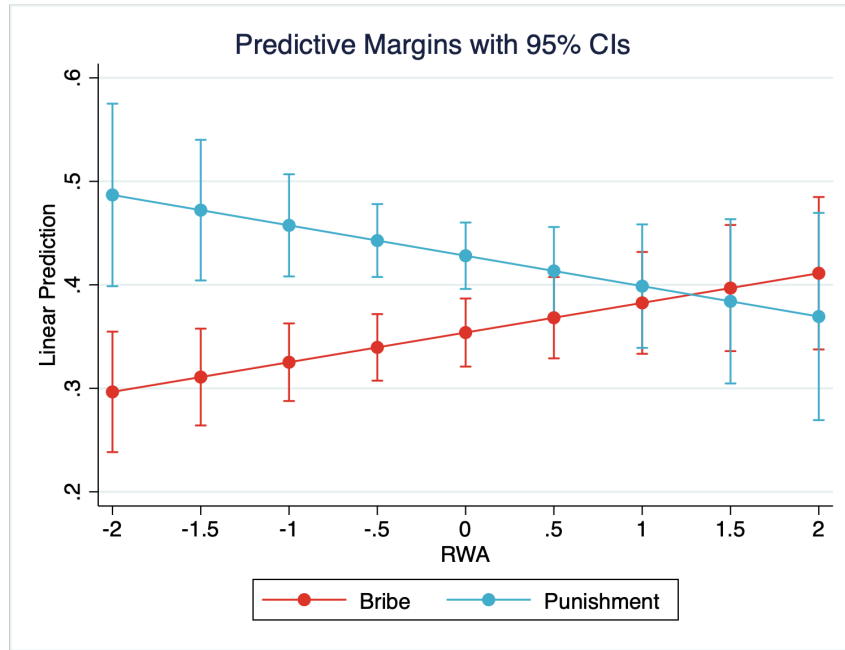
A Appendix

A.1 Additional Figures and Tables



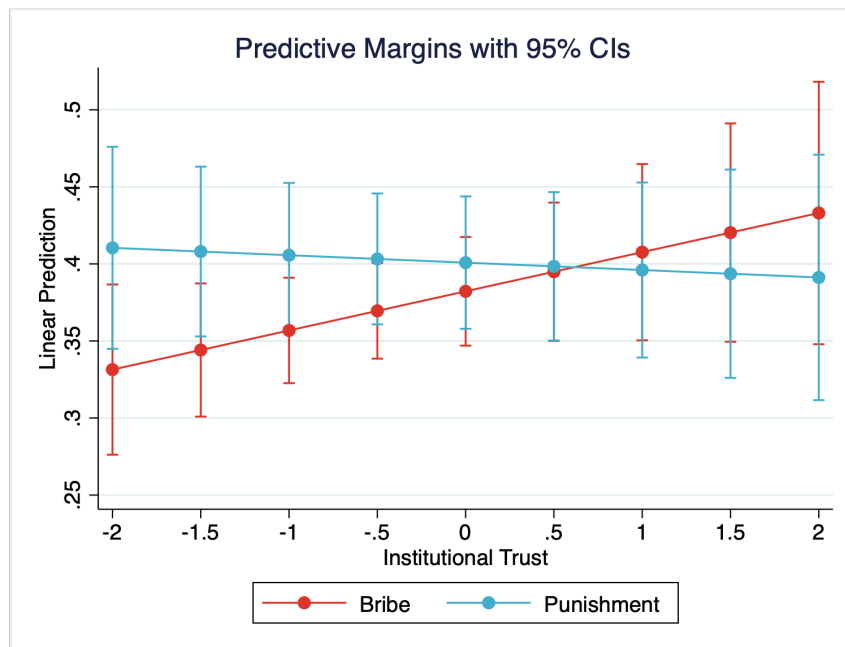
Notes: The graph depicts the predicted marginal effects of the standardized RWA score on interpersonal trust based on the random effects panel estimation used to estimate the results in Table 4. The dependent variable is the expected share returned (i.e., trust), and the independent variables are the standardized RWA score, amount dummies, and a treatment indicator. Heteroscedasticity robust standard errors are clustered on the individual level.

Figure A1. Predictive margins – RWA



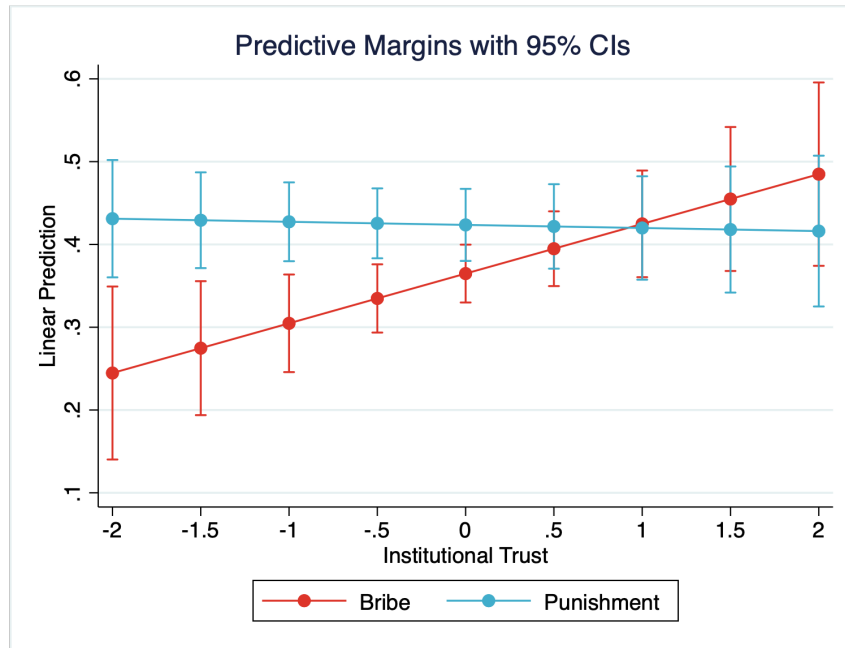
Notes: The graph depicts the predicted marginal effects of the standardized RWA score on interpersonal trust based on the random effects panel estimation used to estimate the results in Table 4. The dependent variable is the expected share returned (i.e., trust), and the independent variables are the standardized RWA score, amount dummies, a treatment indicator as well as individual controls. Heteroscedasticity robust standard errors are clustered on the individual level.

Figure A2. Predictive margins – RWA with controls



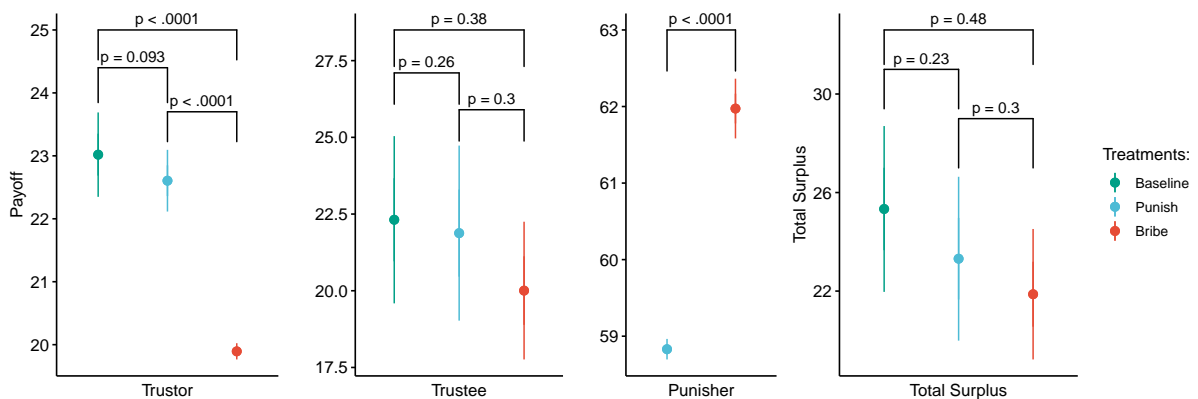
Notes: The graph depicts the predicted marginal effects of the standardized institutional trust on interpersonal trust based on the random effects panel estimation used to estimate the results in Table 5. The dependent variable is the expected share returned (i.e., trust) and the independent variables are the standardized institutional trust, amount dummies, and a treatment indicator. Heteroscedasticity robust standard errors are clustered on the individual level.

Figure A3. Predictive margins – institutional trust



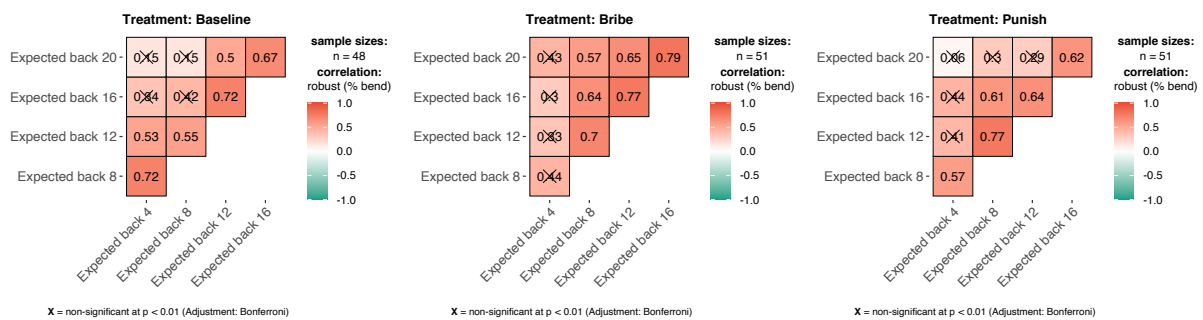
Notes: The graph depicts the predicted marginal effects of the standardized institutional trust on interpersonal trust based on the random effects panel estimation used to estimate the results in Table 5. The dependent variable is the expected share returned (i.e., trust) and the independent variables are the standardized institutional trust, amount dummies, a treatment indicator, as well as individual controls. Heteroscedasticity robust standard errors are clustered on the individual level.

Figure A4. Predictive margins – institutional trust with controls



Notes: The graph depicts the mean potential end-of-game payoff for each participant in each role in each treatment calculated by matching strategy responses to actual trustor decisions for all participants. Whiskers represent the 95-percent confidence interval around the mean. Brackets and p-values represent the pairwise comparisons of means based on a Mann-Whitney-U test.

Figure A5. Payoffs and surplus – Mann-Whitney-U p-values



Notes: The heatmap depicts robust percentage bend correlation coefficients (Wilcox, 1994) between all return expectations elicited by the strategy method. The crossed-out coefficients are insignificant at the 1%–level when adjusting for multiple comparisons using the Bonferroni method.

Figure A6. Correlation of strategy method answers

Table A1. Summary statistics: means

	All		Baseline		Punish		Bribe	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Female	0.59	0.49	0.52	0.50	0.67	0.48	0.59	0.50
Age	22.06	2.92	21.69	3.15	21.98	2.40	22.49	3.18
Study Semester	3.85	3.30	2.94	2.44	3.73	3.07	4.82	3.95
Errors in ControlQ	0.62	1.24	0.60	1.42	0.53	1.09	0.74	1.22
RWA Score	-0.00	1.00	-0.14	0.96	0.15	1.00	-0.02	1.04
Institutional Trust	0.00	1.00	0.06	0.92	-0.00	1.12	-0.07	0.99
Altruism	-0.00	0.86	0.09	0.75	0.11	0.97	-0.20	0.82
Trustor Sent Amount	13.36	5.75	12.67	5.80	13.41	5.92	13.96	5.56
Observations	150		48		51		51	

Notes: This table contains mean and standard deviations of the entire sample. Female is a dummy indicating whether a subject identified as female or not, Age the age in years, Study Semester the number of semesters studied, Errors in ControlQ is the number of errors a participant made when answering the control questions, RWA Score the standardized score on the Right Wing Authoritarianism scale, Institutional Trust the mean standardized institutional trust, Altruism the altruism score calculated according to Falk et. al (2022), and Trustor Sent Amount the amount of money sent by the trustors.

Table A2. Summary statistics: treatment differences

	Baseline vs Punish		Bribe vs Punish		Baseline vs Bribe	
	Difference	p	Difference	p	Difference	p
Female	0.15	0.14	0.08	0.42	0.07	0.51
Age	0.29	0.61	-0.51	0.36	0.80	0.21
Study Semester	0.79	0.16	-1.10	0.12	1.89***	0.01
Errors in ControlQ	-0.07	0.78	-0.21	0.37	0.14	0.61
RWA score	0.30	0.13	0.17	0.40	0.13	0.53
Institutional Trust	-0.07	0.77	0.06	0.78	-0.13	0.51
Altruism	0.03	0.87	0.31*	0.08	-0.28*	0.08
Trustor Sent Amount	0.75	0.53	-0.55	0.63	1.29	0.26
Observations	99		102		99	

Notes: This table contains differences in means and p-values of pairwise t-tests of equality of means between all pairs of treatments. Female is a dummy indicating whether a subject identified as female or not, Age the age in years, Study Semester the number of semesters studied, Errors in ControlQ is the number of errors a participant made when answering the control questions, RWA Score the standardized score on the Right Wing Authoritarianism scale, Institutional Trust the mean standardized institutional trust, Altruism the altruism score calculated according to Falk et. al (2022), and Trustor Sent Amount the amount of money sent by the trustors. * p < 0.1, ** p < 0.05, *** p < 0.01

Table A3. Treatment effects on trust: 4 excluded

	Baseline vs Punish		Bribe vs Punish		Bribe vs Baseline	
	(1)	(2)	(3)	(4)	(5)	(6)
Punishment	0.011 (0.026)	0.028 (0.026)				
Bribe			-0.042* (0.024)	-0.075*** (0.027)	-0.031 (0.024)	-0.019 (0.031)
Individual controls	No	Yes	No	Yes	No	Yes
Decision Amount	Yes	Yes	Yes	Yes	Yes	Yes
Random Effects	Yes	Yes	Yes	Yes	Yes	Yes
N	396	396	408	408	396	396
Unique N	99	99	102	102	99	99

Notes: Estimates are from a random effects model. The dependent variable is the trustee's expectation about the fraction of money returned. Punishment and Bribe are dummy variables indicating treatment status. In column 1 and 2 the omitted sample is Bribe, in 3 and 4, Baseline and in columns 5 and 6, Punishment. All responses to the sending amount of 4 are omitted. Individual controls include gender, age, number of semesters studied, one dummy per field of study, and number of errors per control question answered. Heteroscedasticity robust standard errors clustered on the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A4. Treatment effects on trustworthiness: 4 excluded

	Baseline vs Punish		Bribe vs Punish		Bribe vs Baseline	
	(1)	(2)	(3)	(4)	(5)	(6)
Punishment	-0.013 (0.025)	-0.007 (0.027)				
Bribe			-0.065** (0.027)	-0.073** (0.029)	-0.078*** (0.026)	-0.082*** (0.027)
Individual controls	No	Yes	No	Yes	No	Yes
Decision Amount	Yes	Yes	Yes	Yes	Yes	Yes
Random Effects	Yes	Yes	Yes	Yes	Yes	Yes
N	396	396	408	408	396	396
Unique N	99	99	102	102	99	99

Notes: Estimates are from a random effects model. The dependent variable is the fraction of money returned by the trustee. Punishment and Bribe are dummy variables indicating treatment status. In column 1 and 2 the omitted sample is Bribe, in 3 and 4, Baseline and in columns 5 and 6, Punish. All responses to the sending amount of 4 are omitted. Individual controls include gender, age, number of semesters studied, field of study, and understanding of the game. Heteroscedasticity robust standard errors clustered on the individual level.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A5. The impact of altruism in the bribe treatment

	Sent Back		Sent as Bribe	
	(1)	(2)	(3)	(4)
Altruism	0.019 (0.021)	0.019 (0.017)		0.022 (0.018)
Sent Back			-0.165** (0.072)	-0.168** (0.070)
Amount Dummies	Yes	Yes	Yes	Yes
Random Effects	Yes	Yes	Yes	Yes
N	204	204	204	204
Unique N	51	51	51	51

Notes: The dependent variable in the first column is the fraction of money returned by the trustee. In columns 2 – 4 it is the amount sent to the punisher by the trustee. Altruism is the Altruism score based on the two questions and associated weights reported in Falk et al. (2018). Items are standardized within this sample, before weights are applied. Sent Back is the amount sent back by the Trustee to the Trustor. Only data from the bribe treatment is used for the estimation. All responses to the sending amount of 4 are omitted. Heteroscedasticity robust standard errors clustered on the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A6. Payoff and surplus

	Control	Punish	Bribe
Trustor			
Payoff	23.02 (0.33)	22.61 (0.24)	19.89 (0.06)
Trustee			
Payoff	22.31 (1.35)	21.88 (1.42)	20.01 (1.12)
Punisher			
After punishment (w/o bribe)		58.83 (0.07)	58.12 (0.08)
Payoff		58.83 (0.07)	61.97 (0.19)
Total Surplus	25.33 (1.67)	23.31 (1.66)	21.87 (1.32)

Notes: The table presents mean payoffs calculated by matching strategies to empirical distribution of sending behavior within each treatment. Standard errors of the mean are presented in brackets below each value.

A.2 Experimental Materials

General Instructions

We warmly welcome you to this economic study. Thank you very much for your participation!

We guarantee that, at no time, another participant of the experiment is informed about your identity. Also, the experimenters are not able to assign identities to the decisions. All information provided by you will be treated confidentially and will not be disclosed to third parties. The data is used exclusively for scientific purposes.

You will receive guaranteed 6 euro for your appearance. If you read the following explanations carefully, you will be able to earn additional money - depending on your decisions and/or the decisions of the other participants. Thus, it is very important that you read these explanations carefully. If you have any questions, please direct them to us.

During the study, you are not allowed to talk to the other participants or use your mobile phone. Non-compliance of this rule will result in exclusion from the study and all payments.

During the study, we do not talk about euros, but about points. Hence, your total income is first calculated in points. The number of points you earn during the study will then be converted into euros at the end and rounded to the nearest 10 cents, where the following applies:

100 points = 10.00 Euro
(1 point = 10 Cents)

On the following pages we explain the exact procedure of the study.

If you have any questions during the study, please raise your hand and we will come to you.

Overview study process

Here is an explanation of the general process of the study. Following this, you will receive a detailed description of how each of your decisions will exactly look like.

In total, there are three different roles in the study, which we call "participant A", "participant B" and "participant C" for purposes of simplicity. You will learn later in which role you will make decisions.

In the study, participants A and B will interact as a pair. participants A get 20 points at the beginning. These participants then must decide, how many points they send to participant B. Thereby 0, 4, 8, 12, 16 or 20 points can be sent.

We will then **triple** the amount, which participant A sent to participant B. This means that participant B will receive three times the amount sent by Participant A.

After that, participant B makes his decision how much he sends back. Here, the amount can be freely chosen, but not more than the amount received. Please note: This amount will **NOT be tripled**. This means that participant A receives exact the amount that participant B sends back. Additionally, participant B can send points to participant C who initially got an endowment of 60 points.

Afterward, participant C learns what happened previously. I.e., he sees how many points participant B has sent back to participant A for each possible send amount. He also sees how many points participant B has sent to him and must decide whether he accepts them or not. If yes, he receives them. If not, 80 % of the points go back to participant B; nobody receives the remaining 20 % of the points. Afterwards, participant C can deduct points from participant B if he considers it appropriate. As a reminder: Participant C has an endowment of 60 points. For every point spent by participant C, participant B will be deducted two points. He can spend any number of points to deduct participant B's points. However, participant B cannot have less than 0 points at the end of the game. No one then receives the points deducted and spent.

At the end of the study, all participants learn all decisions of the other relevant participants, i.e., how much participant A sent to participant B, how much the latter sent back and how many points he sent to participant C. They also learn whether participant C has accepted these points and how many points he has spent to deduct points from participant B.

Summary:

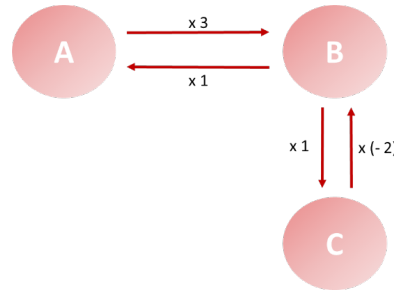


Figure 1: Overview of participants and point multipliers

Participant A receives the amount he did not send to participant B, plus the amount he receives back from participant B.

Participant B receives the tripled amount he received from Participant A minus the amount he sends back to Participant A. Also, the points he sends to Participant C are subtracted if the latter accepts, otherwise he gets back 80% of the points. In addition, his payout will be reduced by the points deducted from Participant C.

Participant C receives 60 points plus the points he receives from participant B if he accepts. The points he spent to deduct points from participant B are then subtracted.

Payout overview for entire study

In the following you will make decisions as a participant with role A, role B as well as a participant with role C.

For the payout, you will be randomly assigned two participants and one role, for which your decision and that of your participants will then be paid out in the end. That is, only one of your roles, either role A, B or C, will be selected and paid out in the end.

Important: However, your decisions in a role will have no effect on your potential payout in another role, as you will always be randomly assigned other participants for each role in any case.

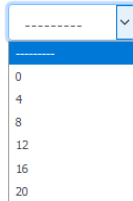
Since you do not know which role will be selected, you should make every decision thoughtfully.

Procedure on the computer

Description participant A

As participant A, you send either 0, 4, 8, 12, 16, or 20 points to a participant B. You make the decision as follows:

Wie viele Punkte wollen Sie an Teilnehmer B senden?



Click on the scroll-down menu and select the option you prefer. To confirm your choice, click on "Next".

After that, we would like you to tell us how many points you expect to receive back for each potential amount sent (i.e., 0, 4, 8, 12, 16, or 20) and whether you expect Participant B to send anything to Participant C.

As an example, assume that you have sent 4 points to participant B. Thus, participant B receives 12 points. How many points do you think he will send back to you? How many points will he send to participant C?

The participant with role B will also make a decision for all possible amounts sent to him. He will only be informed **afterwards** which amount you have actually sent. His suitable decision will then be selected for this amount.

You can receive **additional points** for the correctness of your answer. You will receive an additional 5 points for each of your answers if it differs from participant B's answer by no more than 10%.

Example:

You state that for an amount of 12 points sent by you, you expect participant B to send 10 points back to you and not to send any point to participant C. Participant B's response was 11 to you and 2 to Participant C. Therefore, you receive an additional 5 points (out of a possible 10) at the end of the study.

(For help: 10% of 11 points are 1.1 points. I.e., for each answer between 9.9 and 12.1 you get the additional 5 points. 10% of 2 points are 0.2 points, here your answer should have been between 1.8 and 2.2)

You make the decisions as follows:

Ihr gesendeter Betrag	4
Betrag, den Teilnehmer B erhält ...	12
Wie viel erwarten Sie zurück? (Geben Sie bitte 0 ein, wenn Sie denken, dass Teilnehmer B nichts sendet)	<input type="text"/>

Erwarten Sie, dass Teilnehmer B etwas an Teilnehmer C sendet? Wenn ja, wie viel? (Geben Sie bitte 0 ein, wenn Sie denken, dass Teilnehmer B nichts sendet)	<input type="text"/>
---	----------------------

Please enter the number of points you expect to be returned to you in the upper box. In the lower box, indicate whether you expect participant B to send anything to participant C. Participant B can enter any number between 0 and the maximum amount available (in the example above, "12"), and so can you. The amounts in both boxes added up cannot exceed the total number of points available (12 in the example above). To confirm your choice, click on "Next".

Description participant B

Now you must make the decision how much you want to send back to participant A and at the same time if you want to send something to participant C. As described for participant A, you must first make a decision for each possible amount sent. For the potential payout, your answer will be matched with the real amount sent by participant A.

You make the decisions as follows:

Ihr zugesendeter Betrag von Teilnehmer A	4
Ihr erhaltener Betrag (3x zugesendeter Betrag)	12
Wie viel senden Sie zurück an Teilnehmer A? (Geben Sie bitte 0 ein, wenn Sie nichts an Teilnehmer A senden möchten)	<input type="text"/>

Möchten Sie etwas an Teilnehmer C senden? Wenn ja, wie viel? (Geben Sie bitte 0 ein, wenn Sie nichts an Teilnehmer C senden möchten)	<input type="text"/>
---	----------------------

Bitte beachten Sie: Teilnehmer C hat eine Anfangsausstattung von 60 Punkten.

In both boxes you can type any number between 0 and the maximum amount available. You can enter any number between 0 and the maximum available amount (in the above example "12"). The amounts in both boxes added up also cannot exceed the total number of points available.

To confirm your choice, click on "Next".

Description participant C

As a participant with role C, you already have 60 points. You will see the decisions of a random participant B from the previous round displayed on the screen in a random order.

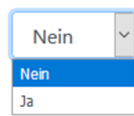
For these decisions you must decide if you want to accept the points sent by participant B. If you do not accept the points, they will go back to participant B with a 20% subtraction.

You also need to decide whether you want to deduct points from participant B. For every point you spend, two points will be deducted from the participant with role B.

If you are later assigned role C, the matching decisions to the actual choice of participant A will then be applied.

You make the decisions as follows:

Nehmen Sie die Punkte von Teilnehmer B an?



Select in the drop-down menu by clicking whether you accept the points sent to you ("Yes"), or not ("No").

(This question is only displayed to you if participant B has sent some points)

Below you select how many points you want to deduct from participant B:

Wie viele Punkte möchten Sie ausgeben um Teilnehmer B Punkte abzuziehen (Zur Erinnerung: Teilnehmer B bekommt die doppelte Punktzahl abgezogen)?

You can enter any amount in the box. However, the payout of participant B cannot fall below 0 points!

When you have understood these instructions, please turn to the screen, and click "Next". You will now be asked a few control questions to help you fully understand the study.

Room for calculations (also on the next page):

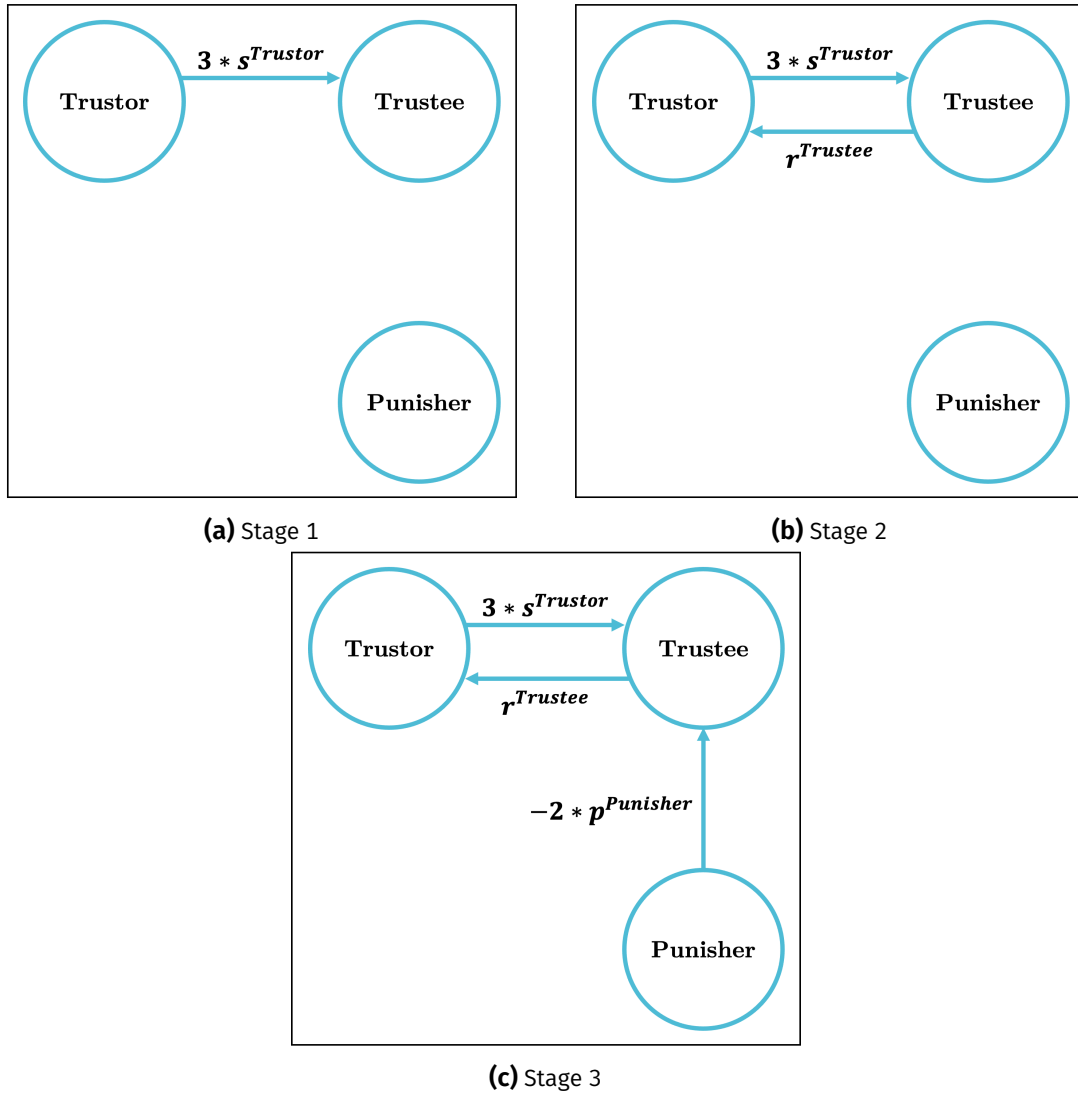


Figure A7. Punish treatment

A.2.2 Treatment Description.

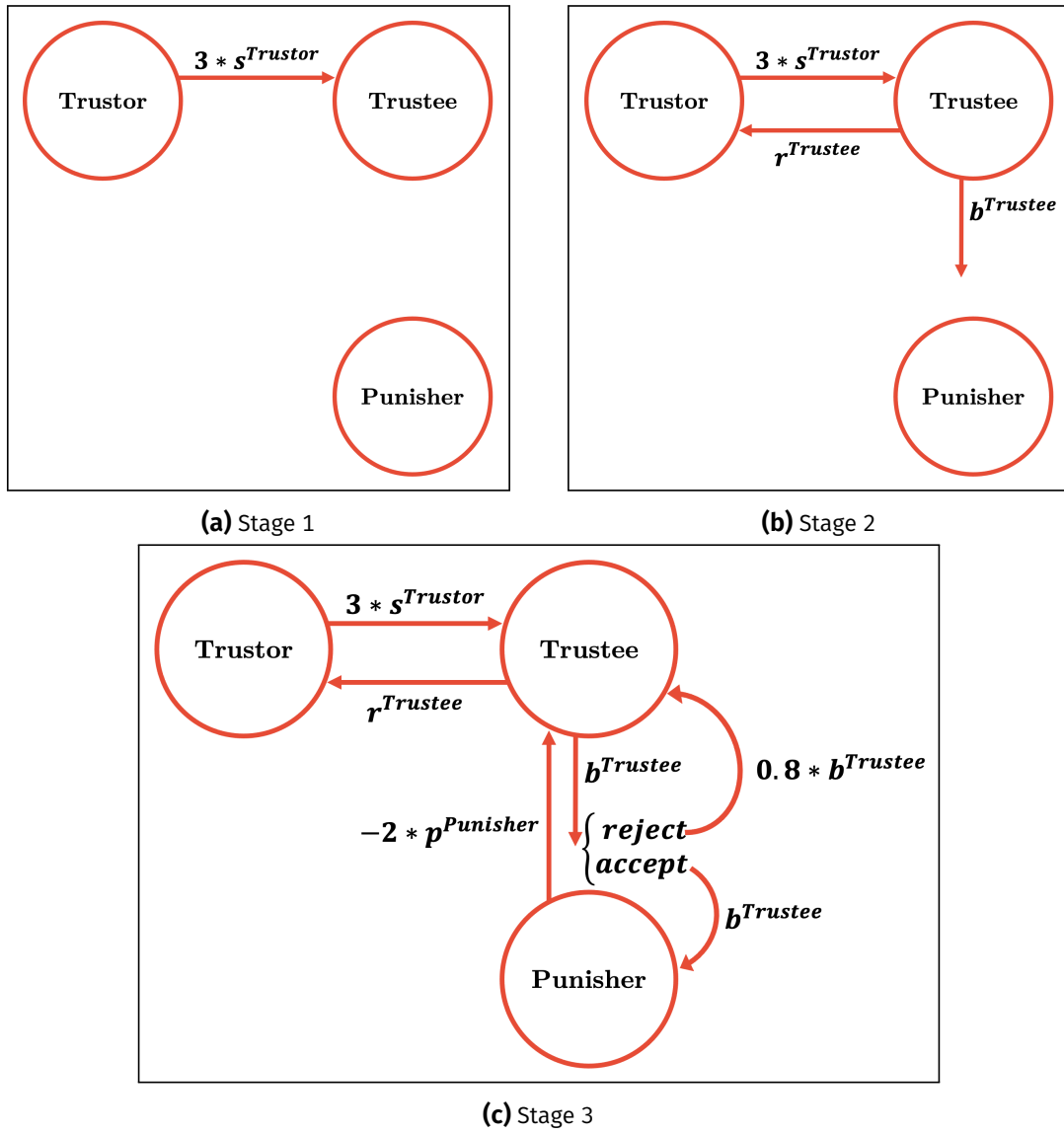


Figure A8. Bribe treatment

A.2.3 Survey Items.

Right-Wing Authoritarianism. *Items are translated by the authors of this paper. For the original German wording see Beierlein et al. (2014).*

Participants answered the following items on a five-point Likert scale: (1) absolutely do not agree, (2) agree only a little, (3) agree somewhat, (4) mostly agree, (5) agree fully.

- Strenuous actions should be taken against outsiders and slackers in society.
- Agitators should clearly feel that they are unwanted in society.
- Societies' rules should be enforced without mercy.
- We need strong leaders to live safely in society.
- Humans should leave important decisions in society to leaders.
- We should be thankful for leading figures that tell us exactly what we can do.
- Traditions should be cared for and maintained.
- Proven behavioral patterns should not be questioned.
- It is always best to do things in the usual manner.

Institutional Trust. Participants answered the following items on a five-point Likert scale, where only the endpoints were given as (1) meaning complete distrust and (5) meaning complete trust.

- federal president
- federal government
- state government
- federal parliament
- courts
- political parties
- armed forces
- police

References

- Abbink, K., Irlenbusch, B., & Renner, E. (2002). An Experimental Bribery Game. *Journal of Law, Economics, and Organization*, 18(2), 428–454. <https://doi.org/10.1093/jleo/18.2.428>. [8, 9]
- Abbink, K., Dasgupta, U., Gangadharan, L., & Jain, T. (2014). Letting the briber go free: An experiment on mitigating harassment bribes. *Journal of Public Economics*, 111, 17–28. <https://doi.org/10.1016/j.jpubeco.2013.12.012>. [8, 9]
- Acemoglu, D., & Robinson, J. A. (2008). Persistence of power, elites, and institutions. *American Economic Review*, 98(1), 267–293. <https://doi.org/10.1257/aer.98.1.267>. [2]
- Algan, Y., & Cahuc, P. (2013). Trust and Growth. *Annual Review of Economics*, 5(1), 521–549. <https://doi.org/10.1146/annurev-economics-081412-102108>. [32]
- Altemeyer, B. (1981). *Right-Wing Authoritarianism*. University of Manitoba Press. [21]
- Altemeyer, B. (1996). *The authoritarian specter*. Harvard University Press. [21]
- Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics*, 9(3), 193–208. <https://doi.org/10.1007/s10683-006-9122-4>. [29]
- Balafoutas, L., Grechenig, K., & Nikiforakis, N. (2014). Third-party punishment and counter-punishment in one-shot interactions. *Economics Letters*, 122(2), 308–310. <https://doi.org/10.1016/j.econlet.2013.11.028>. [2, 4]
- Balafoutas, L., & Nikiforakis, N. (2012). Norm enforcement in the city: A natural field experiment. *European Economic Review*, 56(8), 1773–1785. <https://doi.org/10.1016/j.euroecorev.2012.09.008>. [2]
- Banerjee, R. (2016). Corruption, norm violation and decay in social capital. *Journal of Public Economics*, 137, 14–27. <https://doi.org/10.1016/j.jpubeco.2016.03.007>. [2, 5]
- Bartling, B., Fehr, E., Huffman, D., & Netzer, N. (2021). *The Complementary Nature of Trust and Contract Enforcement*, University of Zurich. <http://www.econ.uzh.ch/static/wp/econwp377.pdf>. [3]
- Bauer, P. C., Keusch, F., & Kreuter, F. (2019). Trust and cooperative behavior: Evidence from the realm of data-sharing (V. Capraro, Ed.). *PLoS ONE*, 14(8), e0220115. <https://doi.org/10.1371/journal.pone.0220115>. [5]
- Beierlein, C., Asbrock, F., Kauff, M., & Schmidt, P. (2014). *Die Kurzskala Autoritarismus (KSA-3): Ein ökonomisches Messinstrument zur Erfassung dreier Subdimensionen autoritärer Einstellungen*, GESIS – Leibniz-Institut für Sozialwissenschaften. <https://doi.org/10.6102/zis228>. [4, 10, 23, 48]
- Bellemare, C., & Kröger, S. (2007). On representative social capital. *European Economic Review*, 51(1), 183–202. <https://doi.org/10.1016/j.euroecorev.2006.03.006>. [5]
- Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3), 141–164. <https://doi.org/10.1257/jep.30.3.141>. [7]
- Ben-Ner, A., & Halldorsson, F. (2010). Trusting and trustworthiness: What are they, how to measure them, and what affects them. *Journal of Economic Psychology*, 31(1), 64–79. <https://doi.org/10.1016/j.joep.2009.10.001>. [9]
- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, 10(1), 122–142. <https://doi.org/10.1006/game.1995.1027>. [3, 5, 6]
- Bicchieri, C., & Maras, M. (2022). Intentionality matters for third-party punishment but not compensation in trust games. *Journal of Economic Behavior & Organization*, 197, 205–220. <https://doi.org/10.1016/j.jebo.2022.02.026>. [2]
- Bicchieri, C., Xiao, E., & Muldoon, R. (2011). Trustworthiness is a social norm, but trusting is not. *Politics, Philosophy and Economics*, 10(2), 170–187. <https://doi.org/10.1177/1470594X10387260>. [5, 8]
- Burnham, T., McCabe, K., & Smith, V. L. (2000). Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior and Organization*, 43(1), 57–73. [https://doi.org/10.1016/S0167-2681\(00\)00108-6](https://doi.org/10.1016/S0167-2681(00)00108-6). [2]

- Chang, E. C., & Chu, Y. H. (2006). Corruption and trust: Exceptionalism in Asian democracies? *Journal of Politics*, 68(2), 259–271. <https://doi.org/10.1111/j.1468-2508.2006.00404.x>. [10]
- Charness, G., Cobo-Reyes, R., & Jiménez, N. (2008). An investment game with third-party intervention. *Journal of Economic Behavior and Organization*, 68(1), 18–28. <https://doi.org/10.1016/j.jebo.2008.02.006>. [2, 4, 5, 10, 15, 28, 30–32]
- Charness, G., Du, N., & Yang, C. L. (2011). Trust and trustworthiness reputations in an investment game. *Games and Economic Behavior*, 72(2), 361–375. <https://doi.org/10.1016/j.geb.2010.09.002>. [32]
- Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9, 88–97. <https://doi.org/10.1016/j.jbef.2015.12.001>. [10]
- Costa-Gomes, M. A., Huck, S., & Weizsäcker, G. (2014). Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect. *Games and Economic Behavior*, 88, 298–309. <https://doi.org/10.1016/j.geb.2014.10.006>. [29]
- Costa-Gomes, M. A., & Weizsäcker, G. (2008). Stated Beliefs and Play in Normal-Form Games. *Review of Economic Studies*, 75(3), 729–762. <https://doi.org/10.1111/j.1467-937X.2008.00498.x>. [29]
- Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, 46(2), 260–281. [https://doi.org/10.1016/S0899-8256\(03\)00119-2](https://doi.org/10.1016/S0899-8256(03)00119-2). [9, 29]
- Cronk, L. (2007). The influence of cultural framing on play in the trust game: a Maasai example. *Evolution and Human Behavior*, 28(5), 352–358. <https://doi.org/10.1016/j.evolhumbehav.2007.05.006>. [2]
- Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1), 67–80. <https://doi.org/10.1007/s00199-006-0153-z>. [7]
- Dannenber, A., & Gallier, C. (2020). The choice of institutions to solve cooperation problems: a survey of experimental research. *Experimental Economics*, 23(3), 716–749. <https://doi.org/10.1007/s10683-019-09629-8>. [30]
- Dunning, D., Anderson, J. E., Schlösser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality and Social Psychology*, 107(1), 122–141. <https://doi.org/10.1037/a0036673>. [29]
- Dunning, D., Fetchenhauer, D., & Schlösser, T. M. (2012). Trust as a social and emotional act: Noneconomic considerations in trust behavior. *Journal of Economic Psychology*, 33(3), 686–694. <https://doi.org/10.1016/j.joep.2011.09.005>. [29]
- Engelmann, D., & Strobel, M. (2004). Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments. *American Economic Review*, 94(4), 857–869. <https://doi.org/10.1257/0002828042002741>. [31]
- Engl, F., Riedl, A., & Weber, R. (2021). Spillover Effects of Institutions on Cooperative Behavior, Preferences, and Beliefs. *American Economic Journal: Microeconomics*, 13(4), 261–299. <https://doi.org/10.1257/mic.20180336>. [12]
- European Bank for Reconstruction and Development, & World Bank. (2011). Life in Transition Survey (LiTS) 2010. [10]
- Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2022). The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences. *Management Science*, (Articles in Advance), 1–16. <https://doi.org/10.1287/mnsc.2022.4455>. [10, 28]
- Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2), 63–87. [https://doi.org/10.1016/S1090-5138\(04\)00005-4](https://doi.org/10.1016/S1090-5138(04)00005-4). [2, 4]
- Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, 13(1), 1–25. <https://doi.org/10.1007/s12110-002-1012-7>. [2, 32]

- Fehr, E., Fischbacher, U., von Rosenblatt, B., Schupp, J., & Wagner, G. G. (2002). A Nationwide Laboratory Examining Trust and Trustworthiness by Integrating Behavioural Experiments into Representative Surveys. *Schmollers Jahrbuch: Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 122(4), 519–542. [5]
- Fehr, E., & Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, 90(4), 980–994. <https://doi.org/10.1257/aer.90.4.980>. [32]
- Fehr, E., & Schneider, F. (2010). Eyes are on us, but nobody cares: are eye cues relevant for strong reciprocity? *Proceedings of the Royal Society B: Biological Sciences*, 277(1686), 1315–1323. <https://doi.org/10.1098/rspb.2009.1900>. [5]
- Fehr, E., & Williams, T. (2018). *Creating an Efficient Culture of Cooperation*, University of Zurich, Department of Economics, Working Paper No. 267. <https://doi.org/10.2139/ssrn.3062528>. [2]
- Fiedler, M., & Haruvy, E. (2017). The effect of third party intervention in the trust game. *Journal of Behavioral and Experimental Economics*, 67, 65–74. <https://doi.org/10.1016/j.socec.2016.10.003>. [2, 5, 31]
- Frey, B. S., Benz, M., & Stutzer, A. (2004). Introducing procedural utility: Not only what, but also how matters. *Journal of Institutional and Theoretical Economics*, 160(3), 377–401. <https://www.jstor.org/stable/40752468>. [30]
- Frey, B. S., & Stutzer, A. (2005). Beyond outcomes: Measuring procedural utility. *Oxford Economic Papers*, 57(1), 90–111. <https://doi.org/10.1093/oep/gpi002>. [30]
- Gächter, S., Herrmann, B., & Thöni, C. (2004). Trust, voluntary cooperation, and socio-economic background: survey and experimental evidence. *Journal of Economic Behavior & Organization*, 55(4), 505–531. <https://doi.org/10.1016/j.jebo.2003.11.006>. [5]
- Gambetta, D. (1988). Can We Trust Trust. In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations* (pp. 213–237). Basil Blackwell. [2, 9]
- Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring Trust. *Quarterly Journal of Economics*, 115(3), 811–846. <https://doi.org/10.1162/003355300554926>. [3]
- Gneezy, U., Van Veldhuizen, R., & Saccardo, S. (2019). Bribery: Behavioral drivers of distorted decisions. *Journal of the European Economic Association*, 17(3), 917–946. <https://doi.org/10.1093/jeea/jvy043>. [2, 7, 8]
- 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>. [10]
- Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2014). On cooperation in open communities. *Journal of Public Economics*, 120, 220–230. <https://doi.org/10.1016/j.jpubeco.2014.10.001>. [2]
- Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon. [21]
- Haisley, E. C., & Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior*, 68(2), 614–625. <https://doi.org/10.1016/j.geb.2009.08.002>. [7]
- Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardaroas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesoronol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science*, 312(5781), 1767–1770. <https://doi.org/10.1126/science.1127333>. [2]
- Herreros, F. (2023). The State and Trust. *Annual Review of Political Science*, 26(1), 603–626. <https://doi.org/10.1146/annurev-polisci-051921-102842>. [31]
- Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial Punishment Across Societies. *Science*, 319(5868), 1362–1367. <https://doi.org/10.1126/science.1153808>. [5]
- Iriberri, N., & Rey-Biel, P. (2011). The role of role uncertainty in modified dictator games. *Experimental Economics*, 14(2), 160–180. <https://doi.org/10.1007/s10683-010-9261-5>. [31]
- Jordan, J., McAuliffe, K., & Rand, D. (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, 19(4), 741–763. <https://doi.org/10.1007/s10683-015-9466-8>. [2, 4]

- Kaufmann, D. (2005). Myths and Realities of Governance and Corruption. *Global competitiveness report 2005-06* (pp. 81–98). World Economic Forum. <https://doi.org/10.2139/ssrn.829244>. [2]
- King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, 308(5718), 78–83. <https://doi.org/10.1126/science.1108062>. [2]
- Knack, S., & Keefer, P. (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *The Quarterly Journal of Economics*, 112(4), 1251–1288. <https://doi.org/10.1162/003355300555475>. [32]
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>. [7]
- Li, C., Turmunkh, U., & Wakker, P. P. (2019). Trust as a decision under ambiguity. *Experimental Economics*, 22(1), 51–75. <https://doi.org/10.1007/s10683-018-9582-3>. [9, 29]
- Mishler, W., & Rose, R. (2001). What are the origins of political trust? Testing institutional and cultural theories in post-communist societies. *Comparative Political Studies*, 34(1), 30–62. <https://doi.org/10.1177/0010414001034001002>. [10]
- Murtin, F., Fleischer, L., Siegerink, V., Aassve, A., Algan, Y., Boarini, R., Gonzalez, S., Lonti, Z., Grimalda, G., Vallve, R. H., Kim, S., Lee, D., Putterman, L., & Smith, C. (2018). Trust and its determinants: Evidence from the Trustlab experiment. *OECD Statistics Working Papers*, 33, 1–75. <https://doi.org/https://doi.org/10.1787/869ef2ec-en>. [11]
- Muthukrishna, M., Francois, P., Pourahmadi, S., & Henrich, J. (2017). Corrupting cooperation and how anti-corruption strategies may backfire. *Nature Human Behaviour*, 1(7), 1–5. <https://doi.org/10.1038/s41562-017-0138>. [5]
- Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, 92(1-2), 91–112. <https://doi.org/10.1016/j.jpubeco.2007.04.008>. [7]
- Nikiforakis, N., & Mitchell, H. (2014). Mixing the carrots with the sticks: third party punishment and reward. *Experimental Economics*, 17(1), 1–23. <https://doi.org/10.1007/s10683-013-9354-z>. [2]
- Nyarko, Y., & Schotter, A. (2002). An Experimental Study of Belief Learning Using Elicited Beliefs. *Econometrica*, 70(3), 971–1005. <https://doi.org/10.1111/1468-0262.00316>. [29]
- 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>. [2]
- Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature*, 444(7120), 718–723. <https://doi.org/10.1038/nature05229>. [2, 4, 32]
- Sapienza, P., Toldra-Simats, A., & Zingales, L. (2013). Understanding trust. *Economic Journal*, 123(573), 1313–1332. <https://doi.org/10.1111/ecoj.12036>. [3, 9, 14, 17]
- Schwerter, F., & Zimmermann, F. (2020). Determinants of trust: The role of personal experiences. *Games and Economic Behavior*, 122, 413–425. <https://doi.org/10.1016/j.geb.2020.05.002>. [5, 12]
- Stenner, K., & Haidt, J. (2018). Authoritarianism is not a momentary madness, but an eternal dynamic within liberal democracies. *Can it happen here? authoritarianism in america*. [21]
- Walkowitz, G. (2021). Dictator game variants with probabilistic (and cost-saving) payoffs: A systematic test. *Journal of Economic Psychology*, 85(April), 102387. <https://doi.org/10.1016/j.joep.2021.102387>. [31]