I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16346

Behavioural Responses to Unfair Institutions: Experimental Evidence on Rule Compliance, Norm Polarisation, and Trust

Simon Columbus
Lars P. Feld
Matthias Kasper
Matthew D. Rablen

JULY 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16346

Behavioural Responses to Unfair Institutions: Experimental Evidence on Rule Compliance, Norm Polarisation, and Trust

Simon Columbus
University of Copenhagen

Lars P. Feld
Albert-Ludwigs-Universität Freiburg

Matthias Kasper
Walter Eucken Institute, Freiburg

Matthew D. Rablen
University of Sheffield and IZA

JULY 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Behavioural Responses to Unfair Institutions: Experimental Evidence on Rule Compliance, Norm Polarisation, and Trust*

This study investigates the effects of unfair enforcement of institutional rules on public good contributions, personal and social norms, and trust. In a preregistered online experiment (n = 1,038), we find that biased institutions reduce rule compliance compared to fair institutions. However, rule enforcement – fair and unfair – reduces norm polarisation compared to no enforcement. We also find that social heterogeneity lowers average trust and induces ingroup favouritism in trust. Finally, we find consistent evidence of peer effects: higher levels of peer compliance raise future compliance and spillover positively into norms and trust. Our study contributes to the literature on behavioural responses to institutional design and strengthens the case for unbiased rule enforcement.

JEL Classification: H41, C72, C91, C92

Keywords: public goods, compliance, social norms, trust, audits, biased rule enforcement, polarisation

Corresponding author:
Matthew D. Rablen
Department of Economics
University of Sheffield
9 Mappin Street
Sheffield, S1 4DT
United Kingdom
E-mail: m.rablen@sheffield.ac.uk

* We thank the Forum Ordnungspolitik e.V. for financial support and Arrita Domi and Berit Heling for help with implementing the experiment. We also thank participants of the 2023 IAREP/SABE conference for valuable comments. This study was pre-registered at https://osf.io/qaedu/?view only=262ca0dcde3e41ad98778c2bb1141be5. Columbus, Kasper, and Rablen contributed equally and share first authorship.

1 Introduction

Rules are a central element of human civilisation, needed for preserving social order and cohesion, and as the basis for any form of large-scale cooperation (e.g. Hayek, 1973; Bicchieri, 2006; North, 1990). Laws prescribe formal sanctions in order to deter rule violations. Early economic analyses of the law assumed that the expected cost of such formal sanctions needed to outweigh the benefits of breaking the law to be deterrent (Becker, 1968). However, laws also have an "expressive function" (Sunstein, 1996), signalling or establishing social norms about what is considered to be appropriate and what is not (Benabou and Tirole, 2011; Lane et al., 2023; Sunstein, 1996). Social norms, in turn, are enforced by informal sanctions, including expressions of disapproval, ridicule, and social ostracism (Bicchieri, 2006; Posner, 1997; Fehr and Gächter, 2000; Masclet et al., 2003; Balafoutas and Nikiforakis, 2012; Balafoutas et al., 2014; Molho et al., 2020). Consequently, even laws proscribing weak or infrequent formal sanctions may deter noncompliance by signalling social norms, especially if the rules and their enforcement are perceived as legitimate (e.g., because they are established by democratic means; Dal Bó et al., 2010; Markussen et al., 2014; Tyran and Feld, 2006). Conversely, illegitimate legal institutions may be inefficient or even counterproductive if they fail to signal, or even undermine, cooperative social norms.

In this study we examine the causal effects of fair and unfair rule enforcement on compliance, social norms, and trust in heterogeneous groups. To study these relationships, we use a public goods game and introduce variation in group composition by assigning participants to homogeneous or heterogeneous groups, where heterogeneous groups comprise two distinct subgroups. Subgroup membership is merely a label, i.e., an uninformative signal. To isolate the pure effect of heterogeneity, we assign group membership randomly. We then study the effects of rule enforcement in homogeneous and heterogeneous groups. Specifically, all treatments include a rule that requires participants to contribute a share of their endowment to the public good. By introducing fair (i.e., equal-probability) and unfair (i.e., subgroup-biased) monitoring of rule violations, we are able to identify the causal effects of institutional fairness.

Recent history reveals various examples where law enforcement institutions have applied and enforced rules unfairly. In the US, for example, the Lois Lerner scandal of 2013 – whereby the US tax authority (IRS) was accused of political bias in targeting right-leaning nonprofit organisations for scrutiny – fuelled a belief among right-wing groups of

a leftward bias in the machinations of the state and ultimately resulted in substantial budget cuts for the IRS (Kiel and Eisinger, 2018). More recently, the IRS has faced accusations of racial bias in tax audits. Specifically, a widely noted study finds that Black US taxpayers are three to five times more likely to be audited than non-Black taxpayers (Elzayn et al., 2023). The Netherlands witnessed a childcare benefits scandal in which claimants were targeted for audits based on their foreign ancestry, a practice which has been described as "discriminatory" and arising from "institutional bias" (Dutch Data Protection Authority, 2020). The scandal led to the resignation of the third cabinet of prime minister Mark Rutte in 2021.[1] These cases are notable because they involve accusations that the monitoring of rule violations was biased, rather than the penalties themselves.

It is important to note that institutional bias might – in some instances – be economically efficient. For example, a recent study finds that the returns from tax audits are substantially higher at the upper end of the income distribution (Boning et al., 2023). Our study, however, investigates institutional bias that is not rooted in such efficiency considerations. Instead, our focus in this paper is on the effects of arbitrary unfairness in rule enforcement. There is initial indication that such institutional unfairness might undermine the willingness to follow rules. For example, a recent experimental study finds that non-compliance is a vehicle for retribution after "unfair" tax audits that overestimate the taxpayer's true income (Lancee et al., 2023). Prior work has documented a positive correlation between tax evasion and the frequency of tax audits, suggesting that taxpayers might respond to what they perceive as excessive audit risk by reducing their compliance (Mendoza et al., 2017). More generally, people are more likely to obey the law when they think that the system of government treats them and others fairly (e.g., Tyler and Blader, 2000; Tyler and Huo, 2002).

However, the effects of institutional bias might extend beyond rule compliance and spill over into interactions among individuals, undermining trust and cooperative norms. Experimental studies show that impartial sanctioning institutions can have positive spillovers on trust and social norms in subsequent, unregulated interactions (Cassar et al., 2014; Engl et al., 2021; Peysakhovich and Rand, 2016; Stagnaro et al., 2017). When these institutions are corrupt or biased, however, this may undermine trust. For example, survey evidence and experiments show that exposure to a corrupt institution reduces trust towards other

[1] As a further example, in the UK, the London ("Metropolitan") police have twice been found "institutionally racist" by judicial inquiries (Macpherson, 1999; Casey, 2023), and suffer low levels of trust within London's ethnic minority communities (Atkinson, 2023).

individuals (Spadaro et al., 2023). At a global level, weak institutions are associated with greater levels of dishonesty in individual interactions (Gächter and Schulz, 2016), and trust in institutions is causally linked to generalised trust in strangers (Sønderskov and Dinesen, 2016). Thus, the quality and impartiality of institutions may influence rule compliance not just through deterrence, but also by promoting or undermining social norms and trust among individuals.

Biased rule enforcement might also contribute to polarisation in personal and social norms. When rules are enforced unfairly across social groups, individuals might develop diffuse or multi-modal empirical and normative expectations (Dimant et al., 2023) instead of coordinating on one normative standard (Krupka and Weber, 2013). More specifically, there are at least two channels by which biased rule enforcement may contribute to polarisation. First, unequal enforcement of rules may signal that one group's rule violations are more acceptable than another group's rule violations. Second, differences in the probabilities of sanctions for rule violations might induce group differences in rule compliance (Kasper and Alm, 2022a).[2] People frequently infer social norms from observed behaviour (Li et al., 2021; Welch et al., 2005; Lindström et al., 2018; Tworek and Cimpian, 2016) and may thus conclude that groups whose members behave differently also vary in their normative beliefs. Such perceptions may also give rise to increasing in-group favouritism if members of different groups perceive each other as less trustworthy (Balliet et al., 2014). Thus, biased institutions may contribute to normative polarisation and undermine trust in heterogeneous populations.

To tease out the separate effects of group membership and institutional fairness, we compare outcomes under fair rule enforcement, biased rule enforcement, and under no rule enforcement at all. Specifically, our experimental treatments introduce variation in three dimensions. First, we vary the composition of social groups, i.e., we introduce a "homogeneous" treatment in which players are indistinguishable, and a "heterogeneous" treatment in which players are randomly assigned to "red" or "blue" subgroups. Second, we introduce variation in the existence of rule enforcement. In particular, all experimental treatments include a rule that requires participants to contribute 50% of their endowment

[2]Further harmful effects of biased rule enforcement include reductions in productivity (Glover et al., 2017), loss of intrinsic motivation in the workplace (Kim and Rubianty, 2011; Hartmann and Slapničar, 2012), increased aggressiveness and willingness to behave unethically (Neuman, 2004), an increased inclination among members of the disadvantaged group to harm members of other groups (Zizzo and Oswald, 2001; Grosch and Rau, 2020; Lancee et al., 2023), and the desire for retribution (Greenberg, 1990; Tyler and Lind, 2001).

to the public good. Some treatments then implement an audit scheme, where contributions to the public good are randomly audited, while others do not. Finally, we vary the fairness of rule enforcement. To this end, we distinguish between treatments with fair institutions, which audit all players with the same probability, and unfair institutions, where specific subgroups of players are audited more frequently than others.

Our findings suggest that unbiased rule enforcement increases compliance even in the presence of a strong norm of cooperation. Surprisingly, however, we find no evidence that biased rule enforcement undermines cooperation, social norms, or trust. Instead, we find that the introduction of audits – irrespective of their fairness – increase "exact" rule compliance by reducing freeriding and crowding out cooperation beyond the required minimum contribution. Normative expectations followed this pattern, which means that both fair and unfair audits reduce the polarisation of social norms. Beyond these treatment effects, we find consistent evidence for peer effects across all outcome variables. In particular, a higher number of compliant peers in the first round of the public goods game translates into higher compliance levels in the final round of the game, stronger personal and social norms, as well as higher levels of trust.

Our findings contribute to different strands of research in economics and the behavioral sciences. For example, our results complement current work on the causal effect of laws on social norms (Lane et al., 2023) by showing that the form and enforcement of laws may affect the distribution of normative expectations. Similarly, our study provides a broader perspective on behavioural responses to unfair rule enforcement (Lancee et al., 2023) by demonstrating that the existence of biased enforcement (rather than personal experience of unfair treatment) affects rule compliance. Our results also support current work finding nuanced effects of incentives on prosocial behaviour (Graf et al., 2023) and relate to recent studies on the effects of minimal group identities (Espín et al., 2023; Dimant, in press) and the effects of social information on empirical expectations (Dimant et al., 2023). While we believe our paper is the first to study uninformative group membership and institutional bias in the public goods game, we contribute to a wider literature on heterogeneity in a public goods setting. Some studies vary the initial endowments (Kingsley, 2016) or the ability of players to monitor and punish other players (Boosey and Isaac, 2016; Nikiforakis et al., 2010; Burton-Chellew and Guérin, 2021), whereas others sort players into either identity-homogeneous or identity-heterogeneous groups (Bicskei et al., 2016; Charness et al., 2014; Drouvelis et al., 2021; Martinangeli and Martinsson, 2020). We also enlarge an experimental literature utilising biased procedures outside the public good game context

(Zizzo and Oswald, 2001; Grosch and Rau, 2020) and connect to the growing literature on social norm elicitation (Bicchieri and Xiao, 2009; Krupka and Weber, 2013; Dimant, 2022; Gächter et al., 2023), to a wider literature on procedural justice (Tyler and Lind, 2002; Tyran and Feld, 2006), and to the academic debate on the drivers of political polarisation (see, e.g., Boxell et al., 2017; McCoy et al., 2018; Carothers and O'Donohue, 2019; Di Tella et al., 2021; Levy, 2021; Dimant, in press). Finally, we provide further evidence of peer effects on cooperation, social norms, and trust (Isler and Gächter, 2022; Gächter et al., 2017).

The plan of the paper is as follows: Section 2 describes the experiment, performed using the online platform Prolific. Section 3 presents the results, and Section 4 concludes with some policy implications for the balance between efficiency (prediction) and equity (randomness) in law enforcement.

2 Experimental Design

Our experimental design seeks to understand the impact of group heterogeneity, $H \in \{0,1\}$, and institutional bias in enforcement, $B \in \{0,1\}$, on three outcomes: rule compliance, norms, both personal ($N_{personal}$) and social (N_{social}), and trust (T). To capture these determinants, our experiment comprises three stages: a public goods game, a norm elicitation task, and a trust game. We endow the public goods game with a minimum contribution rule R, given which rule compliance is determined by the level of contributions, g. To organize ideas, we may represent these outcomes as a system

$$g := g\left(I_E, H, B, T, N_{personal}, N_{social}; \mathbf{\Phi}_g\right); \tag{1}$$

$$N_z := N_z\left(H, B; \mathbf{\Phi}_z\right) \qquad z \in \{\text{personal,social}\}; \tag{2}$$

$$T := T(H, B; \mathbf{\Phi}_T); \tag{3}$$

In (1), $I_E \in \{0,1\}$ is an indicator for institutional rule enforcement – the presence or absence of which regulates the monetary incentives for rule-breaking, as emphasised in, e.g., Becker (1968). In each equation (1)-(3) $\mathbf{\Phi}$ is a vector of all other determinants. In particular, the elements of $\mathbf{\Phi}_g$ include rule-following behaviour of an automatic or morally driven form, as discussed in detail in Gächter et al. (2023), and kindness or other prosocial "warm glow" motives, which may interact negatively with monetary incentives (e.g., Gneezy and Rustichini, 2000).

We highlight three points from this representation. First, norms and trust are viewed as outcomes in their own right, but also as ingredients into the contribution decision. Second, institutional bias may affect all three outcomes directly, but also affect contribution (compliance) outcomes indirectly via norms and trust. Thus, our findings for compliance reflect both these direct and indirect channels. In our findings, the direct channel shall predominate. Third, we introduce bias in an initial public goods game – manipulating the argument B in equation (1) – but detect its impact on trust in equation (3) in a subsequent unenforced trust game. Thus, measured variation in trust in equation (3) is a spillover from a regulated context to one that is unregulated, as documented in, e.g., Engl et al. (2021).

We now describe the three experimental stages in detail (see Figure 1): In the public goods game we implement a between-subjects design with four experimental treatments. Specifically, we introduce variation in the group composition (*homogeneous* versus *heterogeneous groups*), the absence or presence of audits (*no audits* versus *audits*), and the fairness of audits (*random audits* versus *biased audits*). This results in the following treatments which are detailed below:[3]

1. Homogeneous groups without audits (BaseHom)
2. Heterogeneous groups without audits (BaseHet)
3. Heterogeneous groups with random audits (AuditHet)
4. Heterogeneous groups with biased audits (BiasedHet)

2.1 Public Goods game

The basis of our experiment is a standard public goods game with ten rounds. Participants are assigned randomly to groups of $n = 6$ members, which remain fixed for the duration of the experiment. In every round each player receives an endowment of $E = 10$ points. Each player decides independently how to allocate these points between a private account and a group account. Points allocated to the private account yield one point each for the player. Points allocated to the group account are tripled and redistributed equally across

[3]We also implemented a second baseline treatment with homogeneous groups and random audits but, due to a coding error, the data are uninformative and excluded from the analyses. This treatment was not necessary for any of our preregistered hypotheses.

all players, so that every point contributed to the group account, $g_i \in \{0, 1, \ldots, 10\}$, yields $\mu = 0.5$ points for each group member. Thus, individual payoffs, π_i, are determined by

$$\pi_i = E - g_i + \mu \sum_{j=1}^{n} g_j. \tag{4}$$

All treatments include a contribution rule. Specifically, participants are told that they must make a minimum contribution R of five points, or 50 percent of their endowment, to the group account. However, participants are instructed that they may contribute any amount between zero points and ten points and that each group member has the same choice to make.

2.1.1 Group composition

We introduce variation in the group composition across treatments. Specifically, in the treatment with *homogeneous* groups (BaseHom), players interact within groups of six players throughout the game. In contrast, in the treatments with *heterogeneous* groups (BaseHet, AuditHet, BiasedHet), each player is randomly assigned to a "red" or a "blue" subgroup, so that each group of six players comprises two subgroups with three players each. In all treatments the group composition, including the player's subgroup, remain constant throughout the game, and each player knows the colour of their subgroup.

2.1.2 Audits

We also introduce variation across treatments in the *audit* mechanism, i.e., the institutional mechanism to check contributions to the group account. In the two *baseline* treatments (BaseHom and BaseHet), contributions to the public good are not audited. In contrast, in the audit treatments (AuditHet and BiasedHet) players face a probability $p \in (0, 1)$ of being audited, an event indicated by $a \in \{0, 1\}$. If a player is audited and the audit reveals that the contribution is less than the required minimum contribution of five points, the player pays a fine f that is $s = 2$ times the difference between the player's contribution and the required minimum contribution, or $f_i = s(R - g_i)$.

Thus, in the audit treatments, payoffs are determined by

$$\pi_i = \begin{cases} E - g_i + \mu \sum_{j=1}^{n} g_j & \text{if } g_i \geq R; \\ E - g_i + \mu \sum_{j=1}^{n} g_j - \mathbf{1}_{a=1} \times f_i & \text{if } g_i < R. \end{cases}$$

In expectation, this simplifies to

$$\mathbf{E}(\pi_i) = E - g_i + \mu \sum_{j=1}^{n} g_j - ps \max\{R - g_i, 0\}. \tag{5}$$

In the treatment with *random audits* (AuditHet), all players are audited with probability $p = .2$. In contrast, in the treatment with *biased audits* (BiasedHet), players in the blue subgroup are audited with a low probability of $p = .1$ (BiasedHetL), whereas players in the red subgroup are audited with a high probability of $p = .3$ (BiasedHetH). The audit probabilities (of both subgroups) are common knowledge in all audit treatments. We design the audit mechanism so that sanctions are imperfect, i.e., the sanctions are non-deterrent for rationally self-interested agents (Engel, 2013; Tyran and Feld, 2006).

2.2 Social Norms

After participants have completed the public goods game, we elicit social norms by adapting methods from Bicchieri and Xiao (2009) and Dimant (2022). Specifically, we first assess personal normative beliefs $N_{personal}$ by asking participants, *"Personally, how many points do you think would be the appropriate contribution to the group account?"* Participants use a slider with range 0–10 to indicate their personal normative beliefs.

Subsequently, we elicit normative expectations in the form of expectations about the distribution of responses to the above question. To this end, we ask participants to indicate how many out of ten participants in the same treatment $n \in \{0, 1, \ldots, 10\}$ they believe stated each possible level of personal normative belief $N_{personal} \in \{0, 1, \ldots, 10\}$. Participants must allocate exactly ten points (one for each other player) across the eleven possible responses for the appropriate contribution to the group account. We define the social norm as the mean normative expectation, and the degree of polarisation of the social norm as the within-person standard deviation of the normative expectation.[4]

2.3 Trust

We elicit trust towards other group members (in treatments with *homogeneous* groups), respectively towards members of both subgroups (in treatments with *heterogeneous* groups) through sender decisions in a trust game (Berg et al., 1995). In the first part of the game,

[4]We do not incentivise the elicitation of normative expectations as we are not aware of a scoring rule that allows incentive-compatible elicitation of beliefs about distributions of ordinal variables. Dimant (2022) introduces a method to elicit normative expectations when these may be more or less uncertain. However, this method incentivises beliefs about the modal personal normative belief in the population. In contrast, we seek to elicit beliefs about the shape of the distribution of personal normative beliefs.

all players receive an endowment of $E = 10$ points and act as a sender towards a randomly selected member of their group. They may send any amount $M \in \{0, 1, \ldots, 10\}$ to the receiver and the amount sent is tripled. The amount not sent remains in the sender's possession. In the treatments with *heterogeneous* groups we use the strategy method to elicit trust towards a randomly selected receiver from the "red" and the "blue" subgroups. In the second part, all players again receive a ten-point endowment and act as the receiver to decide how much, up to a maximum of $3M$, to return to the sender. The amount not returned remains in the possession of the receiver. We use the strategy method to elicit receivers' decisions for each $M \in \{0, 1, \ldots, 10\}$.

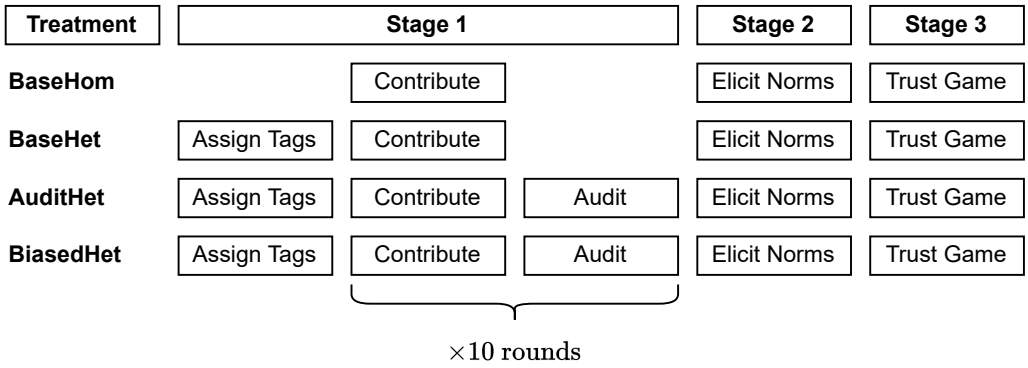

Figure 1: The experimental setup.

2.4 Experimental procedure

Figure 1 illustrates the structure of the experiment. After entering the experiment, all participants receive detailed instructions on the public goods game. Participants in the *heterogeneous* treatments are informed about the existence of two subgroups within their group and the colour they have been assigned. Subsequently, all participants must correctly answer four comprehension questions on the rules of the public goods game and the computation of their payoffs to move on. Participants in the baseline treatments continue directly to the first contribution decision. Participants in the audit treatments receive additional instructions on the audit mechanism and must pass another set of comprehension questions to move on. Specifically, players have to answer four questions on the audit probabilities in both subgroups as well as the fines for noncompliance. Subsequently, participants are randomly assigned to groups of six players. The groups do not change throughout the public goods game.

Participants then proceed to the first contribution decision, where they decide how much of their endowment of $E = 10$ points (1 point = £0.10) they want to contribute to the public good. After each contribution decision, participants in the *baseline* treatments learn the contributions of the other players as well as their earnings before advancing to the next round. In the *audit treatments* participants are selected for an audit with probability p. If an audit occurs and the player contributed less than five points to the group account, the player receives a fine f that is deducted from the earnings in that round. All players are informed about whether they were audited or not, whether the audit resulted in a fine, and how much they earned in this round. Participants also receive information about the contributions, audits, and fines of all other group members. In the treatments with *heterogeneous* groups this information is presented together with the colours of the other players. Players' IDs are randomised each round to prevent individual reputation building.

This procedure is repeated for ten rounds, though participants do not know the number of rounds. Once participants have completed the final round of the public goods game, one round is randomly selected and the players' earnings in this round are converted to Pounds Sterling and paid out to the participants. The maximum bonus payment for the public goods game is £2.

After the end of the public goods game, all players indicate their personal normative beliefs with respect to their group. Then, the players indicate their normative expectations for ten other players in their treatment as described above.

Finally, participants play two trust games. First, all participants take the role of the sender and decide how many points M (1 point = £0.05) of their 10 point endowment to send to a randomly selected recipient from their group. In the treatments with *heterogeneous* groups we use the strategy method to elicit trust towards a randomly selected player from the "red" and the "blue" subgroups. Subsequently, all players, now taking the role of receivers and again endowed with ten points, can return any integer amount up to $3M$ to the sender they have been matched with. We use the strategy method to elicit receivers' decisions for each $M \in \{0, 1, \ldots, 10\}$. Once participants have made their decisions, one game is randomly selected (i.e., either the game in which the player was the sender, or the game in which the player was the receiver), players' earnings in this game are converted to monetary amounts and are paid out to the participants. The maximum bonus payment for the trust game is £2.

2.5 Data

We ran the experiment on Prolific (https://prolific.co) in April 2023. On average, the study lasted between 15 minutes (in the treatments *without audits*) and 20 minutes (in the treatments *with audits*). Participants were paid the equivalent of £9.00 per hour (£2.25–£3.00) as fixed compensation. Additionally, participants received bonus payments of up to £4.00 (up to £2.00 from a randomly selected round of the public goods game and up to £2.00 from the trust game).

We aimed to recruit 408 participants, or 68 groups of six players, in the treatment with *biased audits* (BiasedHet). In all other treatments, we aimed to recruit 204 participants per treatment. The aspired sample size of $n = 1,020$ is substantially larger than the average sample size in prior experimental work studying public good games ($n_{mean} = 146$, Spadaro et al., 2022) or tax compliance games ($n_{mean} = 235$, Alm and Malézieux, 2021). Our final sample consists of $n = 1,038$ participants (173 groups). We exclude participants who failed to pass either comprehension check or who did not complete all ten rounds of the public goods game. Table A.1 shows the effective sample sizes per treatment. Participants are from the UK and balanced in terms of gender. The mean age is 40 years ($SD = 13.6$).

We take a host of measures to ensure confidence in the quality of our data. First, we rely on Prolific, which is considered to produce high-quality data compared to other online platforms (Peer et al., 2022; Douglas et al., 2021).

Second, we implement a generous incentive structure that places emphasis on variable compensation. In particular, sanctions in the public goods game are imperfect, i.e., self-interested participants have a financial incentive to ignore the contribution rule and free-ride, even in the treatments with rule enforcement.

Third, we implement a series of carefully designed comprehension checks to ensure participants have understood the rules of the public good games, including the contribution rule, the computation of their payoffs, as well as the composition of the (sub-)groups in their treatments. Moreover, participants in the treatments *with audits* answer additional questions on the audit probability in their group (respectively the audit probability in each subgroup), the fines for noncompliance, and the effects of fines on their earnings before they can proceed to the first contribution decision of the experiment. Participants who fail to answer the comprehension check questions are returned to the instructions until they answer the questions correctly or drop out of the experiment. Table A.2 provides information on comprehension check performance. Among participants who passed

the comprehension checks, the median number of attempts to complete the check questions is 1, suggesting that participants who contributed in the experiment understood the instructions well.

Fourth, all instructions are adapted from prior work with only minimal modifications. Specifically, the instructions to the public goods and trust games are adapted from Thielmann et al. (2021), while the measures of social norms are adapted from Bicchieri and Xiao (2009) and Dimant (2022).

2.6 Preregistration and Open Data

The procedure and key hypothesis tests were preregistered on the Open Science Framework, https://osf.io/qaedu/?view_only=262ca0dcde3e41ad98778c2bb1141be5. The experimental files, data, and code are available at https://osf.io/6by3c/?view_only=aa1919f1dbab427b97a94f5a26934041. All analyses were conducted using R 4.0.4 (R Core Team, 2022) and tidyverse (Wickham et al., 2019). Regression analyses were conducted using the estimatr package (Blair et al., 2022). We describe all preregistered hypotheses, deviations from the preregistration, and hypothesis tests in Appendix B. In the following, we use the term 'explore' to distinguish analyses which were not included in the preregistration.

3 Results

This section presents our results on the effects of institutional fairness on public good contributions, rule compliance, personal normative beliefs (personal norms) and normative expectations (social norms), as well as trust. All treatment comparisons are based on Wald tests with robust standard errors clustered at the group level.

Table 1: Estimated marginal means and cluster-robust standard errors.

Treatment	Contribution	Compliance	Personal Norm	Social Norm	Trust
BaseHom	6.16(0.27)	0.77(0.03)	6.23(0.31)	5.53(0.23)	6.75(0.20)
BaseHet	6.30(0.24)	0.80(0.03)	6.29(0.23)	5.70(0.19)	5.27(0.18)
AuditHet	6.22(0.21)	0.88(0.02)	6.23(0.21)	5.66(0.18)	5.37(0.15)
BiasedHet	6.02(0.16)	0.84(0.01)	6.11(0.15)	5.50(0.14)	5.33(0.13)

Table 1 presents estimated marginal means with cluster-robust standard errors for all outcome variables. It reveals five important results. First, group heterogeneity and rule enforcement do not affect average contributions. Second, biased rule enforcement reduces average rule compliance relative to unbiased enforcement. Third, personal norms reflect descriptive norms, i.e., the contributions of other players. Fourth, the contribution rule induces a strong social norm, even when it is not enforced. Fifth, group heterogeneity reduces trust. We elaborate on these results in the following sections.

3.1 Contributions to the public good

Figure 2 shows mean contributions by round across treatments. As detailed in Table 1, the average contribution exceeds the required minimum contribution in all treatments, consistent with prosocial warm-glow effects coupled with automatic rule-following behaviour. In line with prior evidence from the experimental literature on tax compliance (Alm and Malézieux, 2021), we expected that random audits would increase overall contributions relative to no audits and compared to biased audits. Surprisingly, however, the introduction of audits does not result in higher contribution levels (all $p > .3$, see Table C.2 for additional details). Additionally, in the biased audit treatment, players who faced a high audit probability of 30% contribute only non-significantly more (6.12) than players who face a low audit probability of 10% (5.91, $p = .305$). These findings suggest that institutional parameters, i.e., the existence and fairness of audits, do not affect average public good contributions in groups with strong social norms and complement prior evidence on compliance with minimum contribution requirements absent any enforcement (see, e.g., Galbiati and Vertova, 2014; Dwenger et al., 2016).

Given the absence of treatment effects on average contributions, we explore the distribution of contributions to the public good across treatments (Figure 3). In all treatments a large share of participants follows the contribution rule and allocates exactly five points to the group account. However, the introduction of audits induces a shift in the distribution of contributions at both the lower and the upper ends of the distribution. Specifically, the relative frequency of rule-compliant contributions (i.e., contributions of exactly 5 points) is significantly larger in the audit treatments (Table C.4), while the frequencies of free-riding and full cooperation are lower, although these effects are not consistently significant at conventional levels (see Tables C.5–C.8). Taken together, these results suggest that audits induce exact compliance with rules by deterring free-riding and crowding out decisions that maximise social welfare. Such backfiring effects from increasing the audit probability

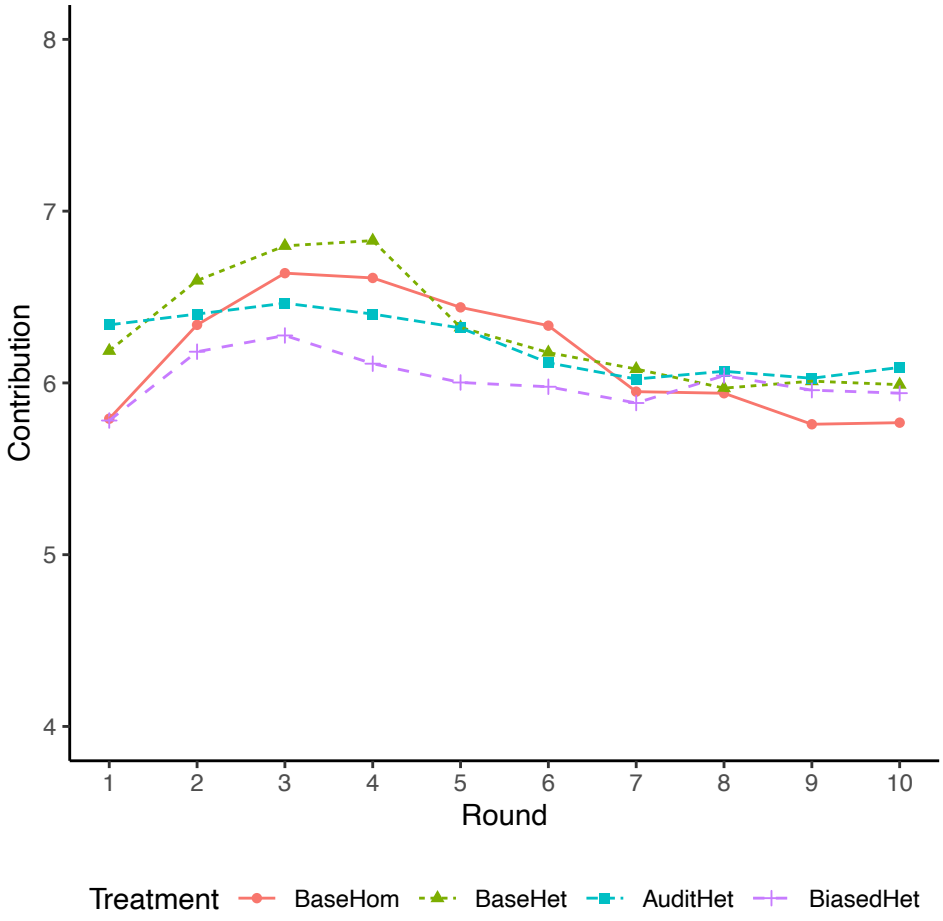

Figure 2: Mean contribution by round.

have been discussed in prior work (Slemrod et al., 2001; Mendoza et al., 2017) and are consistent with a crowding out of intrinsic prosocial motivations by extrinsic (material) incentives (Bénabou and Tirole, 2006).

3.2 Rule compliance

As audits do not affect the level of contributions to the public good, but shift the overall distribution, we next explore the effect of audits on compliance with the contribution rule. We start by analyzing the distribution of compliant contribution decisions (i.e., contributions of five points or more) across treatments. Figure 4 shows the share of compliant decisions across treatments over time.

It reveals two important findings on the effects of audits on heterogeneous groups (see

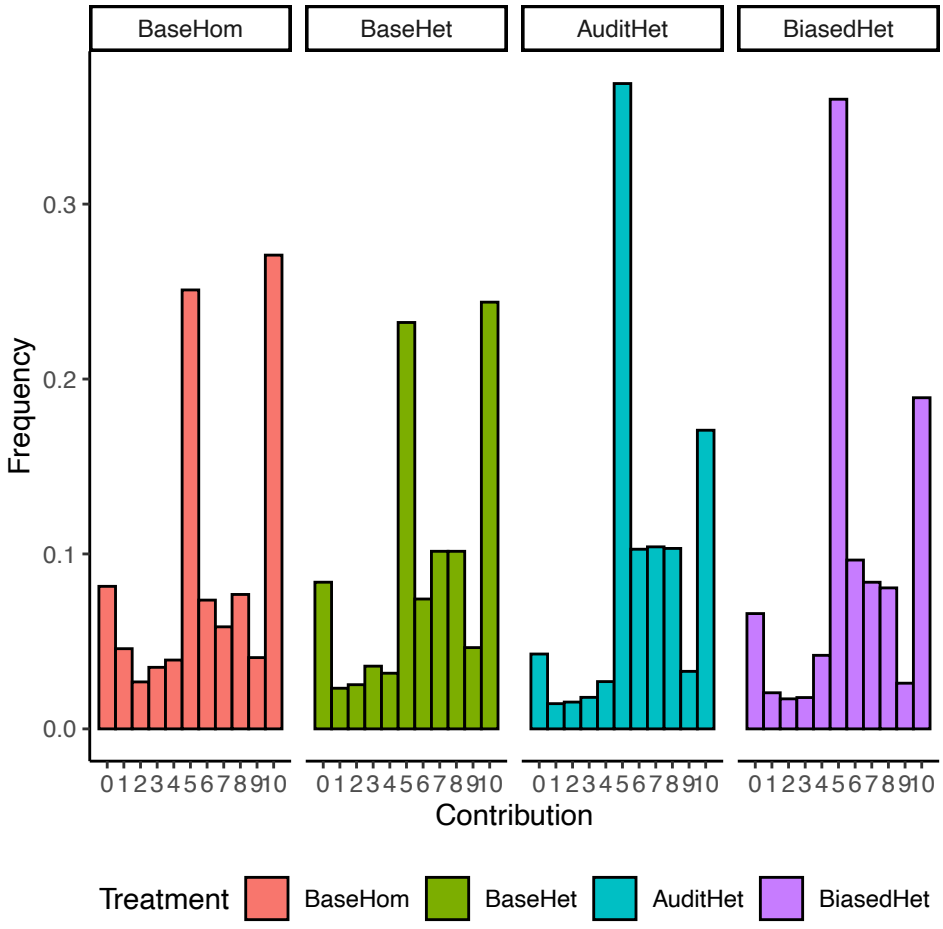

Figure 3: Distribution of contribution decisions by treatment.

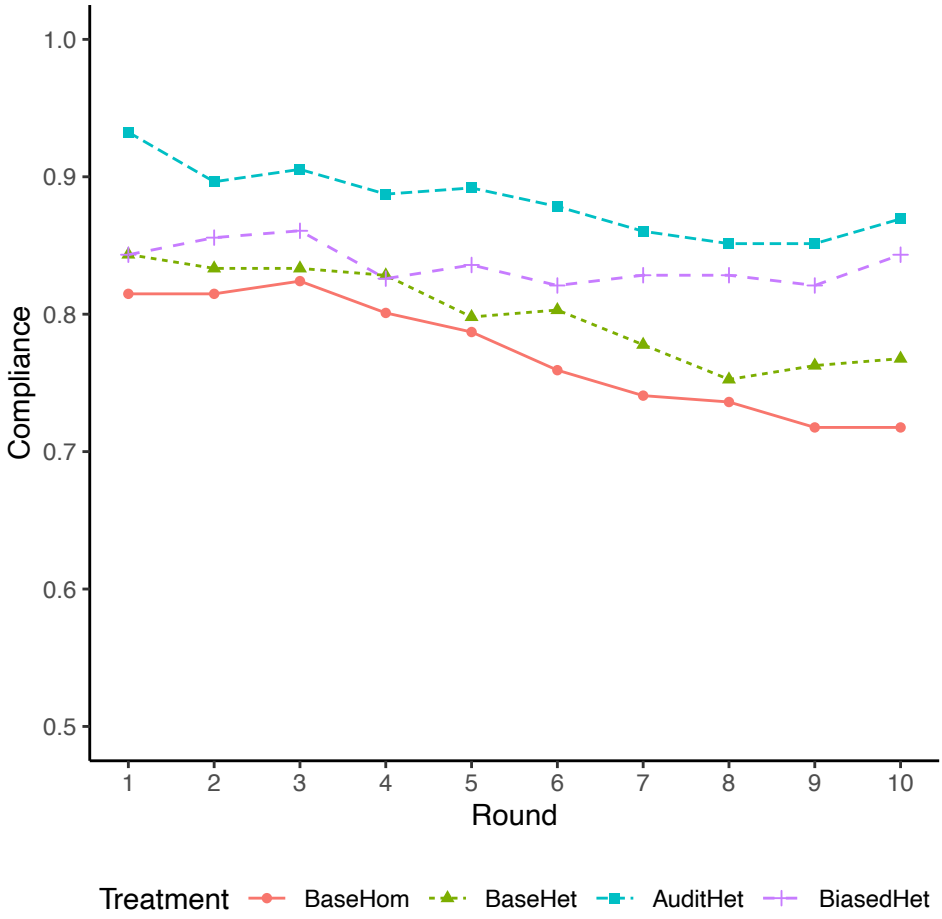

Figure 4: Proportion of compliant players by round.

Tables C.9–C.10). First, random audits (.88) increase the rate of compliance compared to no audits (.80, $p = .012$) and, descriptively, compared to biased audits (.84, $p = .052$). Second, biased audits do not increase compliance compared to no audits ($p = .237$). Additionally, the individual audit probability has no effect: players in the biased audit treatment who faced a high audit probability were not more compliant (.86) than players who faced a low audit probability (.82, $p = .125$).

In addition to the treatment effects on compliance, we also observe significant variation within treatments, which might reflect normative influences within groups. We therefore explore the effects of peer behaviour on rule compliance. In particular, we regress individual compliance rates in the last round of the public good game on the number of compliant peers in one's group in the first round, accounting for treatment fixed effects.

This approach allows us to identify whether reporting behaviour is affected by the (initial) compliance levels of the other group members, which is exogenous due to random composition of groups. Our results show that peer behaviour influences compliance (Table 2), in line with prior literature (Gächter et al., 2023). Specifically, one additional compliant group member in round 1 translates into a 3.8 percentage point increase in compliance in round 10 ($B = .04$, $SE = .02$, $p = .027$). This result also holds for the effect of initial compliance on contributions in round 10 (Table 2).

Table 2: Effects of peer compliance in round 1 on contributions and compliance in round 10, norms, and trust (Model 1), controlling for own compliance in round 1 (Model 2).

		Model 1		Model 2	
Outcome	Predictor	Est. (SE)	p	Est. (SE)	p
Contributions	Peer	0.45(0.14)	0.001	0.42(0.14)	0.002
	Self			2.49(0.27)	$< .001$
Compliance	Peer	0.04(0.02)	0.027	0.04(0.02)	0.058
	Self			0.35(0.04)	$< .001$
Personal Norm	Peer	0.22(0.11)	0.047	0.20(0.11)	0.058
	Self			2.26(0.21)	$< .001$
Social Norm	Peer	0.50(0.09)	$< .001$	0.49(0.09)	$< .001$
	Self			1.17(0.17)	$< .001$
Trust	Peer	0.28(0.08)	0.001	0.27(0.08)	0.001
	Self			0.99(0.23)	$< .001$

Finally, we explore the effects of audits on post-audit rule compliance (Kasper and Alm, 2022a,b). In line with recent work (Kasper and Rablen, 2023), we find that experiencing an audit reduces compliance in the subsequent round. However, this decline in post-audit compliance is conditional on pre-audit compliance levels (Table 3, Figure 5). For non-compliant individuals, experiencing an audit reduces the probability of complying in the next round by 23 percentage points, or about half of their predicted rate of compliance if not audited. In contrast, the decline in post-audit rule compliance is statistically insignificant at conventional levels for compliant players, who reduce their compliance by 6.6 percentage points in the round after an audit. Similar results hold for contributions (Table C.22), which audits reduce by 1.29 points (for non-compliant players), respectively 0.35 points (for compliant players).

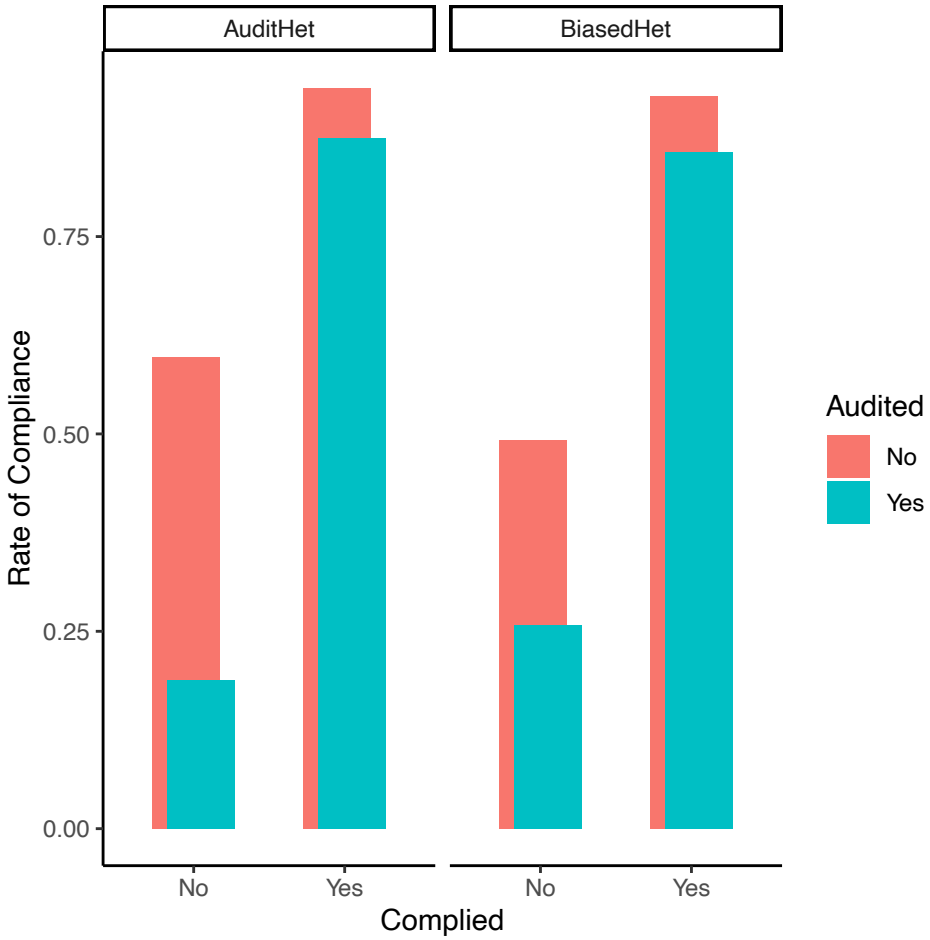

Figure 5: Effects of audits on compliance, by treatment and prior compliance.

Table 3: Effects of audits on post-audit compliance. Models with cluster-robust standard errors and treatment, round, and player fixed effects.

	Model 1		Model 2		Model 3	
Predictor	Est.	p	Est.	p	Est.	p
Audited	-0.09(0.01)	$< .001$	-0.09(0.01)	$< .001$	-0.23(0.05)	$< .001$
Complied			-0.09(0.03)	0.002	-0.12(0.03)	$< .001$
Audited:Complied					0.16(0.05)	0.001

3.3 Personal and social norms

Next, we investigate the effects of institutional fairness on personal and social norms. We expected that random audits would induce prosocial personal normative beliefs and normative expectations, whereas biased audits might undermine such norms. However, Table 1 indicates high levels of personal normative beliefs and normative expectations across all treatments. Consequently, our preregistered regression analyses confirm that the effect of the experimental treatments on personal and social norms is non-significant at conventional levels (all $p > .1$, Wald tests, for details see Tables C.13–C.16).

We also expected that biased audits would lead to more polarised norms. Specifically, we elicited a distribution of normative expectations by asking each participant to indicate their belief about the personal norms of ten randomly selected participants. For each participant, we computed the sample standard deviation across these ten values as an indicator of polarisation. We expected that biased audits ($\sigma = 1.84$) would polarise norms (i.e., lead to a higher standard deviation) compared to random audits ($\sigma = 1.72$, $p = .221$), but this was not the case. However, further exploratory analyses reveal that norms were less dispersed in both audit treatments than in the BaseHet treatment ($\sigma = 2.12$, $p_{AuditHet} < .001$, $p_{BiasedHet} = .011$). In contrast, BaseHet did not differ from the homogeneous baseline treatment (BaseHom) ($\sigma = 2.27$, $p = .268$). Figure 6 displays the average distribution of normative expectations by treatment. It shows that audits reduce the polarisation of social norms: the density of normative expectations is higher around the contribution rule (i.e., 5 points of one's endowment), and lower at both extremes of the distribution in treatments with audits.

Again, we observed significant heterogeneity in personal normative beliefs and normative expectations within treatments. To explore the effect of peer behavior on norms in more detail, we regress personal and social norms on the compliance of the other players

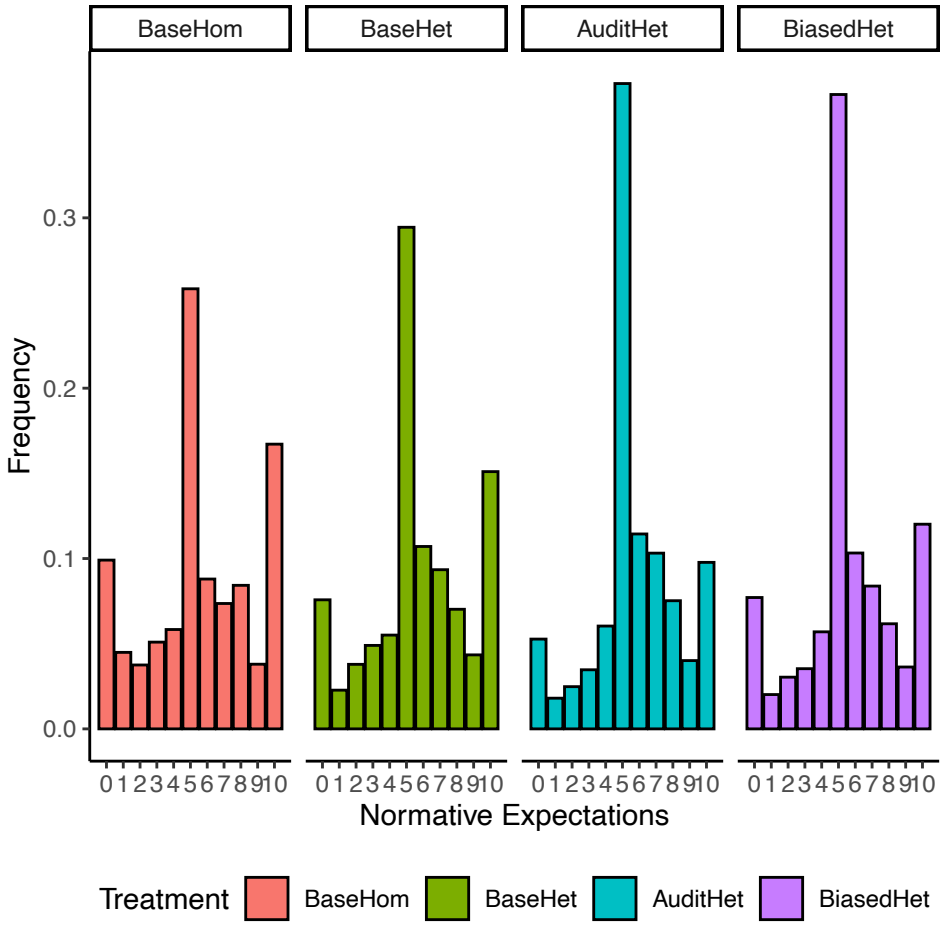

Figure 6: Average distribution of normative expectations by treatment.

in one's group in the first round of the game (Table 2). In line with prior findings (e.g., Gächter et al., 2023), our results suggest that higher initial compliance among peers increases personal ($B = 0.22$, $SE = 0.11$, $p = .047$) and social norms ($B = 0.50$, $SE = 0.09$, $p < .001$). In particular, our regression results show that one additional compliant group member in round 1 increases personal beliefs about the appropriate contribution to the public good (personal norms) by 0.22 points. Similarly, one additional compliant group member in round 1 increases normative expectations about the contributions of other players (social norms) by 0.50 points. These results suggest that descriptive norms (i.e., the behaviour of other group members) do not only inform social but also personal norms.

3.4 Trust

Finally, we analyse the effects of institutional fairness on trust. Figure 7 shows trust towards a randomly selected group member (defined as the mean amount sent in a trust game) across treatments. It illustrates three important findings. First, average trust is higher in the homogeneous treatment than in the heterogeneous treatments, suggesting that the introduction of arbitrary social heterogeneity undermines trust. Second, it provides clear indication of ingroup favouritism: In all treatments with heterogeneous groups, the players exhibit higher levels of trust towards ingroup members (i.e., players of the same colour) than towards outgroup members (i.e., players of the other colour). Third, biased audits did not undermine trust, nor did they increase in-group favouritism.

Our regression results confirm these results (for details, see Tables C.19–C.20). As predicted, trust is higher in the homogeneous treatment (average number of points sent in the trust game, 6.75) than in the heterogeneous baseline treatment (5.27, $p < .001$), the random audit treatment (5.37, $p < .001$), and the biased audit treatment (5.33, $p < .001$). However, contrary to our expectations, trust does not differ across the treatments with heterogeneous groups (all $p > .1$). In these treatments, players exhibit higher levels of trust towards other players from their subgroup (5.64) than towards players of the other subgroup (5.01, $p < .001$), indicating in-group favouritism in trust (for details, see Table C.23). Contrary to our expectations, however, biased audits ($B = -.56$) did not increase the degree of in-group favouritism relative to random audits ($B = -.79$, $p_{interaction} = .173$). We also explore whether disadvantaged players ($B = -0.65$) in the BiasedHet treatment exhibited greater in-group favouritism than advantaged players ($B = -.48$), but the interaction is non-significant ($p = .371$).

We conclude our analysis by exploring the effect of peer behaviour on trust. To this

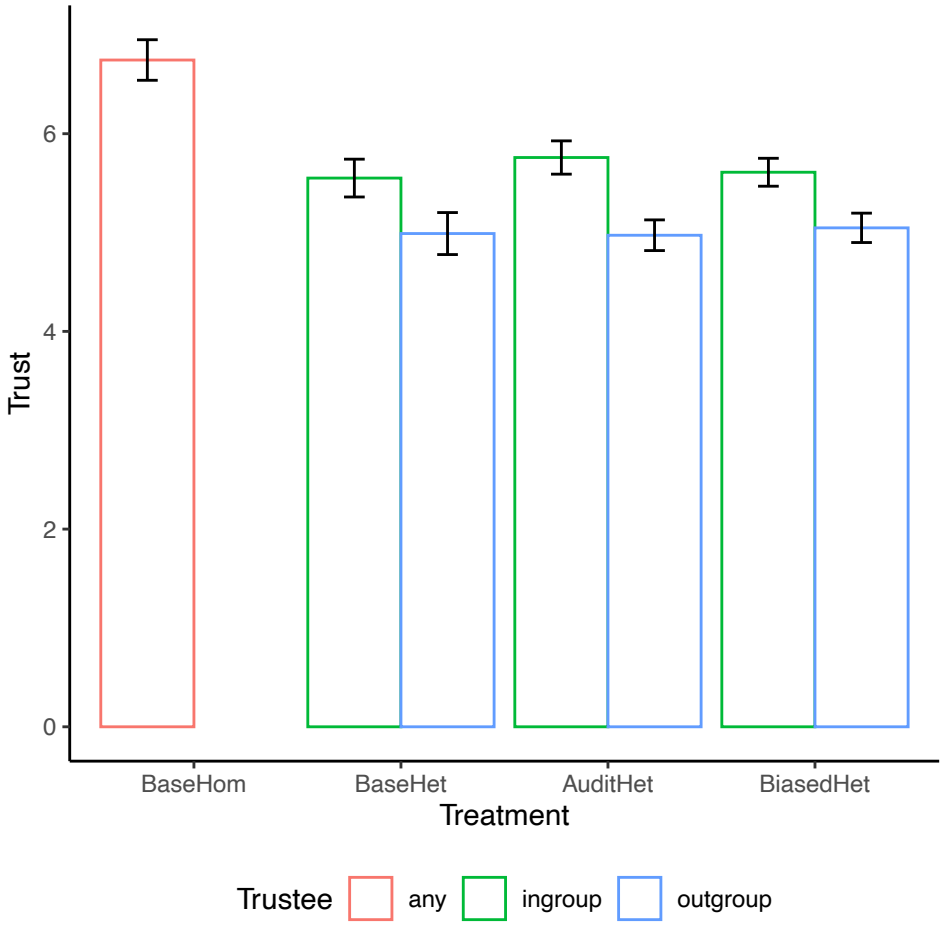

Figure 7: Mean trust by treatment.

end, we regress trust on the compliance of the other players in one's group in the first round of the public goods game. We find that peer behaviour has a strong spillover effect on trust ($B = .28$, $p = .001$, see Table 2), consistent with prior findings in the literature (e.g., Peysakhovich and Rand, 2016). Specifically, one additional compliant group member in round one is associated with an additional 0.28 points transferred to the receiver.

4 Concluding discussion

Human behaviour is guided by formal and informal rules. While laws are typically enforced with the threat of formal legal sanctions, the law also has an expressive function by shaping and communicating informal social norms (Benabou and Tirole, 2011; Lane et al., 2023; Sunstein, 1996). A large body of research investigates the effects of rule enforcement and social norms on social behaviour (Balliet et al., 2011; Bicchieri, 2006; Engl et al., 2021; Fehr and Gächter, 2000; Gächter et al., 2023; Kimbrough and Vostroknutov, 2016; Masclet et al., 2003; Ostrom, 1990; Peysakhovich and Rand, 2016). However, this literature has largely focused on unbiased enforcement of rules in homogeneous populations. In reality, many populations are heterogeneous, and rules are not always applied and enforced in an unbiased way across different social groups.

In this study, we examine experimentally the effects of institutional fairness on rule compliance, personal and social norms, as well as trust in heterogeneous groups playing a public goods game. To isolate the pure effect of heterogeneity, we randomly assign players to visible, but payoff-irrelevant subgroups. We then introduce an institution which punishes rule violations with non-deterrent sanctions. By establishing institutional unfairness, i.e., by overtly biasing the audit probability towards one of the subgroups, we then additionally consider whether unfair rule enforcement undermines compliance and trust and polarises social norms.

Our results provide several new perspectives on the interaction between institutions and individuals in social environments. Contrary to our expectations, we do not find that random audits increase contribution rates in the public goods game. Similarly, we find no evidence of a decline of cooperation over time. These results contrast with earlier work on institutional punishment (Balliet et al., 2011). One explanation for these findings is that the contribution rule itself – i.e., the requirement to contribute 50% of the endowment – established a strong norm of cooperation. Against this background, the introduction of rule enforcement induces more exact rule following. Specifically, audits reduce freeriding,

but they also reduce full contributions. Therefore, our study is the first to provide causal evidence on crowding-out effects of audits (Beer et al., 2020).

Our main finding is that biased rule enforcement undermines the ability of institutions to promote rule compliance. Whereas fair institutions, which audit all players with equal probability, increase compliance compared to enforcement, biased institutions fail to do so. However, institutional unfairness does not undermine contributions to the public good: in our experiment, groups with fair and with unfair institutions contribute to the public good at high levels, as do groups without sanctioning institutions. Importantly, we find no evidence of institutional unfairness inducing social polarisation. Even though disadvantaged players face a threefold audit risk compared to advantaged players, we observe no differences in compliance or contributions between these subgroups. In sum, these results suggest that bias may be pernicious not so much because it leads to over- or under-policing of some groups, but because it undermines the perceived legitimacy of the institution among the entire population.

Unexpectedly, biased rule enforcement reduces, rather than increases, norm polarisation. Using a novel approach to eliciting the entire distribution of each player's normative expectations, we find that normative expectations at the end of the public goods game mirror the shape of actual contributions (i.e., empirical expectations). In all treatments, rule compliance is the modal normative expectation. However, both fair and biased audits increase this pattern and reduce norm polarisation (as indexed by the within-person standard deviation in normative expectations). In contrast, group heterogeneity alone does not lead to greater norm polarisation. These results align with recent work on the expressive function of laws (Lane et al., 2023), but add that even unfair application of the law may not undermine this function.

We also study spillovers from institutions into unregulated interactions among individuals. Prior work has shown that effective institutions can have positive spillovers (e.g., Engl et al., 2021), whereas dysfunctional institutions may have negative spillovers (Peysakhovich and Rand, 2016; Spadaro et al., 2023). In contrast, we find no evidence of either positive or negative spillovers. Instead, and in line with the literature on identity and social preferences (Balliet et al., 2014; Chen and Li, 2009), we find that random assignment to artificial subgroups leads to in-group favouritism in trust. More strikingly, social heterogeneity decreases overall levels of trust: across all treatments with heterogeneous groups, players trust in-group members less than players in the treatment with homogeneous groups. This result is even more remarkable given that these treatments

did not differ in cooperation rates in the public goods game itself, and suggests that even arbitrary social heterogeneity might undermine trust (Dinesen et al., 2020).

Finally, we find robust evidence of peer effects on all outcome variables. Specifically, a higher number of compliant peers in the first round of the public goods game translates into higher contributions and compliance levels in round ten (the final round), stronger personal and social norms, as well as higher levels of trust. This is in line with recent evidence on peer effects on cooperation and social norms (Isler and Gächter, 2022; Gächter et al., 2017). A particularly important, and novel, result of our study is that these peer effects persist over time even in the presence of a sanctioning institution which enforces rule compliance.

Because our results suggest that any action leading even to a superficial perception of sub-group membership can weaken trust, and thereby affect economic outcomes, a policy implication of our findings is that preventing the occurrence or reinforcement of group membership effects has not only moral and social consequences, but also economic ones. To this end, the balance between equity and efficiency in the policing of the law may need to be reconsidered. While there is evidence of tangible efficiency gains from implementing predictive modes of auditing as compared to random modes (Persico and Todd, 2006; Perry et al., 2013), we sense that such approaches risk inadvertently initiating or entrenching "us" versus "them" effects. The efficiency gains from predictive approaches must be weighed against these costs. Therefore, a proper appreciation of these costs might augur for greater use of random, or unbiased, targeting in rule enforcement, albeit not the elimination of all predictive approaches, as proposed by some hard-liners (Harcourt, 2007). We hope future research will shed further light on these issues.

References

Alm, J., Malézieux, A., 2021. 40 years of tax evasion games: A meta-analysis. Experimental Economics 24, 699–750.

Atkinson, E., 2023. More than Half of Ethnic Minority Britons Do Not Trust Metropolitan Police. Independent Digital News & Media Ltd., London.

Balafoutas, L., Nikiforakis, N., 2012. Norm enforcement in the city: A natural field experiment. European Economic Review 56, 1773–1785.

Balafoutas, L., Nikiforakis, N., Rockenbach, B., 2014. Direct and indirect punishment among strangers in the field. Proceedings of the National Academy of Sciences 111, 15924–15927.

Balliet, D., Mulder, L.B., Van Lange, P.A., 2011. Reward, punishment, and cooperation: A meta-analysis. Psychological Bulletin 137, 594–615.

Balliet, D., Wu, J., De Dreu, C.K., 2014. Ingroup favoritism in cooperation: A meta-analysis. Psychological Bulletin 140, 1556–1581.

Becker, G.S., 1968. Crime and punishment: An economic approach. Journal of Political Economy 76, 169–217.

Beer, S., Kasper, M., Kirchler, E., Erard, B., 2020. Do audits deter or provoke future noncompliance? evidence on self-employed taxpayers. CESIfo Economic Studies 66, 248–264.

Bénabou, R., Tirole, J., 2006. Incentives and prosocial behavior. American Economic Review 96, 1652–1678.

Benabou, R., Tirole, J., 2011. Laws and norms. NBER Working Paper, No. 17579.

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

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

Bicchieri, C., Xiao, E., 2009. Do the right thing: But only if others do so. Journal of Behavioral Decision Making 22, 191–208.

Bicskei, M., Lankau, M., Bizer, K., 2016. Negative reciprocity and its relation to anger-like emotions in identity-homogeneous and -heterogeneous groups. Journal of Economic Psychology 54, 17–34.

Blair, G., Cooper, J., Coppock, A., Humphreys, M., Sonnet, L., 2022. estimatr: Fast Estimators for Design-Based Inference. URL: https://CRAN.R-project.org/package=estimatr.

Boning, W.C., Hendren, N., Sprung-Keyser, B., Stuart, E., 2023. A welfare analysis of tax audits across the income distribution. Working Paper.

Boosey, L., Isaac, R.M., 2016. Asymmetric network monitoring and punishment in public goods experiments. Journal of Economic Behavior & Organization 132, 26–41.

Boxell, L., Gentzkow, M., Shapiro, J., 2017. Is the internet causing political polarization? Evidence from demographics. NBER Working Paper, No. 23258.

Burton-Chellew, M.N., Guérin, C., 2021. Decoupling cooperation and punishment in humans shows that punishment is not an altruistic trait. Proceedings of the Royal Society B: Biological Sciences 288, 20211611.

Carothers, T., O'Donohue, A., 2019. Democracies Divided. Brookings Institution Press, Washington, D.C.

Casey, L., 2023. Final Report: An Independent Review into the Standards of Behaviour and Internal Culture of the Metropolitan Police Service. Metropolitan Police Service, London.

Cassar, A., d'Adda, G., Grosjean, P., 2014. Institutional quality, culture, and norms of cooperation: Evidence from behavioral field experiments. Journal of Law and Economics 57, 821–863.

Charness, G., Cobo-Reyes, R., Jiménez, N., 2014. Identities, selection, and contributions in a public-goods game. Games and Economic Behavior 87, 322–338.

Chen, Y., Li, S.X., 2009. Group identity and social preferences. American Economic Review 99, 431–457.

Dal Bó, P., Foster, A., Putterman, L., 2010. Institutions and behavior: Experimental evidence on the effects of democracy. American Economic Review 100, 2205–2229.

Di Tella, R., Gálvez, R., Schargrodsky, E., 2021. Does social media cause polarization? Evidence from access to Twitter echo chambers during the 2019 Argentine presidential debate. NBER Working Paper, No. 29458.

Dimant, E., 2022. Distributions matter: Measuring the tightness and looseness of social norms. SSRN Working Paper, No. 4107802.

Dimant, E., in press. Hate trumps love: The impact of political polarization on social preferences. Management Science .

Dimant, E., Galeotti, F., Villeval, M.C., 2023. Information acquisition and social norm formation. Working Paper.

Dinesen, P.T., Schaeffer, M., Sønderskov, K.M., 2020. Ethnic diversity and social trust: A narrative and meta-analytical review. Annual Review of Political Science 23, 441–465.

Douglas, B.D., Ewell, P.J., Brauer, M., 2021. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. PLoS ONE 18, e0279720.

Drouvelis, M., Malaeb, B., Vlassopoulos, M., Wahba, J., 2021. Cooperation in a fragmented society: Experimental evidence on Syrian refugees and natives in Lebanon. Journal of Economic Behavior & Organization 187, 176–191.

Dutch Data Protection Authority, 2020. Belastingdienst/Toeslagen: De Verwerking van de Nationaliteit van Aanvragers van Ainderopvangtoeslag, z2018-22445. Autoriteit Persoonsgegevens, Amsterdam.

Dwenger, N., Kleven, H., Rasul, I., Rincke, J., 2016. Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany. American Economic Journal: Economic Policy 8, 203–232.

Elzayn, H., Smith, E., Hertz, T., Ramesh, A., Fisher, R., Ho, D.E., Goldin, J., 2023. Measuring and mitigating racial disparities in tax audits. Stanford Institute for Economic Policy Research Working Paper, No. 23-02.

Engel, C., 2013. Deterrence by imperfect sanctions – a public good experiment. Max Planck Institute for Research on Collective Goods, Preprint No. 2013/9.

Engl, F., Riedl, A., Weber, R., 2021. Spillover effects of institutions on cooperative behavior, preferences, and beliefs. American Economic Journal: Microeconomics 13, 261–299.

Espín, Antonio, M., Espinosa, M.P., Vázquez-De Francisco, M.J., Brañas-Garza, P., 2023. Natural identities overcome the minimal group paradigm. Working Paper.

Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. American Economic Review 90, 980–994.

Gächter, S., Gerhards, L., Nosenzo, D., 2017. The importance of peers for compliance with norms of fair sharing. European Economic Review 97, 72–86.

Gächter, S., Molleman, L., Nosenzo, D., 2023. When and why people follow rules. Working paper under preparation.

Gächter, S., Schulz, J.F., 2016. Intrinsic honesty and the prevalence of rule violations across societies. Nature 531, 496–499.

Galbiati, R., Vertova, P., 2014. How laws affect behavior: Obligations, incentives and cooperative behavior. International Review of Law and Economics 38, 48–57.

Glover, D., Pallais, A., Parienté, W., 2017. Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. Quarterly Journal of Economics 132, 1219–1260.

Gneezy, U., Rustichini, A., 2000. A fine is a price. Journal of Legal Studies 29, 1–17.

Graf, C., Sunet, B., Wiepking, P., Merz, E.M., 2023. Social norms offer explanation for inconsistent effects of incentives on prosocial behavior. Journal of Economic Behavior & Organization 211, 429–441.

Greenberg, J., 1990. Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts. Journal of Applied Psychology 75, 561–568.

Grosch, K., Rau, H.A., 2020. Procedural unfair wage differentials and their effects on unethical behavior. Economic Inquiry 58, 1689–1706.

Harcourt, B.E., 2007. Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age. University of Chicago Press, Chicago.

Hartmann, F., Slapničar, S., 2012. Pay fairness and intrinsic motivation: The role of pay transparency. International Journal of Human Resource Management 23, 4283–4300.

Hayek, Friedrich, A., 1973. Law, Legislation and Liberty, Vol. 1: Rules and Order. University of Chicago Press, Chicago.

Isler, O., Gächter, S., 2022. Conforming with peers in honesty and cooperation. Journal of Economic Behavior & Organization 195, 75–86.

Kasper, M., Alm, J., 2022a. Audits, audit effectiveness, and post-audit tax compliance. Journal of Economic Behavior & Organization 195, 87–102.

Kasper, M., Alm, J., 2022b. Does the bomb-crater effect really exist? Evidence from the laboratory. FinanzArchiv 78, 87–111.

Kasper, M., Rablen, M.D., 2023. Tax compliance after an audit: Higher or lower? Journal of Economic Behavior & Organization 207, 157–171.

Kiel, P., Eisinger, J., 2018. How the IRS was Gutted. Pro Publica, New York.

Kim, S.E., Rubianty, D., 2011. Perceived fairness of performance appraisals in the federal government: Does it matter? Review of Public Personnel Administration 31, 329–348.

Kimbrough, E.O., Vostroknutov, A., 2016. Norms make preferences social. Journal of the European Economic Association 14, 608–638.

Kingsley, D., 2016. Endowment heterogeneity and peer punishment in a public good experiment: Cooperation and normative conflict. Journal of Behavioral and Experimental Economics 60, 49–61.

Krupka, E.L., Weber, R., 2013. Identifying social norms using coordination games: Why does dictator game sharing vary? Journal of the European Economic Association 11, 495–524.

Lancee, B., Rossel, L., Kasper, M., 2023. When the agency wants too much: Experimental evidence on unfair audits and tax compliance. Working Paper.

Lane, T., Nosenzo, D., Sonderegger, S., 2023. Law and norms: Empirical evidence. American Economic Review 113, 1255–1293.

Levy, R., 2021. Social media, news consumption, and polarization: Evidence from a field experiment. American Economic Review 111, 831–870.

Li, X., Molleman, L., van Dolder, D., 2021. Do descriptive social norms drive peer punishment? Conditional punishment strategies and their impact on cooperation. Evolution and Human Behavior 42, 469–479.

Lindström, B., Jangard, S., Selbing, I., Olsson, A., 2018. The role of a "common is moral" heuristic in the stability and change of moral norms. Journal of Experimental Psychology: General 147, 228–242.

Macpherson, W., 1999. The Stephen Lawrence Inquiry: Report of an Inquiry, Cm 4262-1. Home Office, London.

Markussen, T., Putterman, L., Tyran, J.R., 2014. Self-organization for collective action: An experimental study of voting on sanction regimes. Review of Economic Studies 81, 301–324.

Martinangeli, A.F., Martinsson, P., 2020. We, the rich: Inequality, identity and cooperation. Journal of Economic Behavior & Organization 178, 249–266.

Masclet, D., Noussair, C., Tucker, S., Villeval, M.C., 2003. Monetary and nonmonetary punishment in the voluntary contributions mechanism. American Economic Review 93, 366–380.

McCoy, J., Rahman, T., Somer, M., 2018. Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. American Behavioral Scientist 62, 16–42.

Mendoza, J.P., Wielhouwer, J.L., Kirchler, E., 2017. The backfiring effect of auditing on tax compliance. Journal of Economic Psychology 62, 284–294.

Molho, C., Tybur, J.M., Van Lange, P.A., Balliet, D., 2020. Direct and indirect punishment of norm violations in daily life. Nature Communications 11, 3432.

Neuman, J.H., 2004. Injustice, stress, and aggression in organizations, in: Griffin, R.W., O'Leary-Kelly, A.M. (Eds.), The Dark Side of Organizational Behavior. John Wiley & Sons, New York, pp. 62–102.

Nikiforakis, N., Normann, H.T., Wallace, B., 2010. Asymmetric enforcement of cooperation in a social dilemma. Southern Economic Journal 76, 638–659.

North, D.C., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, Cambridge.

Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., Damer, E., 2022. Data quality of platforms and panels for online behavioral research. Behavior Research Methods 54, 1643–1662.

Perry, W.L., McInnis, B., Price, C.C., Smith, S., Hollywood, J.S., 2013. Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. RAND Corporation, Santa Monica, CA.

Persico, N., Todd, P., 2006. Generalising the hit rates test for racial bias in law enforcement, with an application to vehicle searches in Wichita. Economic Journal 116, F351–F367.

Peysakhovich, A., Rand, D.G., 2016. Habits of virtue: Creating norms of cooperation and defection in the laboratory. Management Science 62, 631–647.

Posner, R.A., 1997. Social norms and the law: An economic approach. American Economic Review 87, 365–369.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project .org/.

Slemrod, J., Blumenthal, M., Christian, C., 2001. Taxpayer response to an increased probability of audit: Evidence from a controlled experiment in Minnesota. Journal of Public Economics 79, 455–483.

Sønderskov, K.M., Dinesen, P.T., 2016. Trusting the state, trusting each other? The effect of institutional trust on social trust. Political Behavior 38, 179–202.

Spadaro, G., Molho, C., Van Prooijen, J.W., Romano, A., Mosso, C.O., Van Lange, P.A., 2023. Corrupt third parties undermine trust and prosocial behaviour between people. Nature Human Behaviour 7, 46–54.

Spadaro, G., Tiddi, I., Columbus, S., Jin, S., Ten Teije, A., CoDaTeam, Balliet, D., 2022. The Cooperation Databank: Machine-readable science accelerates research synthesis. Perspectives on Psychological Science 17, 1472–1489.

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

Sunstein, C.R., 1996. On the expressive function of law. University of Pennsylvania Law Review 144, 2021–2053.

Thielmann, I., Böhm, R., Ott, M., Hilbig, B.E., 2021. Economic games: An introduction and guide for research. Collabra: Psychology 7, 19004.

Tworek, C.M., Cimpian, A., 2016. Why do people tend to infer "ought" from "is"? The role of biases in explanation. Psychological Science 27, 1109–1122.

Tyler, T.R., Blader, S.L., 2000. Cooperation in Groups: Procedural Justice, Social Identity, and Behavioral Engagement. Psychology Press, Oxford.

Tyler, T.R., Huo, Y.J., 2002. Trust in the Law: Encouraging Public Cooperation with the Police and Courts. Russell Sage Foundation, New York.

Tyler, T.R., Lind, E.A., 2001. Retribution and revenge, in: Sanders, J., Hamilton, V. (Eds.), Handbook of Justice Research in Law. Kluwer, New York, pp. 31–63.

Tyler, T.R., Lind, E.A., 2002. Procedural justice, in: Sanders, J., Hamilton, V.L. (Eds.), Handbook of Justice Research in Law. Springer, Boston, MA, pp. 65–92.

Tyran, J.R., Feld, L.P., 2006. Achieving compliance when legal sanctions are non-deterrent. Scandinavian Journal of Economics 108, 135–156.

Welch, M.R., Xu, Y., Bjarnason, T., Petee, T., O'Donnell, P., Magro, P., 2005. "But everybody does it...": The effects of perceptions, moral pressures, and informal sanctions on tax cheating. Sociological Spectrum 25, 21–52.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. Journal of Open Source Software 4, 1686.

Zizzo, D., Oswald, A., 2001. Are people willing to pay to reduce others' incomes? Annals of Economics and Statistics 63-64, 39–65.

Appendix A: Sample Sizes

1,571 participants began the experiment, of whom 1,128 passed all comprehension questions and started the public goods game. $n = 1,038$ participants (i.e., 173 groups) completed all rounds of the public goods game and are included in the final data set. Table A.1 shows effective sample sizes for all outcome variables.

Table A.1: Effective sample sizes for contributions and compliance (public goods game – PGG), personal and social norms, and trust (trust game – TG).

Treatment	Started	Passed	PGG	Norms	TG
BaseHom	311	228	216	216	216
BaseHet	292	222	198	198	196
AuditHet	358	252	222	222	220
BiasedHet	610	426	402	402	400

Table A.2: Mean and median number of comprehension check attempts among participants in the final sample.

	PGG		Audits	
Treatment	Mean	Median	Mean	Median
BaseHom	3.564815	2.0		
BaseHet	4.065657	1.5		
AuditHet	3.545045	1.0	1.963964	1
BiasedHet	4.037313	1.0	1.878110	1

Appendix B: Preregistered Hypothesis Tests

We preregistered a total of 23 hypotheses. Below, we list each hypothesis as stated in the preregistration, declare any deviations from the preregistration, and provide the key statistical test of the hypothesis. The full preregistration is available at https://osf.io/qaedu/?view_only=262ca0dcde3e41ad98778c2bb1141be5.

> H1: Random audits increase contributions to the public good compared to no audits (due to the higher audit probability). (AuditHom + AuditHet > BaseHom + BaseHet)

Due to a coding error, data from the AuditHom treatment were not usable. We therefore only compare AuditHet with BaseHet. The difference is not significant (Wald test, $B = .07$, $SE = .32$, $p = .825$; Kruskal-Wallis test, $\chi^2(1) = .07$, $p = .791$).

> H2: Biased audits increase contributions to the public good compared to no audits (due to the higher audit probability), or decrease contributions compared to no audits (due to lower legitimacy). (BiasedHet $\neq$ BaseHet)

The difference is not significant (Wald test, $B = .28$, $SE = .29$, $p = .332$; Kruskal-Wallis test, $\chi^2(1) = .93$, $p = .352$).

> H3: Biased audits decrease contributions to the public good compared to random audits (due to lower legitimacy). (AuditHet > BiasedHet)

The difference is not significant (Wald test, $B = .21$, $SE = .27$, $p = .432$; Kruskal-Wallis test, $\chi^2(1) = .49$, $p = .482$).

> H4: Under biased audits, a higher individual audit probability increases contributions to the public good (BiasedHetL < BiasedHetH)

The difference is not significant (Wald test, $B = .21$, $SE = .20$, $p = .305$).

> H5: Random audits increase personal normative beliefs compared to no audits. (AuditHom + AuditHet > BaseHom + BaseHet)

Due to a coding error, data from the AuditHom treatment were not usable. We therefore only compare AuditHet with BaseHet. The difference is not significant (Wald test, $B = .06$, $SE = .31$, $p = .838$; Kruskal-Wallis test, $\chi^2(1) = .42$, $p = .517$).

> H6: Biased audits reduce personal normative beliefs compared to no audits. (BaseHet > BiasedHet)

The difference is not significant (Wald test, $B = .18$, $SE = .28$, $p = .361$; Kruskal-Wallis test, $\chi^2(1) = .84$, $p = .361$).

> H7: Biased audits reduce personal normative beliefs compared to random audits. (AuditHet > BiasedHet)

The difference is not significant (Wald test, $B = .12$, $SE = .26$, $p = .653$; Kruskal-Wallis test, $\chi^2(1) = .09$, $p = .762$).

H8: Under biased audits, a higher individual audit probability reduces personal normative beliefs. (BiasedHetL < BiasedHetH)

The difference is not significant (Wald test, $B = .34$, $SE = .22$, $p = .131$).

H9: Random audits increase average normative expectations compared to no audits. (AuditHom + AuditHet > BaseHom + BaseHet)

Due to a coding error, data from the AuditHom treatment were not usable. We therefore only compare AuditHet with BaseHet. The difference is not significant (Wald test, $B = .04$, $SE = .27$, $p = .877$; Kruskal-Wallis test, $\chi^2(1) = .04$, $p = .841$).

H10: Biased audits increase average normative expectations compared to no audits (due to higher rates of cooperation) or decrease average normative expectations (due to lower legitimacy). (BiasedHet ≠ BaseHet)

The difference is not significant (Wald test, $B = .19$, $SE = .24$, $p = .426$; Kruskal-Wallis test, $\chi^2(1) = .78$, $p = .377$).

H11: Biased audits reduce average normative expectations compared to random audits. (AuditHet > BiasedHet)

The difference is not significant (Wald test, $B = .15$, $SE = .23$, $p = .512$; Kruskal-Wallis test, $\chi^2(1) = .40$, $p = .526$).

H12: The individual audit probability does not affect normative expectations (BiasedHetL = BiasedHetH)

The difference is not significant (Wald test, $B = .12$, $SE = .16$, $p = .447$).

H13: Biased audits increase the within-person variance in normative expectations compared to random audits. (Var(AuditHet) < Var(BiasedHet))

For ease of interpretation, we report the within-person standard deviation rather than the variance. The difference is not significant (Wald test, $B = .12$, $SE = .10$, $p = .221$; Kruskal-Wallis test, $\chi^2(1) = 1.22$, $p = .270$).

H14: Audits increase trust (compared to no audits), because audits signal that the institution aims to deter noncompliance and higher compliance levels result in higher levels of trust. (AuditHom > BaseHom)

Due to a coding error, data from the AuditHom treatment were not usable. We therefore compare AuditHet with BaseHet. The difference is not significant (Wald test, $B = -.10$, $SE = .23$, $p = .684$; Kruskal-Wallis test, $\chi^2(1) = .03$, $p = .865$).

H15: Biased audits decrease trust (compared to no audits), because unfair treatment reduces trust relative to fair treatment. (BiasedHet < BaseHet)

The difference is not significant (Wald test, $B = -.06$, $SE = .23$, $p = .796$; Kruskal-Wallis test, $\chi^2(1) = .00$, $p = .976$).

H16: Biased audits decrease trust (compared to random audits), because unfair treatment reduces trust relative to fair treatment. (BiasedHet < AuditHet)

The difference is not significant (Wald test, $B = .04$, $SE = .20$, $p = .853$; Kruskal-Wallis test, $\chi^2(1) = .01$, $p = .914$).

H17: Participants show in-group favouritism: In heterogeneous groups, there is more trust within subgroups (i.e., between pairs of 'red'-'red' and 'blue'-'blue' players) than across subgroups (i.e., between pairs of 'red'-'blue', respectively 'blue'-'red' players).

The difference is significant in the expected direction (Wald test, $B = .62$, $SE = .10$, $p < .001$).

H18: Unfair treatment increases in-group favouritism. (BiasedHet > AuditHet)

The interaction between subgroup (in-group vs. out-group) and treatment was not significant ($B = .22$, $SE = .16$, $p = .173$).

H19: Tag-based heterogeneity reduces overall contributions to the public good. (BaseHom > BaseHet).

The difference is not significant (Wald test, $B = .14$, $SE = .36$, $p = .702$; Kruskal-Wallis test, $\chi^2(1) = .06$, $p = .806$).

H20: Tag-based heterogeneity reduces personal normative beliefs. (BaseHom > BaseHet).

The difference is not significant (Wald test, $B = .07$, $SE = .38$, $p = .864$; Kruskal-Wallis test, $\chi^2(1) = .05$, $p = .815$).

H21: Tag-based heterogeneity reduces average normative expectations. (BaseHom > BaseHet)

The difference is not significant (Wald test, $B = .17$, $SE = .32$, $p = .600$; Kruskal-Wallis test, $\chi^2(1) = .60$, $p = .438$).

H22: Tag-based heterogeneity increases the within-person variance in normative expectations. (Var(BaseHom) < Var(BaseHet))

The difference is not significant (Wald test, $B = -.15$, $SE = .13$, $p = .268$; Kruskal-Wallis test, $\chi^2(1) = 1.94$, $p = .164$).

H23: Tag-based heterogeneity reduces trust. (BaseHom > BaseHet)

The difference is significant in the expected direction (Wald test, $B = -1.47$, $SE = .27$, $p < .001$; Kruskal-Wallis test, $\chi^2(1) = 18.41$, $p < .001$).

Appendix C: Additional Statistical Details

Model results for contributions

Table C.1: Full model results for estimated marginal means of contributions. Standard errors are clustered at the group level.

Treatment	Est.	SE	t	p
BaseHom	6.16	0.27	22.73	$< .001$
BaseHet	6.30	0.24	26.11	$< .001$
AuditHet	6.22	0.21	29.11	$< .001$
BiasedHet	6.02	0.16	37.89	$< .001$

Table C.2: Treatment comparisons for contributions. Wald tests are based on regressions with cluster-robust standard errors; Kruskal-Wallis tests are for group means.

	Wald test					Kruskal-Wallis test		
Contrast	Est.	SE	df	t	p	χ^2	df	p
BaseHet - AuditHet	0.07	0.32	1034	0.22	0.825	0.07	1	0.791
BaseHet - BiasedHet	0.28	0.29	1034	0.97	0.331	0.93	1	0.335
AuditHet - BiasedHet	0.21	0.27	1034	0.79	0.432	0.49	1	0.482

Models results for exact rule compliance

Table C.3: Full model results for estimated marginal means of exact rule compliance (i.e., contributions of exactly ten points. Standard errors are clustered at the group level.

Treatment	Est.	SE	t	p
BaseHet	0.23	0.02	10.44	$< .001$
BaseHom	0.25	0.02	12.82	$< .001$
AuditHet	0.37	0.03	13.67	$< .001$
BiasedHet	0.36	0.02	17.32	$< .001$

Table C.4: Treatment comparisons for exact rule compliance. Wald tests are based on regressions with cluster-robust standard errors; Kruskal-Wallis tests are computed on group means.

	Wald test					Kruskal-Wallis test		
Contrast	Est.	SE	df	t	p	χ^2	df	p
BaseHet - AuditHet	-0.14	0.03	1034	-3.90	$< .001$	12.36	1	$< .001$
BaseHet - BiasedHet	-0.13	0.03	1034	-4.19	$< .001$	12.43	1	$< .001$
AuditHet - BiasedHet	0.01	0.03	1034	0.26	0.792	0.19	1	0.661

Model results for freeriding

Table C.5: Full model results for estimated marginal means of freeriding (i.e., contributions of exactly zero points). Standard errors are clustered at the group level.

Treatment	Est.	SE	t	p
BaseHet	0.08	0.02	4.46	$< .001$
BaseHom	0.08	0.02	4.64	$< .001$
AuditHet	0.04	0.01	4.45	$< .001$
BiasedHet	0.07	0.01	6.09	$< .001$

Table C.6: Treatment comparisons for freeriding. Wald tests are based on regressions with cluster-robust standard errors; Kruskal-Wallis tests are computed on group means.

	Wald test					Kruskal-Wallis test		
Contrast	Est.	SE	df	t	p	χ^2	df	p
BaseHet - AuditHet	0.04	0.02	1034	1.94	0.052	2.25	1	0.134
BaseHet - BiasedHet	0.02	0.02	1034	0.83	0.409	0.86	1	0.354
AuditHet - BiasedHet	-0.02	0.01	1034	-1.60	0.110	0.52	1	0.472

Model results for full cooperation

Table C.7: Full model results for estimated marginal means of full cooperation (i.e., contributions of exactly ten points). Standard errors are clustered at the group level.

Treatment	Est.	SE	t	p
BaseHet	0.24	0.03	7.11	$< .001$
BaseHom	0.27	0.04	7.39	$< .001$
AuditHet	0.17	0.03	5.37	$< .001$
BiasedHet	0.19	0.02	8.05	$< .001$

Table C.8: Treatment comparisons for full cooperation. Wald tests are based on regressions with cluster-robust standard errors; Kruskal-Wallis tests are computed on group means.

	Wald test					Kruskal-Wallis test		
Contrast	Est.	SE	df	t	p	χ^2	df	p
BaseHet - AuditHet	0.07	0.05	1034	1.57	0.118	4.99	1	0.025
BaseHet - BiasedHet	0.05	0.04	1034	1.31	0.189	3.29	1	0.070
AuditHet - BiasedHet	-0.02	0.04	1034	-0.47	0.639	1.00	1	0.317

Model results for compliance

Table C.9: Full model results for estimated marginal means of compliance. Standard errors are clustered at the group level.

Treatment	Est.	SE	t	p
BaseHom	0.77	0.03	26.37	$< .001$
BaseHet	0.80	0.03	29.65	$< .001$
AuditHet	0.88	0.02	47.34	$< .001$
BiasedHet	0.84	0.01	56.94	$< .001$

Table C.10: Treatment comparisons for compliance. Wald tests are based on regressions with cluster-robust standard errors; Kruskal-Wallis tests are computed on group means.

	Wald test					Kruskal-Wallis test		
Contrast	Est.	SE	df	t	p	χ^2	df	p
BaseHet - AuditHet	-0.08	0.03	1034	-2.51	0.012	4.61	1	0.032
BaseHet - BiasedHet	-0.04	0.03	1034	-1.18	0.237	0.82	1	0.365
AuditHet - BiasedHet	0.05	0.02	1034	1.94	0.052	4.29	1	0.038

Model results for compliance, round 10

Table C.11: Full model results for estimated marginal means of compliance in round 10. Standard errors are clustered at the group level.

Treatment	Est.	SE	t	p
BaseHet	0.77	0.04	18.65	$< .001$
BaseHom	0.72	0.04	17.08	$< .001$
AuditHet	0.87	0.03	31.91	$< .001$
BiasedHet	0.84	0.02	42.95	$< .001$

Table C.12: Treatment comparisons for compliance in round 10. Wald tests are based on regressions with cluster-robust standard errors; Kruskal-Wallis tests are computed on group means.

	Wald test					Kruskal-Wallis test		
Contrast	Est.	SE	df	t	p	χ^2	df	p
BaseHet - AuditHet	-0.10	0.05	1034	-2.06	0.040	4.61	1	0.032
BaseHet - BiasedHet	-0.08	0.05	1034	-1.66	0.098	0.82	1	0.365
AuditHet - BiasedHet	0.03	0.03	1034	0.78	0.437	4.29	1	0.038

Model results for personal normative beliefs

Table C.13: Full model results for estimated marginal means of personal norms. Standard errors are clustered at the group level.

Treatment	Est.	SE	t	p
BaseHom	6.23	0.31	20.24	$< .001$
BaseHet	6.29	0.23	27.35	$< .001$
AuditHet	6.23	0.21	30.11	$< .001$
BiasedHet	6.11	0.15	40.45	$< .001$

Table C.14: Treatment comparisons for personal norms. Wald tests are based on regressions with cluster-robust standard errors; Kruskal-Wallis tests are for group means.

	Wald test					Kruskal-Wallis test		
Contrast	Est.	SE	df	t	p	χ^2	df	p
BaseHet - AuditHet	0.06	0.31	1034	0.20	0.838	0.42	1	0.517
BaseHet - BiasedHet	0.18	0.28	1034	0.65	0.517	0.84	1	0.361
AuditHet - BiasedHet	0.12	0.26	1034	0.45	0.653	0.09	1	0.762

Model results for normative expectations

Table C.15: Full model results for estimated marginal means of normative expectations. Standard errors are clustered at the group level.

Treatment	Est.	SE	t	p
BaseHom	5.53	0.26	21.71	$< .001$
BaseHet	5.70	0.20	29.28	$< .001$
AuditHet	5.66	0.18	31.24	$< .001$
BiasedHet	5.50	0.14	38.58	$< .001$

Table C.16: Treatment comparisons for normative expectations. Wald tests are based on regressions with cluster-robust standard errors; Kruskal-Wallis tests are for group means.

	Wald test					Kruskal-Wallis test		
Contrast	Est.	SE	df	t	p	χ^2	df	p
BaseHet - AuditHet	0.04	0.27	1034	0.15	0.877	0.04	1	0.841
BaseHet - BiasedHet	0.19	0.24	1034	0.80	0.426	0.78	1	0.377
AuditHet - BiasedHet	0.15	0.23	1034	0.66	0.512	0.40	1	0.526

Model results for polarisation of social norms

Table C.17: Full model results for estimated marginal means of within-person standard deviations of social norms. Standard errors are clustered at the group level.

Treatment	Est.	SE	t	p
BaseHet	2.12	0.09	24.15	$< .001$
BaseHom	2.27	0.10	23.45	$< .001$
AuditHet	1.72	0.07	24.27	$< .001$
BiasedHet	1.84	0.07	27.94	$< .001$

Table C.18: Treatment comparisons for within-person standard deviations of social norms. Wald tests are based on regressions with cluster-robust standard errors; Kruskal-Wallis tests are computed on group means.

	Wald test					Kruskal-Wallis test		
Contrast	Est.	SE	df	t	p	χ^2	df	p
BaseHet - AuditHet	0.40	0.11	1034	3.54	$< .001$	10.43	1	0.001
BaseHet - BiasedHet	0.28	0.11	1034	2.56	0.011	5.62	1	0.018
BiasedHet - AuditHet	0.12	0.10	1034	1.23	0.221	1.22	1	0.270

Model results for trust

Table C.19: Full model results for estimated marginal means of trust. Standard errors are clustered at the group level.

Treatment	Est.	SE	t	p
BaseHom	6.74	0.20	33.20	$< .001$
BaseHet	5.27	0.18	29.01	$< .001$
AuditHet	5.37	0.15	36.25	$< .001$
BiasedHet	5.33	0.13	39.63	$< .001$

Table C.20: Treatment comparisons for trust. Wald tests are based on regressions with cluster-robust standard errors; Kruskal-Wallis tests are computed on group means.

	Wald test					Kruskal-Wallis test		
Contrast	Est.	SE	df	t	p	χ^2	df	p
BaseHet - AuditHet	-0.10	0.23	1028	-0.41	0.684	0.03	1	0.865
BaseHet - BiasedHet	-0.06	0.23	1028	-0.26	0.796	0.00	1	0.975
AuditHet - BiasedHet	0.04	0.20	1028	0.19	0.853	0.01	1	0.914

Model results for audit probability

Table C.21: Differences between high (HetH) and low (HetL) audit probability subgroups in heterogeneous treatments, on contributions, compliance, personal norms, and normative expectations. Standard errors are clustered at the group level.

	Contrast	Est.	SE	df	t	p
Contributions	HetH - HetL	0.205	0.199	400	1.028	0.305
Compliance	HetH - HetL	0.038	0.024	400	1.553	0.121
Personal Norms	HetH - HetL	0.338	0.224	400	1.512	0.131
Normative Expectations	HetH - HetL	0.118	0.155	400	0.761	0.447

Post-audit effects on contributions

Table C.22: Effects of audits on post-audit contributions. Models with cluster-robust standard errors and treatment, round, and player fixed effects.

	Model 1		Model 2		Model 3	
Predictor	Est.	p	Est.	p	Est.	p
audited	-0.49(0.08)	$< .001$	-0.49(0.08)	$< .001$	-1.29(0.24)	$< .001$
complied			0.00(0.15)	0.987	-0.18(0.15)	0.230
audited:complied					0.93(0.26)	$< .001$

Model results for in-group favouritism

Table C.23: Test of in-group favouritism in trust across treatments all heterogeneous treatments (Model 1), moderation by treatment across treatments AuditHet and BiasedHet (Model 2), and moderation by audit probability in treatment BiasedHet. Model 1 includes treatment and participant fixed effects. In all models, standard errors are clustered at the group level.

	Model 1		Model 2		Model 3	
Var	Est.(SE)	p	Est.(SE)	p	Est.(SE)	p
Outgroup	-0.62(0.1)	$< .001$	-0.79(0.12)	$< .001$	-0.48(0.14)	0.001
BiasedHet			-0.15(0.22)	0.497		
Outgroup:BiasedHet			0.22(0.16)	0.173		
High Prob.					0.60(0.27)	0.030
Outgroup:High Prob.					-0.18(0.19)	0.371
(Intercept)			5.76(0.17)	$< .001$	5.31(0.19)	$< .001$

Model results for tag-based heterogeneity

Table C.24: Effects of tag-based heterogeneity on key outcome variables, comparing BaseHet and BaseHom.

Contrast	Est.	SE	df	t	p
Contributions	0.13	0.34	1538	0.37	0.712
Compliance	0.03	0.04	1034	0.72	0.471
Personal Norms	0.00	0.39	1538	0.01	0.993
Normative Expectations	0.13	0.72	1538	0.18	0.859
SD Normative Expectation	-0.09	0.54	1538	-0.16	0.871
Trust	-1.62	0.27	1528	-5.90	$< .001$