The most recent version of this working paper can be downloaded following this link.

Divided We Act:
The Role of Social Sanctions in a Polarized World

Eugen Dimant[a,b], Michele Gelfand[c], Anna Hochleitner[d], Silvia Sonderegger[e]

[a] *University of Pennsylvania*
[b] *CESifo*
[c] *Stanford University*
[d] *Centre for Applied Research at NHH (SNF), FAIR*
[e] *University of Nottingham, CeDEx*

Abstract

Political polarization is increasingly shaping the social and economic makeup of societies, yet its effects are difficult to measure. Naturally occurring data cannot plausibly separate the influence of holding polarized views from the impact of being surrounded by them. Additionally, social sanctions sustain cooperation by discouraging socially undesirable behavior; however, in polarized societies, the concern is that their deterrent effect may weaken, which is difficult to establish causally. We address these identification challenges through a large, pre-registered, representative U.S.-based experiment (N = 2,400) that exogenously varies decision environments between more and less polarized conditions. This allows us to study how polarization affects behavior, particularly the effectiveness of social sanctions. In our study, participants allocate money between politically opposed recipients in private and public settings. In public, their choices can be punished by partisan Observers across three environments: tight (low variance), loose (high variance), and polarized (U-shaped). We find that polarization reduces the deterrent effect of sanctions. In polarized settings, allocation choices are less equitable because participants (correctly) anticipate that they will be punished regardless of equity. Our findings show that polarization undermines social sanctions, threatening cooperation, equity, and cohesion in divided societies.

Keywords: Equitable Behavior, Polarization, Sanctions, Social Punishment.

JEL Codes: C91, D01

We are grateful to Alexander Cappelen, Jamie Druckman, Ernst Fehr, Catalina Franco, Folco Panizza, and Klaus Schmidt, as well as seminar participants at the University of Bonn, LMU, NYU, Stanford University, and the Vienna University of Economics and Business, and conference participants at the 2026 ASSA, the 2025 CESifo Behavioral Economics Conference, and the 2025 MIT polarization workshop for their excellent feedback. Financial support from Stanford University and the German Research Foundation (DFG) under Germany's Excellence Strategy – EXC 2126/1– 390838866 is gratefully acknowledged. This project was preregistered for both data collection waves (before and after the 2024 U.S. Presidential Election) at AsPredicted #196812 and AsPredicted #208802 and received IRB from the University of Pennsylvania (No. 850822).

Email addresses: edimant@sas.upenn.edu (Eugen Dimant), gelfand1@stanford.edu (Michele Gelfand), anna.hochleitner@snf.no (Anna Hochleitner), silvia.sonderegger@nottingham.ac.uk (Silvia Sonderegger).

This version: January 9, 2026

1 Introduction

At least since the classic contributions of thinkers such as Émile Durkheim and other foundational social theorists, social sanctions have been recognized as a cornerstone of social order (Durkheim, 1893).[1] In well-functioning societies, social sanctions foster norms of cooperation, equity, fairness, and deter socially harmful behavior. Yet, as modern societies become increasingly shaped by political polarization, this foundational mechanism may come under strain. This raises a central concern: when the moral fabric of society becomes politically fractured, can social sanctions still effectively promote socially desirable behaviors? Or do they instead begin to reflect and reinforce partisan divisions? We investigate this question by examining both theoretically and empirically how polarization affects the role of social sanctions in promoting equitable behavior.

The answer is *a priori* ambiguous. Consider an individual tasked with deciding how to allocate resources between two recipients who are identical in every respect except for their opposing political views. Clearly, favoring one recipient attracts social punishment from the other recipient's supporters, and vice versa. What is the effect of increased polarization in this context? Two possibilities emerge. First, partisans who are more polarized may react more strongly to any allocation that disadvantages their preferred recipient, thereby increasing the decider's incentives to divide resources equally. In this scenario, greater polarization *strengthens* the role of social sanctions in discouraging inequity. Second, as polarization intensifies, even an equal division of resources might provoke social backlash. Partisans on both sides might disregard the equitable norm and punish allocations that do not disproportionately favor their preferred recipient. In this case, greater polarization *weakens* the effectiveness of social sanctions in deterring inequitable behavior. The deciders anticipate that they will face social punishment even if they allocate the resources equally, and this reduces their incentives to act equitably.

We capture this ambiguity through a theoretical framework in which, similar to Tirole (2023), people seek to select actions that align with their private preferences while simultaneously attempting to minimize expected social sanctions. Individuals form expectations about the material punishment they might incur from different actions, captured by an *expected punishment schedule*. Crucially, the nature of the social environment (e.g., degree of polarization) determines the shape of the expected punishment schedule, which in turn influences allocation decisions. For instance, compared to a punishment schedule that increases sharply in any deviation from the equitable allocation, a flatter punishment schedule makes individuals less inclined to act equitably and more likely to indulge their partisan biases. Intuitively, that's because a flatter punishment schedule implies that the equitable choice is punished similarly to inequitable allocations. A central takeaway from our theoretical analysis is that the behavioral effects of operating in a more/less polarized environment depend on how polarization affects the shape of the expected punishment schedule.

We investigate our research question empirically within a context that is particularly well-suited to studying polarization: political identity in the United States. Political polarization

[1] See also Boyd and Richerson (1992); Fehr and Gächter (2000) for classic contributions on the role of social punishment in promoting cooperation, as well as the vast literature thereafter (for overviews see, e.g., Bicchieri et al., 2021; Balafoutas et al., 2024; Gelfand et al., 2024).

is increasingly shaping societies, yet its effects are difficult to measure. Naturally occurring data suffer from an identification problem because they cannot separate the influence of holding polarized views from the impact of being surrounded by them. We overcome this problem by running an experiment that varies the participants' environment (more/less polarized) exogenously. This allows us to study how polarization affects behavior. Our representative online experiment (n=2400) recruited U.S. citizens through quota sampling based on self-reported political affiliation. Data was collected in two waves – before and after the 2024 U.S. presidential election. The election can be seen as an external shock that potentially intensifies partisan sentiments, which enables us to causally identify its impact on behavior.

We consider a setup where Deciders have to allocate a fixed sum of money between two recipients with opposing political identities (one identified as a Trump supporter and the other as a Trump opponent). First, Deciders make allocation decisions privately, thereby revealing their intrinsic preferences free from external social pressures. In a second stage, Deciders learn that their choices will be evaluated by an Observer, who can punish by deducting points from their monetary payoff. This monetary punishment allows the Observer to tangibly express their disapproval of the Decider's choice. The key feature of our large-scale experiment is that we exogenously vary the environment (more/less polarized) in which subjects operate. Deciders learn that their Observer will be randomly drawn from a known distribution of Observers who differ with respect to their personal values, i.e., what they believe to be the appropriate allocation choice. We consider three between-subject experimental treatments: (i) *tight* (hump-shaped, low variance), (ii) *loose* (hump-shaped, high variance), and (iii) *polarized* (U-shaped). In all treatments, the mean appropriate allocation is the equitable one. However, the distribution of what the Observers consider appropriate reflects progressively increasing polarization.

Our data suggest that, as expected, social punishment plays a disciplining role. When the Deciders' decisions are unobserved, political identity strongly shapes their allocation choices, but when an Observer is introduced who may punish the Decider, allocation choices become more equitable. However, and most importantly for our purposes, the magnitude of this disciplining role critically depends on the nature of the environment, which changes across treatments. The Deciders' allocation choices are most equitable in the tight treatment and least equitable in the polarized treatment. Our findings, therefore, illustrate that all else being equal, a more polarized environment — or the perception of a more polarized environment — weakens the disciplining effect of social punishment. To investigate the mechanism behind these results, we follow our theoretical framework and elicit the Deciders' expected punishment schedules, comparing them across treatments. Expected punishment schedules can, in principle, assume many different shapes. We show that, in all our treatments, the majority of Deciders expect punishment to be minimized at the equitable allocation. However, a central finding is that Deciders expect the punishment schedule to be *flatter* in more polarized environments. That is, while extreme allocations are expected to trigger similar punishment in all three treatments, Deciders believe that choosing the equitable allocation will attract a relatively harsher punishment in more polarized environments. This aligns with the finding that in polarized environments, the disciplining effect of Observers is weakest and supports the notion that differences in the expected punishment schedule across environments drive our treatment effects.

A further advantage of our analysis is that we also elicit the Observers' punishment functions, characterizing how different allocation choices are actually punished. This allows us to verify whether the Deciders' beliefs about the shape of the punishment schedule in different environments are in fact correct. We find that *actual* punishment typically grows proportionately to the distance between the observed allocation and the Observer's preferred allocation. Polarized Observers punish the equitable allocation approximately half as much as an allocation that leaves nothing to their preferred recipient. As a result, in polarized environments, the expected punishment schedule becomes approximately flat: Deciders face the same expected punishment irrespective of their choice. We find that, while Deciders correctly expect the punishment schedule to be flatter in more polarized environments, they underestimate the strength of the effect. This suggests that if the Deciders' expectations fully reflected actual punishment behavior, the treatment effect would be stronger. Comparing results before and after the 2024 U.S. presidential election, we find that in the pre-election period, Deciders in the unobserved scenario choose more unequal allocations, suggesting that animosity between the two factions may have been stronger before the election. Importantly, however, our key results about the impact of polarization on the effectiveness of social sanctions apply across both waves. This confirms that the mechanism we identify is robust and not confined to the particular circumstances of the moment.

Overall, our results have wide-ranging implications. They indicate that, if individuals become intrinsically more polarized over time, this has two distinct effects on behavior. One is the direct effect of more polarized preferences. The other is the *indirect* effect we identify: In polarized societies, the role of social sanctions in discouraging inequitable and biased behavior is reduced, and therefore individuals become more inclined to indulge their partisan preferences. The indirect effect thus acts as a multiplier for the direct effect. Our findings further suggest that by presenting society as polarized, both traditional and social media may inadvertently encourage partisan behavior, which in turn could deepen existing divisions.

Our paper contributes to the literature in several ways, starting from the literature on polarization. The current literature has generally focused on documenting polarization (e.g., Iyengar and Westwood, 2015a; Levy, 2021; Braghieri et al., 2024; Holliday et al., 2024; Kish Bar-On et al., 2024; Musolff and Zimmermann, 2025), identifying the causes of polarization (e.g. Iyengar et al., 2019; Boxell et al., 2022) or investigating the consequences of one's own partisanship on own behavior (Lees and Cikara, 2020; Broockman et al., 2023; Bauer et al., 2024; Dimant, 2024; Druckman et al., 2024; Voelkel et al., 2024; DellaVigna and Kim, 2025).

In this paper, we offer a different perspective, by studying how polarization *shapes* societal functions. Our approach views polarization as a societal condition that translates into behavior and, through that, feeds back into societal functions at a macro level. One important way in which polarization affects societal functioning is by reshaping how individuals interpret social behavior and the norms that govern it. Work on partisan sorting shows that ideological, partisan, and social identities have become increasingly aligned (Levendusky, 2009; Mason, 2015; Della-Posta, 2020), narrowing the scope for shared standards of evaluation. At the same time, research on endogenous identity formation demonstrates that group attachments respond to conflict and incentives for social approval (Atkin et al., 2021; Jia and Persson, 2021). Building on these insights, we argue that polarization changes both the meaning and effectiveness of social sanctions:

when observers evaluate behavior through partisan lenses, sanctions lose their normative clarity, reducing their deterrent effect and reinforcing partisan boundaries.

Methodologically, our approach extends existing research on descriptive norms and strategic behavior by thoroughly analyzing the full distribution of normative cues rather than relying solely on mean-based measures. Recent studies indicate that individuals' responses depend to a large degree on the variance and shape of the behavioral and normative distributions they encounter (d'Adda et al., 2020; Dimant, 2023; Dimant et al., 2025). By experimentally manipulating these distributional characteristics and observing the resulting shifts in behavior, our study deepens the understanding of how social environments influence individual decisions. Crucially, our experimental approach allows us to isolate and decompose the mechanisms by which polarization – and the associated normative uncertainty – directly influence both decision-making and actual behavior toward both partisan in- and out-group members.

Finally, by incorporating a temporal dimension – collecting data both before and after the 2024 U.S. presidential election – our study addresses a key gap in the literature on how exogenous political shocks influence normative decision-making. Existing research has primarily focused on perceptions or time trends of polarization (e.g., Holliday et al., 2024; Musolff and Zimmermann, 2025), or examines how individuals' own partisan attitudes shape their perceptions of – and behavior toward – rival groups (Lees and Cikara, 2020; Broockman et al., 2023; Dimant, 2024; Druckman et al., 2024; Holliday et al., 2024; Voelkel et al., 2024). By collecting pre- and post-election data, we leverage this exogenous political shock to not only measure perceptions and normative beliefs but to explicitly investigate the dynamic effects of polarization on identity-driven preferences, fairness norms, and strategic interactions across partisan divides in response to this unpredictable event. This gives us an additional exogenous layer of exploration.

We thus treat polarization as a socially perceived condition, one that individuals believe to be more or less pronounced, rather than solely a personal or group-level trait. In doing so, we highlight how individuals' micro-level partisan preferences are embedded within a broader macro-level environment, wherein the perceived degree of societal polarization or unfairness exerts normative pressure on behavior (see, e.g., Baxter-King et al., 2022; Dimant and Kimbrough, 2024; Dimant et al., 2024; Panizza et al., 2026). As a result, even small changes in how polarized people perceive their surroundings, especially during major political events, can alter fairness norms and punishments in ways that amplify or diminish cooperation across partisan lines.

The remainder of this paper is structured as follows. In Section 2 we outline our theoretical framework and hypotheses. Section 3 describes the experimental design. Section 4 presents the empirical results using our treatment variation, while Section 5 discusses changes driven by the presidential elections. Section 6 concludes.

2 Theoretical framework

We consider a simple theoretical framework where, similar to Tirole (2023), people might disagree as to what is "right" or "wrong", and each individual seeks to align their actions with their personal values while at the same time taking into account the social sanctions that their behavior may attract. More concretely, in our setup an individual (the "Decider") must allocate a resource of size 100 between two other individuals, A and B, who have opposed political preferences. We

first derive the Decider's choice when unobserved, and then consider what happens when the choice is observed (and potentially punished).

Choice is unobserved. Denoting a as the amount allocated to A, Decider i selects $a_i \in [0, 100]$ to maximize

$$-\frac{1}{2}(a_i - \theta_i)^2 \tag{1}$$

where θ_i is i's bliss point, capturing factors such as i's personal values, closeness to A or B's political preferences, emotions, identity etc. Maximization of (1) yields i's action choice:

$$a_i^u = \theta_i \tag{2}$$

where u stands for "unobserved."

Choice is observed. Suppose now that the Decider's allocation choice is observed by an Observer who can subsequently punish the Decider. The Decider knows the distribution from which the Observer's *personal value* (the allocation that the Observer believes to be appropriate) is drawn. Let $p(a \mid v_j)$ denote how much an Observer with personal value v_j punishes action a, and let $\pi(v_j)$ be the probability that the Observer has value v_j – this depends on the distribution of Observer values, which in turn reflects different environments (more/less polarized). Decider i selects a_i to maximize

$$-\frac{1}{2}(a_i - \theta_i)^2 - P(a) \tag{3}$$

where $P_i(a) \equiv \sum_{v_j} \pi(v_j)\, p(a \mid v_j)$ is expected punishment. The first order condition of (3) with respect to a_i is

$$-a_i + \theta_i - P'(a_i) = 0 \tag{4}$$

This is satisfied when $a_i = a_i^o$ given by

$$a_i^o = \theta_i - P'(a_i^o). \tag{5}$$

The sufficient condition for (5) to characterize a maximum is that $-1 - P''(a) < 0$ for all a, which we henceforth assume to be the case. To compare i's choice when observed to the choice when unobserved, we evaluate expression (4) at $a_i = a_i^u$. Substituting for $a_i^u = \theta_i$, we obtain

$$-P'(\theta_i).$$

This shows how the effect of being observed and potentially punished on i's allocation choice depends on the *slope* of the expected punishment schedule when i selects the allocation corresponding to her bliss point. If the expected punishment schedule is upward sloping at that point, then when observed i will decrease her allocation choice. If, on the other hand, the expected punishment schedule is downward sloping when evaluated at i's bliss point, then when observed i will increase her choice.

Next, we focus on the implications of different expected punishment schedules when allocation choices are observed. Consider two expected punishment schedules, P_0 and P_1, and denote i's allocation choice when $P = P_0$ (resp., P_1) as a_0 (resp. a_1). To compare these two allocation

choices, we take the first order condition (4) under $P_i = P_0$ and evaluate it at $a_i = a_1$, obtaining

$$P_1'(a_1) - P_0'(a_1). \tag{6}$$

If (6) is > 0 then $a_1 < a_0$, and vice versa if (6) is < 0. This shows that the comparison between the allocation choices made under different expected punishment schedules depends on the comparison between the *slopes* of the two expected punishment schedules. For example, if $P_0'(a) < P_1'(a)$ for all a, then $a_0 > a_1$. In sum, our model predicts that,

Proposition 0 *The distribution of Observer values affects the Deciders' allocation choices through its effect on expected punishment schedules. This, in turn, determines how the Deciders' choices change when they are observed, and how different allocation choices may arise in different environments (more/less polarized).*

We now introduce a definition that identifies two types of distributions which will be especially relevant in our context.

Definition A function $P(a)$ is said to be U-shaped (resp., hump-shaped) if it has an interior minimum and no interior maximum (resp., an interior maximum and no interior minimum).

For conciseness, in what follows we focus on the case where U- or hump-shaped expected punishment schedules have a minimum/maximum at 50, in which case a *more extreme* allocation is also *more unequal*. As we further discuss in our results section, this is actually the more relevant case empirically.[2]

Proposition 1 Effect of being observed and potentially punished: *Suppose that, when Deciders are observed, expected punishment $P(a)$ is U-shaped (resp. hump-shaped). Then, compared to the case where i's allocation choice is unobserved, i's allocation is weakly less (resp., more) extreme.*

Suppose that i's bliss point differs from 50, i.e. ideally i would like to allocate more to either A or B. An implication of proposition 1 is that, if $P(a)$ is U-shaped, when observed i will choose an allocation that is less extreme compared to the case where she is unobserved. Consider now two U-shaped expected punishment schedules, $P_0(a)$ and $P_1(a)$. Suppose that $P_0(a)$ is steeper than $P_1(a)$, in the sense that $P_0'(a) <$ (resp., $>$) $P_1'(a)$ for all $a <$ (resp., $>$) 50.

Proposition 2 Steepness of expected punishment when observed: *Suppose that Deciders are observed and potentially punished. Compared to the case where expected punishment is $P(a) = P_1(a)$, if $P(a) = P_0(a)$ the following applies: (a) i's allocation choice is less extreme and (b) i's personal preferences play a smaller role in i's allocation decision.*

Corollary 1 *The results in proposition 2 also apply if $P_0(a)$ is U-shaped while $P_1(a)$ is hump-shaped.*

[2] In our preregistration, we adopted the (more general) following definition. Suppose that $P(a)$ is U- or hump-shaped, and let $\widetilde{a}$ denote the action where $P'(a)$ changes sign. We say that an allocation a_1 is more extreme than an allocation a_0 if one of the following applies. Either (i) $a_0 < \widetilde{a}$ and $a_1 < a_0$ or (ii) $a_0 > \widetilde{a}$ and $a_1 > a_0$. In the general case, proposition 2 applies to two expected punishment schedules with the same $\tilde{a}$.

To summarize, our analysis indicates that, when Deciders are observed, a hump-shaped expected punishment schedule generates more extreme allocation choices than a U-shaped one, and a flatter U-shaped schedule generates more extreme allocations than a steeper one. This is outlined in Figure 1 below. Furthermore, when allocation decisions are more extreme, the Decider's own preferences (personal values, political affiliation/closeness) play a larger role in determining allocation choices.

Which expected punishment schedule? The analysis above shows that, depending on the shape of the expected punishment schedule, Deciders may choose more or less extreme allocations. This brings us to the next question: How does the distribution of Observer values affect the shape of the expected punishment schedule? Here, the theoretical predictions are ambiguous. To fix ideas, consider the following setup, where w.l.o.g. we have normalized the marginal cost of punishment to one. When the Decider's allocation decision is a, Observer j chooses punishment $p_j \geq 0$ to maximize

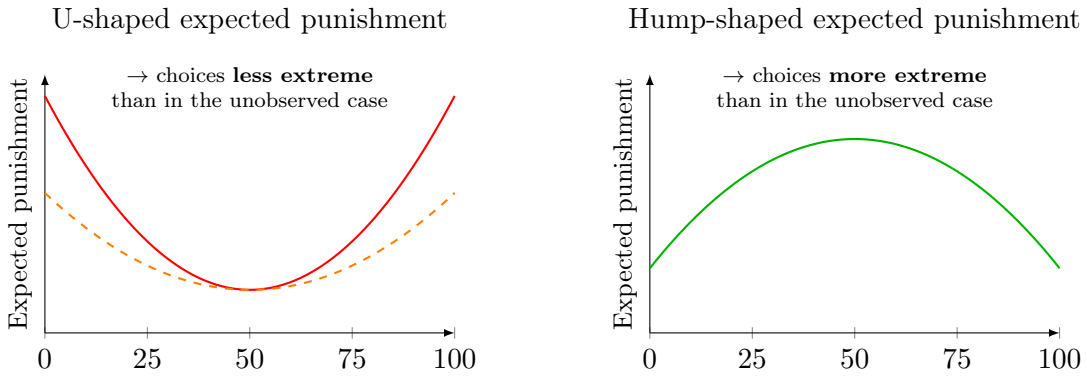
$$h(p_j)d_j - p_j \tag{7}$$

where $d_j \equiv |v_j - a|$ and $h(p)$ is an increasing function, common to all Observers (for simplicity), which satisfies $h(0) = 0$. The first part of the expression (7) captures the psychological utility from imposing a punishment p_j, which is increasing in the absolute difference between the Observer's own value v_j and the Decider's allocation choice a. The second part of the expression (7) is the material cost of punishing. Maximization of (7) yields the individual punishment function $p(a|v_j) = \mathcal{P}(d_j)$, whose shape depends on $h(p)$'s specific functional form. Below we provide some examples.[3]

Example 1: $h(p) = 2\sqrt{p} \Rightarrow \mathcal{P}(d_j) = d_j^2$ (convex).

Example 2: $h(p) = kp - \frac{p^2}{2}$ for some $k > 0 \Rightarrow \mathcal{P}(d_j) = k - \frac{1}{d_j}$ (concave) for $d_j \geq \frac{1}{k}$, $\mathcal{P}(d_j) = 0$ otherwise.

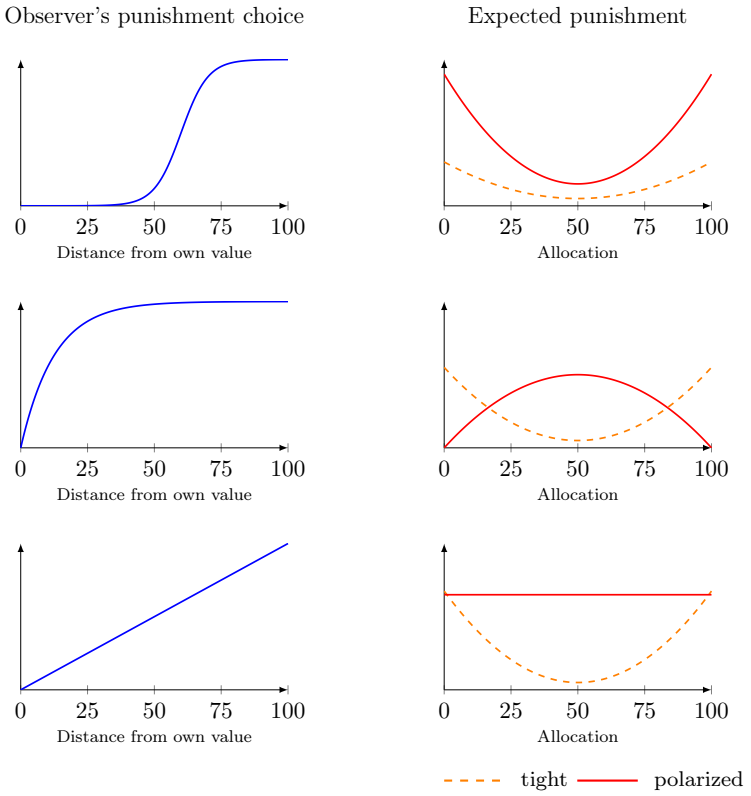
Figure 1: Relationship between allocation choices and expected punishment schedule



Note. The left graph shows the case of a U-shaped expected punishment schedule. This implies that choices will be less extreme than when the Decider is unobserved. The strength of the effect depends on the steepness of the expected punishment schedule. If it is almost flat (see dashed line) there will be hardly any changes. The right graph shows the case of a hump-shaped punishment schedule. Here, choices will be more extreme than when the Decider is unobserved.

[3]For conciseness, we ignore any upper bound on p_j. Allowing the existence of an upper bound is straightforward.

Figure 2: Examples of individual punishment functions and corresponding expected punishment



Note. The blue graphs in the left column depict different punishment functions. The graphs in the right column the corresponding (correct) expected punishment schedules. The red solid lines represent punishment expectations under polarized environments, the orange dashed lines under tight environments.

Example 3: $h(p) = k\,ln(1+p)$ for some $k > 0 \Rightarrow \mathcal{P}(d_j) = kd_j - 1$ (linear) for $d_j \geq \frac{1}{k}$, $\mathcal{P}(d_j) = 0$ otherwise.

To date, there is no consensus (either theoretical or empirical) on the shape of the mapping between personal values and punishment decisions, and hence on the shape of the punishment function.We therefore adopt an agnostic position in that respect.

Clearly enough, the shape of the individual punishment function $\mathcal{P}(d_j)$ mediates the relationship between the distribution of Observer values and the expected punishment schedule. Figure 2 illustrates this point. In the figure, the left column provides examples of individual punishment functions (in blue), where punishment depends on the distance between the observed allocation and the Observer's own value. The right column shows the expected punishment schedules associated with each individual punishment function, in different environments.[4] In red, we have the expected punishment when the distribution of Observer values is *polarized*: the allocation that the Observer finds appropriate is either 100 or 0 (each with equal probability). In orange, we have the expected punishment when the distribution of Observer values is *tight*: the allocation that the Observer finds appropriate is 50 with very high probability, and corresponds to each of the other options with a small, constant probability.

[4]In the Appendix, we explicitly derive the expected punishment schedule corresponding to each case.

Intuitively, suppose that Observers punish only when the observed allocation is sufficiently distant from their own value, as in the top left example depicted in Figure 2. In that case, as shown in the top right picture. expected punishment is U-shaped and minimized when the Decider chooses 50, the middle-of-the-road option that is never too far from any possible Observer value. Note that, in the polarized environment, choosing an extreme allocation generates a lottery where the Decider either incurs a high punishment or no punishment with equal probability. Instead, in the tight environment, the Observer is almost certainly a moderate. Any choice that the Decider may take is therefore expected to attract low punishment, with little variation between more or less extreme allocations. As a result, the expected punishment schedule is flatter than in the polarized environment, and the incentive to avoid extreme allocations to escape high punishment is weaker.

Consider now the bottom left example of Figure 2, where the punishment function is linear. In a polarized environment, this implies that any choice the Decider may take incurs the same punishment. To see this formally, suppose that the slope of the punishment function is β. A Decider who chooses 50 incurs a punishment given by 50β, while a Decider who chooses 100 incurs a punishment of either 0 or 100β with equal probability, meaning that expected punishment is $0.5(100\beta) = 50\beta$, i.e. the same. This implies that there is no punishment reduction from choosing a less extreme allocation, and stands in contrast to the case of a tight environment, where the expected punishment schedule is U-shaped, as depicted in the bottom right picture.

Overall, Figure 2 makes clear that the relationship between the distribution of Observer values and the expected punishment schedule is *a priori* ambiguous. In turn, this implies that the effect of the distribution of Observer values on Deciders' allocations may take different forms. In particular, more polarized environments may induce either more or less extreme allocations relative to less polarized ones.

2.1 Hypotheses

Hypothesis 1 *(a) Being observed affects the Deciders' allocation choices. (b) Suppose that the expected punishment schedule is U-shaped. Then, being observed induces less extreme allocation choices. (c) The opposite applies if the expected punishment schedule is hump-shaped.*

Consider now three environments, characterized by the following distributions of Observer values: (i) Tight: hump-shaped with low variance; (ii) Loose: hump-shaped with high variance; (iii) Polarized: U-shaped. In what follows, we say that the environment becomes progressively *more polarized* as we move from (i) to (ii) to (iii). The first hypothesis simply highlights that, generally, we expect the degree of polarization in the environment (i.e., the distribution of Observer values) to matter.

Hypothesis 2: *The Deciders' allocation decisions are affected by the degree of polarization in their environment.*

The next hypothesis zooms in on the mechanism behind the treatment effect: the shape of the expected punishment schedule.

Hypothesis 3 *Suppose that, as the environment becomes more polarized, the expected punishment schedule becomes flatter (if U-shaped) and/or changes from U-shaped to hump-shaped. Then, in*

more polarized environments, the allocation decisions become more extreme (and vice versa if the expected punishment schedule becomes steeper and/or changes from hump-shaped to U-shaped).

Finally, our last hypothesis concerns the role of the Deciders' own preferences in their allocation choices.

Hypothesis 4: *When the Deciders' allocation decisions are more extreme, they are more affected by personal values and political affiliation.*

3 Experimental Design

To empirically test the hypotheses derived from our theoretical framework, we design an experiment where participants act as "Deciders" and have to allocate money between two individuals. Deciders always take two decisions: First, a decision without being observed. Second, a decision for two new individuals, while knowing that their action will be visible to an "Observer" who can punish them by taking away some of their experimental earnings. Across three treatments, we exogenously vary the distribution from which the Observer is drawn. Specifically, participants are shown what Observers perceive as the "most appropriate" action. This distribution can either be *tight*, i.e., almost all Observers agree on the most appropriate action, *loose* with a greater variance among Observers, or *polarized*, with people holding completely opposite views.

Our design focuses on a setting that is particularly well-suited to study the importance of social norms, punishment, and polarization: political identity in the US. Both Deciders and Observers know the political identity of the two people among whom the money has to be split. This means that Deciders have to trade off two different norms in a highly politicized context: allocating earnings equally or favoring the person who shares their political views. In addition to the experimental variation we leverage the U.S. presidential elections as a shock to partisan-related beliefs and preferences. To do so, we collect data both before the U.S. elections in November 2024 and after the inauguration of Trump in January 2025. Figure 3 visualizes our experimental setup. In the next sections, we describe our design and treatments in more detail.

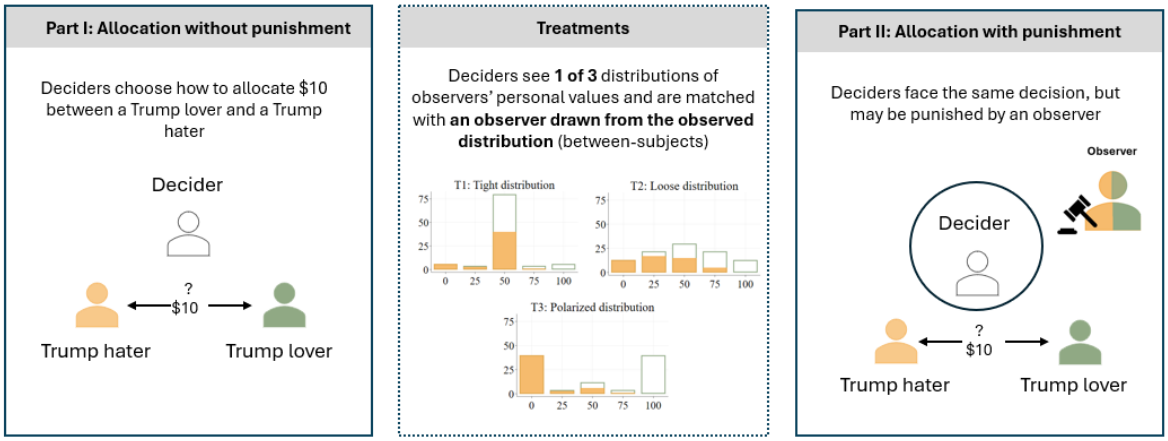
3.1 Allocation decisions

Participants join the study either as Deciders or Observers. At the beginning of the experiment, participants in both roles are asked about their views on Donald Trump and are invited to self-identify as either "Trump lovers" (TL) or "Trump haters" (TH). To strengthen identification with these categories, participants also justify their choice through an open-text response.[5] In the rest of this paper, we refer to the participants' view on Trump as their "political identity".

For Deciders, the experiment consists of two parts. In Part I, they allocate \$10 between two other participants (A and B), with participant A identified as a Trump hater and participant B as a Trump lover. This allocation task presents Deciders with a conflict between two competing

[5]Participants could alternatively indicate neutrality toward Trump. In such cases, we initially recorded their response as neutral, followed by a subsequent question asking them to specify whether they leaned more toward being a Trump lover or hater.

Figure 3: Overview of the experimental design



Note. Deciders have to take two allocation decisions between a Trump hater and lover — one without being observed (Part I) and one with being observed, thereby adding a threat of punishment to the decision (Part II). Between-subjects, we vary whether Observers are drawn from a tight, loose, or polarized distribution (treatments). These distributions are known to Deciders.

norms: first, a fairness norm suggesting an equal 50-50 division since no recipient differs in deservingness; and second, an in-group norm advocating preferential treatment of the recipient who shares the Decider's political identity. Deciders were informed that at the conclusion of the experiment, we would randomly select one participant and implement their chosen allocation between participants A and B.

Following the literature, we further investigate the motives underlying Deciders' allocation choices by eliciting both their personal values (PVs) – what they personally consider the most appropriate action—and their normative expectations (NEs) – their beliefs about what most Trump supporters and opponents deem appropriate (Bicchieri, 2005; Bicchieri and Dimant, 2019; Capraro et al., 2019; Bašić and Verrina, 2024; Charness et al., 2025). Together with PVs and NEs, the allocation decisions from Part I allow us to analyze how individuals balance competing motives in the absence of social influence.

In Part II, we reintroduce a social dimension to the Deciders' decisions. Deciders allocate money between a new pair of participants with opposing political identities. Again, one Decider's decision is randomly chosen and their decision is implemented for a real pair of participants. This time, however, they are informed that their decision will be communicated to an Observer, who can directly punish them by deducting part of their experimental earnings. Across three treatments, we test how Deciders' choices depend on whether their social environment is tight, loose, or polarized. Specifically, we show Deciders different distributions of PVs among Observers — one of which will be randomly matched with the Decider.

3.2 Treatment conditions

Figure 4 illustrates the distribution of Observers' PVs across treatments. All three treatments have identical mean values, with Observers on average favoring the allocation in which both recipients A and B receive $5. However, the treatments differ significantly in the variance of PVs. In T1, where values are tightly clustered around the mean allocation for both Trump

lovers (TL) and haters (TH), the standard deviation is 0.75. In T2, characterized by a looser concentration around the 50-50 allocation, the standard deviation increases to 1.22. Lastly, in T3, where PVs are polarized between Trump lovers and haters, the standard deviation is 1.82. Importantly, all three treatments are symmetric regarding Trump lovers and haters, ensuring we do not portray one political identity as inherently more extreme in their values.

We follow Dimant et al. (2025) in that the distributions of Observer behavior in Part II are constructed through non-random sampling from a previous experimental session.[6] The non-random sampling also implies that the distribution of PVs that we show to Deciders is identical for Wave 1 (pre-election) and Wave 2 (post-election). Deciders are informed that the shown distributions do not reflect general PVs but rather represent the views of a specific subgroup. Crucially, they also know that their Observer will be randomly drawn from the presented distribution, ensuring incentive compatibility.

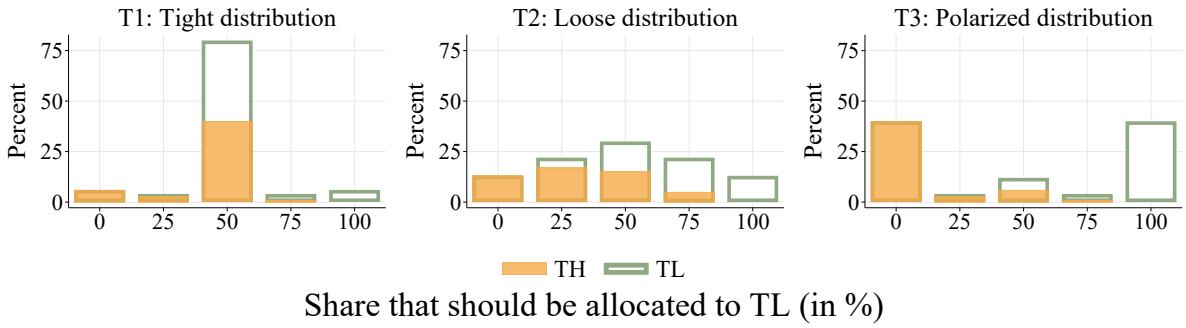
Deciders are shown the distribution and are informed that their allocation decision will be communicated to one randomly drawn Observer from the shown distribution. They are also informed in detail about the Observers' punishment options (see Section 3.3) and answer several control questions to ensure understanding before deciding again how to split \$10 between A and B. Finally, we ask Deciders about their punishment expectations. Specifically, they are asked about the average punishment they expect for each possible allocation decision.

3.3 Punishment decisions

The Observers' task is to decide whether they want to punish Deciders for their allocation decision in Part II. At the beginning of the experiment, we describe the Deciders' allocation decision to Observers, including the political identities of A and B. The political identities of the Decider are not known to the Observer.

For Part II, Deciders are allocated an additional bonus of 100 points, which is converted into USD at the end of the experiment. Observers can take away any amount (in steps of 10) of this bonus by engaging in costly punishment. They have to pay 3 points of their own bonus to

Figure 4: Shown distribution of Observers' PVs



[6]Similar methodological approaches have been employed by other studies (see, e.g., Frey and Meier, 2004; Bicchieri and Xiao, 2009; Krupka and Weber, 2009; Bursztyn et al., 2020), and the implications of using non-random samples in experiments are discussed by Charness et al. (2022) and Bardsley (2000).

take away 10 points from Deciders. We elicit punishment decisions using the strategy method. Observers have to decide how much they want to punish the Decider for each possible allocation decision. Afterward, we match Deciders to Observers and implement the relevant decision.

As for Deciders, we elicit PVs and NEs for Observers to learn more about their punishment motivation. In addition, we apply a novel tool that allows us to measure norm pluralism among Observers (Panizza et al., 2024a). This allows us to evaluate whether Observers perceive several co-existing norms in this scenario. This may be relevant for our context, as previous research indicates that people with singular norms punish norm transgressions more frequently and severely than those who hold multiple norms (Panizza et al., 2024b).

3.4 Sample and data collection

Our experiments and hypotheses were preregistered (see here for Wave 1 and Wave 2) on October 31, 2024, and January 21, 2025. The plan for a second wave was already described in the first preregistration. We programmed the experiment using Qualtrics (2005) and recruited participants online via Prolific. These dates were selected to allow a comparison of behavior before and after the U.S. presidential election. The Wave 1 data collection took place before Election Day, and Wave 2 occurred after Trump's inauguration.

For the Decider study, we recruited ∼900 U.S. participants per wave (N=901 for Wave 1, N=903 for Wave 2), representative in terms of gender and equally distributed across political identities (Trump supporter, opponent, or neutral). These numbers are in line with the preregistered sample size of N=900 per wave. We recruited not only Trump supporters or opponents, but also politically neutral participants because we wanted to study whether a polarized environment may even induce neutral Deciders to behave in a biased way. However, the strongest reactions are to be expected from partisans. We employed quota sampling based on self-reported political party affiliation to achieve an equal distribution across political identity. In total, our Decider sample consists of 37% Republicans, 36% Democrats, and 26% Independents, translating into 39% Trump supporters, 37% Trump opponents, and 24% neutrals.

For the Observer study, again in line with our preregistration, we recruited 300 U.S. participants per wave (N=301 for Wave 1, N=300 for Wave 2), again representative in gender but aiming for an equal distribution between Trump lovers and haters, as we do not include neutral Observers in the distribution shown to Deciders. Using self-reported political party affiliation, we ended up with a sample consisting of 55% Republicans, 41% Democrats and 4% Independents. This translates into 52% Trump lovers, 33% Trump haters, and 15% Neutrals.[7] Sample sizes were informed by results from a pilot run in October 2024.

The median completion time for Deciders (Observers) was 11 (18) minutes, with median earnings of $2.97 ($4.00).[8] In each wave, we also recruited two pairs of participants to receive

[7] To be able to match neutral Observers to Deciders, we used a follow-up question where we asked neutral Observers whether they were leaning more towards being a Trump lover or hater. 59% of neutral Observers indicated to lean towards being a Trump lover, 41% to being a Trump hater. Based on this information we re-classified neutral Observers as Trump lovers and haters to make matching possible.

[8] As completion times were slightly longer than predicted, we paid participants a higher show-up fee than stated in the instructions.

the allocations made by Deciders. These participants were entirely passive, and we recorded only their political identities (Trump supporter or opponent).

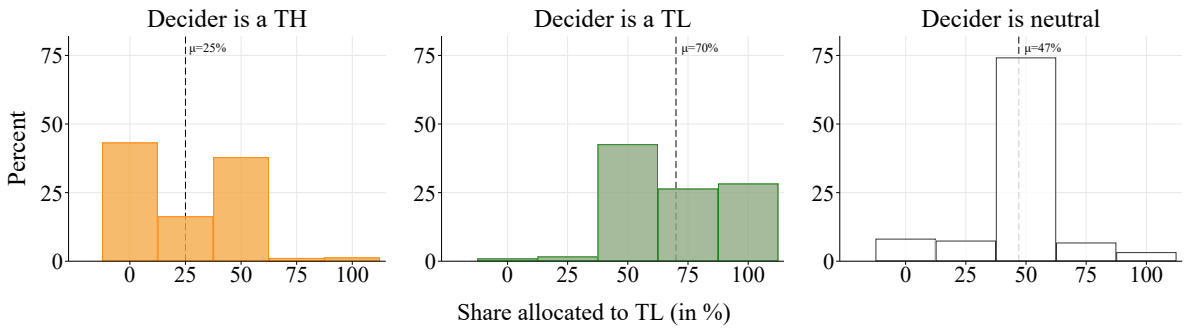
After completing the main experiment, participants responded to an ex-post survey collecting demographic information and additional data on their political views. We also measured participants' perceived closeness to individuals identifying as Trump supporters or opponents using the IOS scale (see Gächter et al., 2015; Baader et al., 2024). In Wave 2, we additionally elicited feelings of power, fear, perceived group status (related to political identity), and political resignation as supplementary control variables.

4 Results

4.1 Descriptives and baseline behavior

Measuring allocations without the threat of punishment provides baseline insights into the Deciders' decisions without any social influence. [9] Overall, the median allocation in our study is an equal 50-50 split. However, Figure 5 shows that allocation decisions vary substantially with the Decider's political identity. While the vast majority of neutral Deciders (74%) choose the 50-50 split, the median distribution for both Trump lovers and haters is to allocate 75% to their political in-group. Consequently, the median allocation decision differs significantly between political identities (Wilcoxon rank sum tests, $p < 0.001$). This pattern shows the direct effect of polarization and is in line with previous research documenting the effect of partisanship on unfair giving in dictator games (see e.g. Iyengar and Westwood, 2015b). The variance in allocations is significantly larger for partisans compared to neutral Deciders (F-test for the equality of standard deviations, $p < 0.001$). This is in line with partisan Deciders trading off different concerns. On the one hand, they may want to allocate the money equally, given that there is no difference in deservingness. On the other hand, they may want to favor their in-group member. Supporting this second motive, we find that partisan Deciders hold very different feelings towards Trump lovers or haters. Specifically, they feel significantly closer towards their political in-group

Figure 5: Allocation decisions without Observers

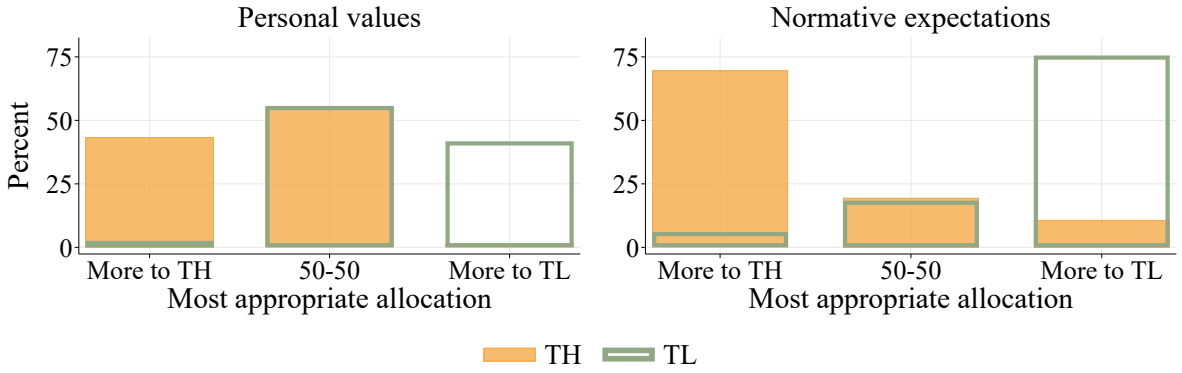


[9] The analysis pools observations from both waves. We discuss changes across waves in Section 5.

members (Wilcoxon signed rank tests, $p < 0.001$). Neutral Deciders, by contrast, do not feel differently close to Trump lovers or haters.

Allocation choices strongly correlate with PVs ($r = 0.65, p < 0.001$), although the share of Deciders stating that an equal split is the most appropriate action is significantly larger than the share of Deciders who actually implement this decision (62% versus 48%, McNemar test $p < 0.001$). Interestingly, Deciders have very wrong expectations about what others think is the most appropriate thing to do. Figure 6 compares actual PVs of Trump haters and lovers to the NEs Deciders hold about them.[10] The graph highlights the stark mismatch between PVs and NEs. While the modal PV for both Trump lovers and haters is to divide the money equally, the modal expectation is that partisans will give more to political in-group members. Expectations thus draw a substantially more polarized picture than what we observe in the data.

Figure 6: Personal values versus normative expectations



4.2 Main results

In line with our theoretical framework, we expect that Deciders change their behavior when being observed. Previous literature has stressed the disciplining effect of social sanctions, predicting less unequal behavior in the presence of an observer (see e.g. Bolton et al., 2021). We go beyond previous findings by moreover testing whether individuals will respond differently to being observed depending on the distribution from which the Observer is drawn.

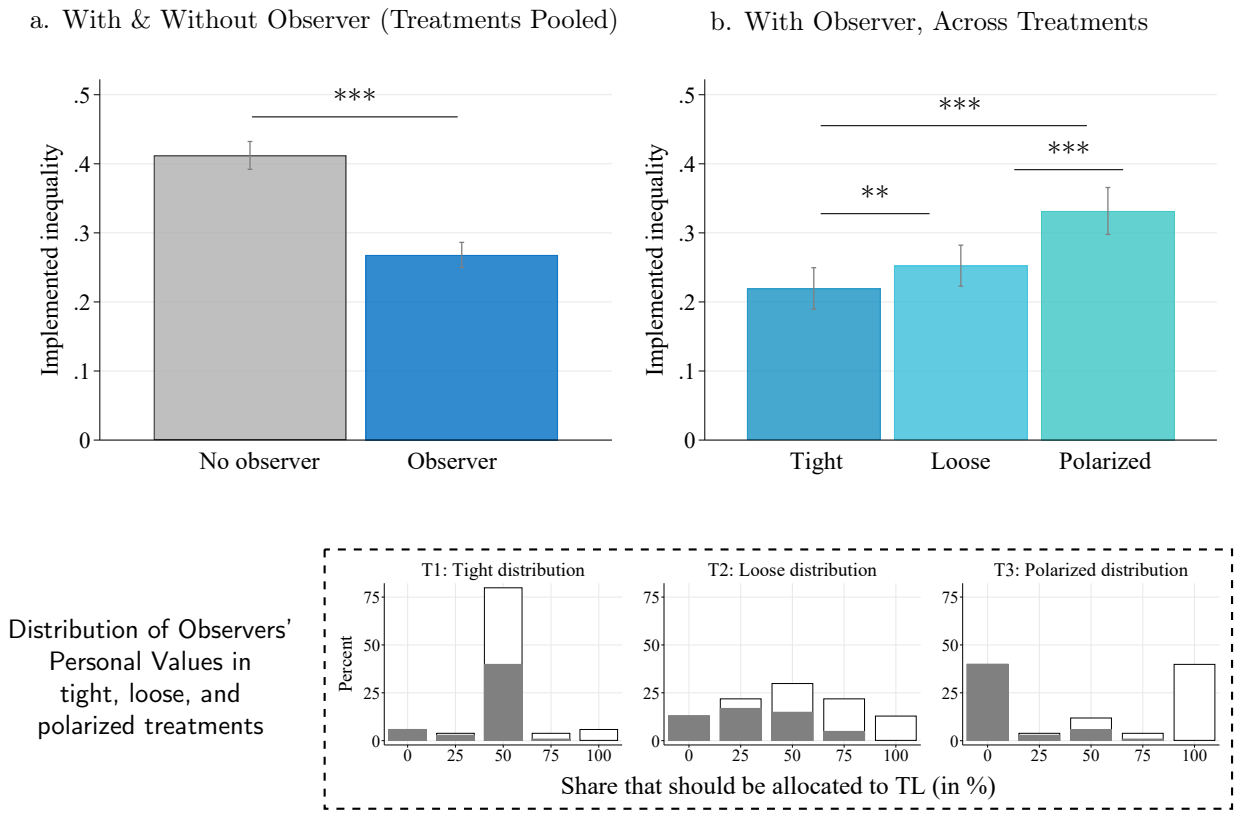
To make decisions between Trump lovers and haters more comparable, we do not simply look at the share allocated to the Trump lover (x_i), but at how much inequality Deciders implement, using the following measure: $D_i = |2x_i - 1|$. This is equivalent to expressing the implemented inequality as the Gini coefficient between two people. This results in three possible outcomes: a Gini of 0 (50-50 split between both A and B), a Gini of 0.5 (giving 75% to one of the two receivers), and a Gini of 1 (giving everything to one receiver).[11]

Figure 7 shows the implemented inequality with and without observability (Panel a.) and across treatments (Panel b.). First, when comparing behavior with and without Observers,

[10] See Appendix Figure B.1 for PVs by political identity.

[11] For robustness, we also look at the share of Deciders choosing the 50-50 split and the variance of allocations. Results for these alternative outcomes confirm our main analysis and can be found in Appendix B.2.

Figure 7: Inequality of Deciders' Allocations



Note. The top graphs show the level of inequality chosen by Deciders. Panel a. compares inequality without and with an observer. Panel b. zooms in on scenarios with an observer, but presents differences across treatments. Inequality is captured by the Gini coefficient based on the allocation decisions. The box beneath displays the distribution of Observer PVs that Deciders were shown in each treatment. Deciders knew their Observer would be randomly selected from the displayed distribution.

we see that being observed is associated with a striking decrease in the average implemented inequality D_i (Wilcoxon signed rank tests for observed vs non-observed, $p < 0.001$). Columns (1) and (2) in Table 1 regress the implemented inequality by Deciders on a dummy for whether their decisions can be observed, resulting in a 0.14 lower Gini coefficient under observability. Our results are consistent with the possibility of punishment having a disciplining effect on Deciders.

Result 1. *The possibility of social sanctions has a disciplining effect on Deciders. They implement less inequality if their decision can be punished by an Observer.*

Crucially, we find that the strength of this disciplining effect changes depending on the environment Deciders operate in. Panel b. in Figure 7 shows the main result of our paper. The implemented inequality differs significantly across environments: It is smallest in the tight and largest in the polarized environment (pairwise Wilcoxon rank sum tests: tight vs loose $p = 0.029$, tight vs polarized $p < 0.001$, loose vs polarized $p = 0.003$). Columns (3) and (4) in Table 1 interact being observed with the distribution of Observers. We can see that compared to the tight environment, observability has a significantly weaker effect in loose and polarized environments. Shifting from a tight to a polarized environment more than halves the effect of

observability. The regression in Panel B zooms in on the pure treatment effect, regressing the implemented inequality in Part II on tight, loose, and polarized environments. Relative to the

Table 1: Implemented inequality across observability and treatments

Panel A. With & Without Observability

Dependent variable:	Implemented inequality			
	(1)	(2)	(3)	(4)
Being observed	-0.14***	-0.14***	-0.19***	-0.20***
	(0.01)	(0.01)	(0.02)	(0.02)
Distributions (baseline=tight)				
Loose			-0.01	-0.01
			(0.02)	(0.02)
Polarized			-0.01	-0.00
			(0.02)	(0.02)
Interactions				
Being observed x Loose			0.04**	0.05**
			(0.02)	(0.02)
Being observed x Polarized			0.11***	0.11***
			(0.02)	(0.02)
Observations	3608	3558	3608	3558
Average inequality (no observer)	0.41	0.41	0.41	0.41
Demographic controls	No	Yes	No	Yes
R^2	0.42	0.44	0.42	0.45

Panel B. Treatment effects on Part II behavior

Dependent variable:	Implemented inequality in Part II	
	(1)	(2)
Distributions (baseline=tight)		
Loose	0.04**	0.04**
	(0.02)	(0.02)
Polarized	0.11***	0.10***
	(0.02)	(0.02)
Wald-tests (loose vs polarized)		
p-values	0.0002	0.0002
Observations	1804	1779
Average inequality (tight)	0.22	0.22
Demographic controls	No	Yes
R^2	0.42	0.43

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) in Panel A are clustered at the individual level. Robust standard errors in Panel B.

Note. Results of OLS regressions. The dependent variable in Panel A is the implemented inequality in decisions with and without an Observer (Gini), resulting in two observations per individual. Being observed is a dummy variable, taking the value 0 for an individual's Part I decision (without an observer) and 1 otherwise. The dependent variable in Panel B is the inequality implemented by Deciders in Part II (Gini), the case with observability. Distributions is a categorical variable that takes the value of 0 for the tight treatment (baseline category), 1 for loose, and 2 for polarized. Baseline controls include PVs in Panels A and B and the Decider's decision without an Observer in Panel B. Demographic controls include age, gender, political identity, party affiliation, risk preferences, voting behavior in the last presidential elections, education, and the measured difference in closeness towards a Trump lover and hater. All specifications include wave fixed effects.

average Gini of 0.22 in the tight environment, the loose environment increases inequality by 0.04 (18%), while the polarized environment raises it by 0.11 (50%). This magnitude is comparable to the well-documented effect of the source of inequality on redistribution: Almås et al. (2020) shows that shifting from luck- to merit-based inequality increases the implemented Gini by about 0.15 in Norway. Shifting from a tight to a polarized environment thus leads to a very meaningful increase in inequality. Wald tests further confirm that the difference between polarized and loose environments is significant in all specifications, with polarized environments leading to a larger implemented inequality.

Result 2. *The disciplining effect of social sanctions depends on the environment: as polarization increases, observability matters less and Deciders choose more unequal allocations.*

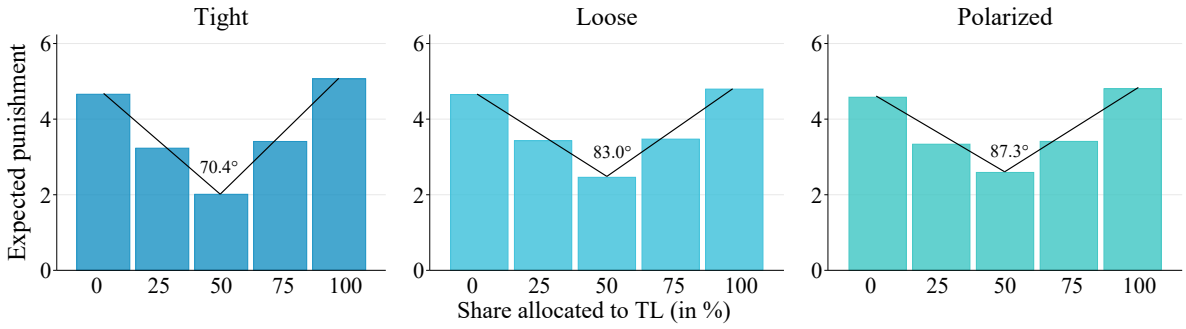
These results are robust to the inclusion of a rich set of demographic controls, as well as to excluding people who needed more than one attempt in control questions (see Appendix Table B.2), and become even more pronounced when using a Tobit specification (Tobin, 1958) to account for the censored nature of our data (see Appendix Table B.1).[12]

4.3 Mechanism: Expected punishment

The key distinction between Part I and Part II in our study is the presence of an Observer that can punish the Decider's decisions. Our theoretical framework argues that different punishment expectations are the key driver of the treatment effect we identified in the last section. Here, we investigate this mechanism empirically.

Figure 8 details the Deciders' punishment expectations. On average, Deciders expect to receive 3.7 punishment points out of a maximum of 10. As shown in Figure 8, expected pun-

Figure 8: Average expected punishment across treatments



Note. The figure shows average expected punishment (maximum punishment = 10) by allocation decision and treatment. The overlaid line shows the steepness of punishment expectations — i.e. the average absolute difference between expectations for extreme versus equal allocations. As indicated by the visualized angles, punishment expectations become flatter as we move from tight to loose and polarized environments.

[12] All our regressions include wave fixed effects. The regressions in the appendix present results both on aggregate and separately for each wave, showing consistent results across waves. There are however, some interesting differences in baseline behavior and punishment when comparing pre- and post-election data. We discuss these results in more detail in Section 5.

ishment follows on average a U-shaped pattern: Deciders anticipate the least punishment for a 50-50 split and the most for extreme allocations. This pattern holds across political identities (for more details, see Appendix Figure B.3). Overall, 73% of Deciders expect punishment to be minimized at the equitable allocation.[13]

To show that punishment expectations are a key driver of our findings, we need to show two things. i) First, that our treatments affect punishment expectations. Since treatments are randomly assigned, this effect is causal. ii) Second, we provide evidence that different punishment expectations correlate systematically with the implemented inequality. Since punishment expectations are endogenous and may also be shaped by other factors beyond treatment assignment, we need to be more cautious about the interpretation of these findings. To address this issue and get at the causal effect of punishment expectations on behavior, in a last step, we bring i) and ii) together, using treatment assignment as an instrument for punishment expectations.

First, we document that treatment assignment affects punishment expectations. Figure 8 shows that punishment expectations become flatter as we move from tight to loose or polarized environments. We define steepness as the average absolute difference between expectations for extreme versus equal allocations.[14] This is visualized in the figure as the angle between lines connecting expectations for extreme and equal allocations. The difference in steepness is statistically significant. Expectations in the tight environment are steeper than in the loose or polarized (Wilcoxon rank sum test: tight vs loose $p < 0.001$, tight vs polarized $p < 0.001$) and again steeper in the loose compared to the polarized environment ($p = 0.02$). Participants thus respond to the shown distributions and adjust their punishment expectations.[15] Columns (1) and (2) in Table 2 confirm this analysis using OLS regressions. Punishment expectations are flatter in loose and especially in polarized environments compared to tight environments. This analysis serves at the same time as the first stage for our iv approach. The F-statistic confirms that our treatment assignment is a relevant instrument for the steepness of punishment expectations.

Result 3. *Punishment expectations vary across environments. They are steeper if the environment is tight or loose compared to polarized environments.*

Next, we look at the OLS results to explore how punishment expectations correlate with implemented inequality. Columns (3) and (4) in Table 2 show that steeper punishment expectations lead to significantly less implemented inequality in Part II. Note that the R^2 is identical to the R^2 in Panel B of Table 1, where we present reduced form results from regressing implemented inequality on treatment indicators. This highlights that we have the same explanatory

[13] Figure 8 shows average expectations. See Appendix Figure B.4 for an illustration of individual heterogeneity and different types of punishment expectations.

[14] Our results are robust to focusing on the absolute difference in expected punishment for equal versus extreme allocations to the political out-group. In fact the correlation between average steepness and focusing only on the steepness with respect to the out-group is $r = 0.93$ and highly significant $p < 0.001$. Results are also identical when looking at aggregate steepness by treatment.

[15] Interestingly, we find that there is an asymmetry in the steepness of punishment expectations. The change in expected punishment between 50-50 and allocating everything to the Trump hater is larger than the change between 50-50 and allocating everything to the Trump lover (Wilcoxon rank sum test, $p < 0.03$). By contrast, we do not find a meaningful relationship between the stated closeness of a Decider towards a Trump lover or hater and the steepness or asymmetry of punishment expectations (Pearson's $r = -0.05$ and $r = 0.01$ respectively). This suggests that punishment expectations reflect genuine judgements rather than mere motivated beliefs.

Table 2: The role of punishment expectations: instrumental variable approach

Dependent variables:	Steepness		Implemented inequality in Part II			
	First stage		**OLS**		**IV**	
	(1)	(2)	(3)	(4)	(5)	(6)
Distributions (baseline=tight)						
Loose	-0.57***	-0.51***				
	(0.16)	(0.16)				
Polarized	-0.90***	-0.85***				
	(0.16)	(0.16)				
Steepness of punishment exp.			-0.01***	-0.01***	-0.11***	-0.12***
			(0.00)	(0.00)	(0.03)	(0.03)
Constant	5.11***	6.33***	0.08***	-0.07	0.54***	0.56***
	(0.16)	(0.52)	(0.02)	(0.06)	(0.12)	(0.19)
Observations	1804	1779	1804	1779	1804	1779
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes	No	Yes
First-stage F	15.20	13.62			15.20	13.62
R^2	0.03	0.05	0.41	0.42		

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Column (1) and (2) report first stage results. The dependent variable is the steepness of punishment expectations. Distributions is a categorical variable that takes the value of 0 for the tight treatment (baseline category), 1 for loose, and 2 for polarized. The steepness of punishment expectations measures for each individual the average absolute difference between expectations for extreme versus equal allocations and varies between 0 and 10. Columns (3) and (4) show OLS results, using the implemented inequality in Part II (as measured by the Gini) as the dependent variable and punishment expectations as the explanatory variable of interest. Columns (5) and (6) show the IV results, using the implemented inequality in Part II (as measured by the Gini) as the dependent variable. Baseline controls include PVs and the Decider's decision without an Observer. Demographic controls include age, gender, political identity, party affiliation, risk preferences, voting behavior in the last presidential elections, education, and the measured difference in closeness towards a Trump lover and hater. All specifications include wave fixed effects.

power independent of whether we use treatment indicators or punishment expectations as the independent variable of interest. Our results become stronger when using a tobit specification instead of OLS and are robust to looking at each wave separately B.4. They are also robust to restricting the analysis to Deciders with U-shaped punishment expectations only (see Appendix Table B.5). In line with our theoretical framework, Appendix Table B.6 shows that Deciders with U-shaped punishment expectations take more extreme and unequal decisions in Part II.

These empirical patterns paint a cohesive picture. We find that Deciders implement less inequality in tight compared to loose and polarized environments. Punishment expectations are steeper in tight environments, and steeper punishment expectations correlate with lower levels of inequality. As a last step, we bring these observations together using treatment assignment as an instrument for punishment expectations. Columns (5) and (6) show the IV estimates for the effect of punishment expectations on implemented inequality. We see that an increase in steepness significantly and causally reduces the implemented inequality. Intuitively, if u-shaped punishment expectations become steeper, deviations from the equal allocation will be punished more. The flatter the expected punishment schedule, the weaker the link between behavior and punishment. In the extreme case of completely flat punishment, the Decider is punished equally, independent of their own decision. Consequently, the disciplining effect of social sanctions disappears and behavior will be more similar to the case without an Observer,

i.e,. more extreme.[16]

The iv estimates are substantially larger than the OLS estimates. While the first stage shows a strong relevance of the instrument and treatment assignment is random, we also need to consider the validity of the exclusion restriction. Treatment assignment should only affect the allocation decision through changes in punishment expectations and not directly. To provide a test of this assumption, we ran a pre-registered follow-up experiment where we compare decisions across three treatments: i) a control treatment, where Deciders are not observed (akin to Part I decisions), ii) a distribution only treatment, where Deciders see the distribution of Observers but are then taking their decision in private, and iii) a punishment treatment, where Deciders are observed from a randomly drawn Observer of a known distribution (akin to Part II decisions). If the distribution that Deciders see in the different treatments had an effect beyond punishment expectations, we would expect to see a change in behavior between i) and ii). Appendix Figure B.5 shows this is not the case. Just being informed about the distribution has zero effect on allocation decisions. Only once the possibility of punishment comes into play, Deciders implement significantly less inequality. This provides strong support that the exclusion restriction holds and that changes in punishment expectations are the key mechanism in our study. In Appendix B.4 we provide more details on the additional data collection.

Result 4. *Steeper punishment expectations lead to a stronger disciplining effect of social norms and lower levels of implemented inequality. The flatter the expected punishment schedule, the weaker this disciplining effect, and the more unequal the Decider's decisions.*

One important implication of social sanctions having a less disciplining effect in loose and especially polarized environments is that there is a larger role for personal attitudes in driving behavior. To test this, we explore the impact of PVs and political preferences on Deciders' behavior across treatments. We use a linear probability model, where the dependent variable is the share allocated to the Trump lover and the independent variables of interest are treatment indicators, PVs (or political preferences), and their interaction. We hereby measure political preferences through the relative closeness Deciders report towards people with different political identities, i.e. how close a Decider indicates to feel to a Trump lover relative to a Trump hater.

Table 3 shows that the more a Decider believes one should allocate to the Trump lover (PVs) or the closer they feel to the Trump lover relative to the Trump hater (closeness), the more they also allocate to the Trump lover. While this relationship between personal attitudes and actions is very intuitive, we see that PVs and closeness matter significantly more in the polarized environment compared to the tight environment, as indicated by the positive interaction term that is significant in all specifications. Wald tests show that these personal attitudes also matter more in the polarized compared to the loose environment (with the exception of column 2, where $p = 0.11$). Finally, while the interaction between loose and PVs/closeness is also positive, it is much smaller and not statistically significant. These findings are robust across waves (see Appendix Table B.8).

Result 5. *PVs and political preferences have a larger effect on the Decider's allocation choices*

[16] Appendix Table B.7 shows that our results are robust to using the standard deviation of the observed treatment distribution as an instrument instead of directly using treatment dummies.

Table 3: Dependent variable = Share allocated to Trump lover

Dependent variable:	Share allocated to Trump lover			
	Personal values		**Political identity**	
	(1)	(2)	(3)	(4)
Distributions (baseline=tight)				
Loose	-0.05	-0.04	-0.01	-0.00
	(0.05)	(0.05)	(0.01)	(0.01)
Polarized	-0.16***	-0.17***	-0.02	-0.02*
	(0.05)	(0.04)	(0.01)	(0.01)
PVs	0.22***	0.16***		
	(0.02)	(0.02)		
Interactions				
Loose x PVs	0.02	0.02		
	(0.02)	(0.02)		
Polarized x PVs	0.07***	0.07***		
	(0.02)	(0.02)		
Closeness			0.03***	0.00
			(0.00)	(0.00)
Interactions				
Loose x Closeness			0.01	0.01*
			(0.00)	(0.00)
Polarized x Closeness			0.02***	0.02***
			(0.00)	(0.00)
Constant	0.08**	0.17***	0.51***	0.51***
	(0.03)	(0.04)	(0.01)	(0.03)
Observations	1804	1779	1804	1779
Average share to TL (tight)	0.5	0.5	0.5	0.5
Baseline controls	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes
R^2	0.43	0.48	0.26	0.33

* p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01. Robust standard errors in parentheses.

Note. Results of OLS regressions. The dependent variable is the share the Decider allocates to the Trump lover. Distributions is a categorical variable that takes the value of 0 for the tight treatment (baseline category), 1 for loose, and 2 for polarized. PVs is a discrete variable that indicates what the Decider thinks should be allocated to the Trump lover and can take values between 0 and 1. Closeness refers to the reported closeness towards a Trump hater relative to a Trump lover and can take values between -6 and 6. The baseline control is the Decider's decision without an Observer. The difference between loose x PVs/closeness and polarized x PVs/closeness is statistically significant in all columns (Wald tests, $p \leq 0.024$). All specifications include wave fixed effects.

in polarized environments.

4.4 Punishment behavior

So far, we have focused on the behavior of Deciders in our study. For Observers, our main outcome of interest is punishment behavior. Of particular interest is to understand if punishment behavior becomes indeed flatter in more polarized environments as expected by Deciders.

From the theoretical framework, we expect the punishment schedule to become flatter with polarization if Observers increase punishment linearly in the distance to their own PV. Figure 9 shows that we find support for this behavior in our data. Observers punish allocations more the further they are away from their own PVs. This result is confirmed when regressing punishment

decisions on the distance between allocation decisions and PVs and the distance squared (see Appendix Table B.11). While the distance is significant in all specifications, the squared distance is insignificant, suggesting a linear relationship, which holds for both Trump lovers and haters.[17]

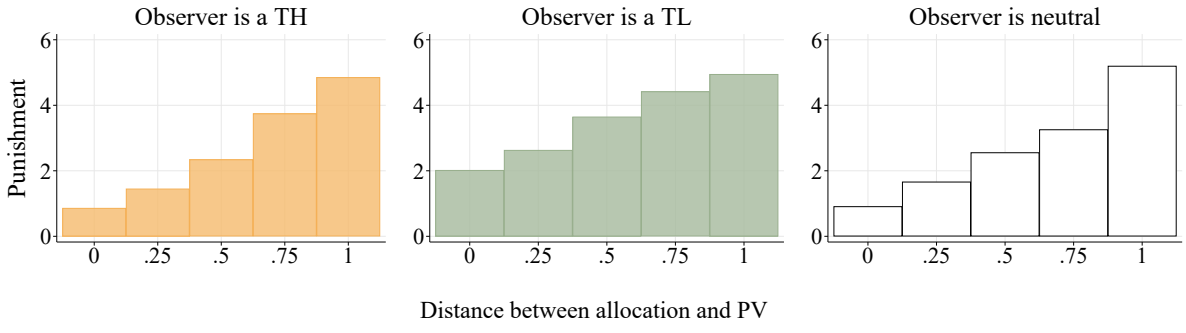
Next, we look at how Observers punish different allocation decisions. Across all Observers, we find that the equal 50-50 split is punished the least ($p < 0.001$, see Appendix Table B.10). In addition, both Trump lovers and haters punish unequal allocations that favor their in-group significantly less than unequal allocations that favor their out-group (see Appendix Figure B.6). This behavior is in line with the Observers' PVs, where — as with Deciders — the modal response to how the money should be split is 50-50 (see Appendix Figure B.7).

To see how punishment changes across treatments, we construct tight, loose, and polarized groups of Observers that match the treatment distributions Deciders saw in the experiment. To do so we calculate sample weights for Observers based on their PVs, in a manner that reproduces the tight, loose, and polarized distributions we used in our three treatments. We then examine the resulting average punishment schedules. Figure 10 shows that in the environment with tight PVs we observe relatively steep punishment schedules: the Observers punish the equal allocation by far the least, while punishing more extreme decisions more harshly. In the loose and particularly the polarized environment, by contrast, punishment schedules are significantly flatter.[18] In fact, it becomes almost completely flat in the polarized environment. This highlights that in a polarized environment, any action you take will upset somebody.

Result 6. *Compared to tight environments, punishment becomes significantly flatter in loose and polarized environments.*

Interestingly, the strength of this effect is not fully anticipated by Deciders. The diamond-shaped markers in Figure 10 show Deciders' punishment expectations in the different environments relative to the actual punishment behavior by Observers. As we can see, Deciders are qualitatively correct to expect the steepest punishment schedule in the tight environment. How-

Figure 9: Punishment depending on the difference between observed action and PVs

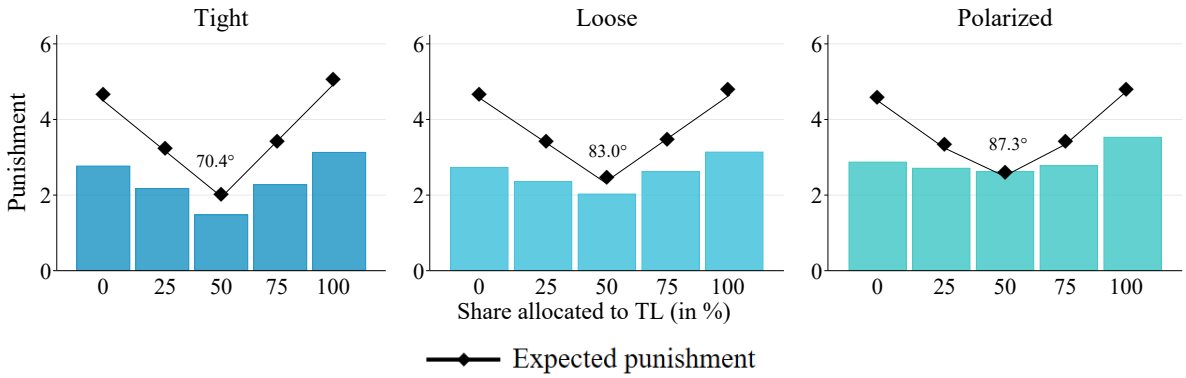


Note. The figure shows punishment behavior as a function of the difference between the implemented allocation by the Decider and the Observer's PV.

[17] At the individual level we see some heterogeneity in how punishment responds to the distance between allocation decision and Observer's PVs. We discuss this heterogeneity in Appendix B.5.

[18] Punishment decisions in the tight environment are significantly steeper than in loose or polarized ones, and steeper in loose than polarized environments (Wilcoxon ranksum tests, $p < 0.001$ for all comparisons).

Figure 10: Punishment across Observers with different distributions of PVs



Note. The figure shows punishment behavior from groups of Observers with a tight, loose, and polarized distribution of personal values. These distributions match the treatments shown to Deciders. The lines with diamond markers represent the Deciders' expected punishment in the different environments (as in Figure 8). The degrees above approximate the angle in the expected punishment schedule, confirming that expectations are much steeper in the tight than in the loose or polarized environment.

ever, overall, they underestimate how much the punishment behavior differs across treatments and, in particular, how much polarization flattens the punishment schedule. This means that if Deciders adjusted their expectations to match the actual punishment schedule, the impact of polarization would intensify, and the disciplinary power of social sanctions may disappear entirely. Under these conditions, individuals' choices would remain the same independent of whether they are observed or unobserved.

5 Pre- and post-election comparisons

We collected data in two waves: one before the 2024 presidential election and one shortly after Trump's inauguration in January 2025. This natural experiment allows us to examine whether shifts in the political environment influence how individuals weigh an equal-sharing norm against in-group norms, and how they respond to social observation and the threat of punishment.

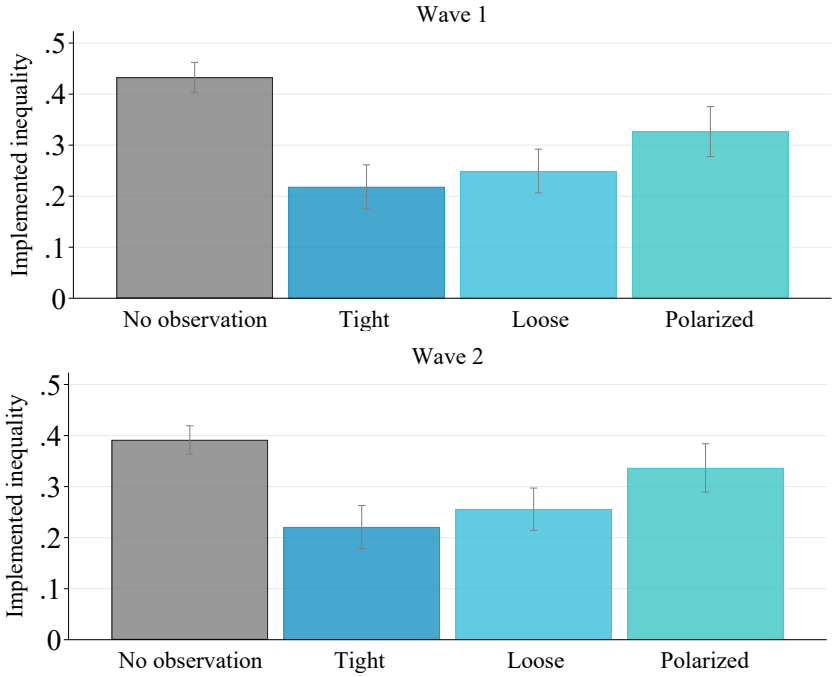
5.1 Differences in behavior and preferences

Overall, the impact of the election outcome on polarization and related behavioral measures is nuanced, which is consistent with existing survey findings (Holliday et al., 2024).

One significant change we observe in Wave 2 is political self-identification: we see significantly more participants self-identifying as Trump lovers — both among Republicans (Wilcoxon rank sum test, $p < 0.03$) and Democrats (Wilcoxon rank sum test, $p < 0.001$), which is in line with participants wanting to belong to the "winning team" (Carsey and Jackson, 2001).

Figure 11 shows that the average implemented inequality without an Observer is smaller in Wave 2 (Wilcoxon rank sum test, $p = 0.07$). This is mainly driven by Trump lovers (Wilcoxon rank sum test, $p = 0.01$). In line with this, Trump lovers report higher levels of closeness to Trump haters after the elections (Wilcoxon rank sum test, $p = 0.01$). There is no significant change in the behavior of Trump haters.

Figure 11: Implemented inequality across waves



Note. Implemented inequality (as measured by the Gini coefficient that results from Deciders' allocation decisions) across waves. The first bar shows the implemented inequality in the case without an observer. The other three bars show the implemented inequality in the case with observability, across treatments.

By contrast, the elections did not change how individuals respond to our treatments. Figure 11 highlights that Deciders respond to tight, loose, and polarized environments in the same way across waves (see Appendix Table B.14 for a formal test). Deciders implement significantly more inequality in loose and especially polarized environments compared to tight environments. This consistency shows that our experiment isolates a more fundamental mechanism that is not affected by political shocks. In line with the constant responses to our treatments, we also find that the elections do not seem to have changed the Deciders' punishment expectations. Interestingly, this expectation is not correct. When looking at Observers, we find significantly more punishment in Wave 2 (Wilcoxon rank sum test, $p = 0.009$), which is again mainly driven by Trump lovers (Wilcoxon rank sum test, $p = 0.003$).[19]

An important point when comparing data across waves is that we are measuring people's responses to a fixed environment. The environment itself remains the same before and after the elections. What may have changed are people's beliefs about the prevailing norm environment in the U.S. and, as a result, their expectations about punishment relative to a representative distribution. In fact, when comparing normative expectations across waves, we find that both Trump supporters and opponents viewed equal allocations as more appropriate after the elections (Wilcoxon rank-sum tests, $p < 0.001$ and $p = 0.003$, respectively). This indicates that the pre-election period amplified perceptions of polarization.

[19] While we do not see a change in punishment expectations per se, we do find that the relative importance of them increases for the Decider's allocation decision (see Appendix Table B.12).

5.2 The role of emotions in Wave 2 decisions

To understand potential drivers of behavior after the election, we elicit several control variables that aim to measure emotional responses. We ask participants six items that assess how they or their group's status and power has changed as a result of the elections (see Appendix C for instructions). From the weighted average of these items, we construct one variable measuring "power". Similarly, we construct a variable measuring "fear" from the weighted average of two items asking about the fear of being treated unfairly after the election and the fear of democracy being threatened. Finally, we have one question asking them whether they were satisfied or disappointed with the campaign of their party.

We find some interesting patterns when exploring the role of emotions in Wave 2. For Deciders who are Trump lovers, we see that the more satisfied they are with their party's campaign, the more inequality they implement. Similarly, for Observers we see that being satisfied with their party's campaign leads Trump lovers to impose harsher punishments (see Appendix Table B.17). For Trump haters, by contrast, the more important feeling is fear. The more inequality they implement, the more they feel afraid (see Appendix Table B.15). Being more afraid also translates into higher punishment expectations for Deciders (see Appendix Table B.16). This highlights that emotions can be a key driver for how individuals trade off different norms in highly politicized contexts.

6 Concluding remarks

Societies depend on the threat of social disapproval to keep partisan impulses in check, yet those impulses often escape restraint. Our theoretical model identifies a subtle cause: citizens' expectations about the shape of the sanction curve. When people believe that both partisan extremes will face stronger condemnation than a balanced stance, fairness becomes the safest strategy. But if they expect disapproval to be evenly distributed across the spectrum, showing partisan loyalty becomes individually rational and the 'safe haven' of fairness disappears. Polarization widens the range of moral viewpoints, which flattens the expected sanction curve and weakens its disciplining effect. By focusing on how people *perceive* the shape of sanctions rather than simply whether sanctions exist, we turn polarization from a background condition into a strategic factor – one that can be studied experimentally and has direct policy relevance.

To bring this insight to data, we conducted a preregistered two-wave online experiment bracketing the 2024 U.S. election. In the experiment, "Deciders" allocated a fixed sum between a Trump supporter and a Trump opponent, first in private and then under the eyes of an "Observer" who could punish their actions. Crucially, the mean norm embodied by the Observer pool was held constant, while its variance and shape were exogenously varied across three environments: tight (narrow, single-peaked), loose (wide, single-peaked), and polarized (bimodal). This design allows us to isolate the causal role of polarization in shaping the disciplining effect of social norms and their welfare implications.

Our results are striking. First, when decisions are private, partisans favor their own camp, confirming that intrinsic motives push allocations away from equity. Second, introducing an Observer discipline that biases, but the effectiveness depends on the environment Deciders are

facing. In tight environments, where Observers' ideals cluster near the 50-50 split, Deciders gravitate toward fairness. Relaxing that consensus - in loose environments - weakens this effect. In a bimodal, polarized environment, finally, there is only a weak disciplining effect of being observed, and behavior is most similar to the unobserved case. Third, individual Observers punish almost perfectly linearly in the distance from their own ideal; aggregating those linear rules across a polarized pool flattens the expected sanction curve to near indifference. Deciders sense this flattening yet underestimate its severity. A post-election wave, which increases actual sanctions for winners and heightens fear among losers, reinforces the same logic: expectations adjust sluggishly, so the binding constraint on cooperation remains the curve people think they face, not the one that is actually applied.

Taken together, our study highlights that fairness in polarized societies depends on how people interpret the strategic landscape they operate within. Social pressure can still promote cooperation, but only when citizens believe that some behaviors will not trigger partisan backlash. By experimentally varying the full distribution of normative cues – rather than just their average – we introduce both a diagnostic tool and a policy lever. The diagnostic tool identifies contexts where the perceived sanction curve is already flat; the lever makes hidden areas of consensus more visible, thereby steepening the curve people believe they face. Future research can test how durable such interventions are, extend the framework to other identity divides, and examine how informal social sanctions interact with formal enforcement. Mapping – and, where possible, reshaping – the strategic landscape of social disapproval is crucial for sustaining cooperation in an era of deepening polarization.

References

Almås, I., Cappelen, A. W., and Tungodden, B. (2020). Cutthroat capitalism versus cuddly socialism: Are americans more meritocratic and efficiency-seeking than scandinavians? *Journal of Political Economy*, 128(5):1753–1788.

Atkin, D., Colson-Sihra, E., and Shayo, M. (2021). How do we choose our identity? a revealed preference approach using food consumption. *Journal of Political Economy*, 129(4):1193–1251.

Baader, M., Starmer, C., Tufano, F., and Gächter, S. (2024). Introducing ios11 as an extended interactive version of the 'inclusion of other in the self'scale to estimate relationship closeness. *Scientific Reports*, 14(1):8901.

Balafoutas, L., Dimant, E., Gächter, S., and Krupka, E. (2024). Social norms: Enforcement, breakdown & polarization. *European Economic Review*, page 104885.

Bardsley, N. (2000). Control without deception: Individual behaviour in free-riding experiments revisited. *Experimental Economics*, 3(3):215–240.

Bašić, Z. and Verrina, E. (2024). Personal norms—and not only social norms—shape economic behavior. *Journal of Public Economics*, 239:105255.

Bauer, K., Chen, Y., Hett, F., and Kosfeld, M. (2024). Group identity and belief formation: a decomposition of political polarization. Working paper.

Baxter-King, R., Brown, J. R., Enos, R. D., Naeim, A., and Vavreck, L. (2022). How local partisan context conditions prosocial behaviors: Mask wearing during covid-19. *Proceedings of the National Academy of Sciences*, 119(21):e2116311119.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. and Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public Choice*, pages 1–22.

Bicchieri, C., Dimant, E., and Xiao, E. (2021). Deviant or wrong? the effects of norm information on the efficacy of punishment. *Journal of Economic Behavior & Organization*, 188:209–235.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208.

Bolton, G., Dimant, E., and Schmidt, U. (2021). Observability and social image: On the robustness and fragility of reciprocity. *Journal of Economic Behavior & Organization*, 191:946–964.

Boxell, L., Conway, J., Druckman, J. N., and Gentzkow, M. (2022). Affective polarization did not increase during the covid-19 pandemic. *Quarterly Journal of Political Science*, 17(4):491–512.

Boyd, R. and Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology*, 13(3):171–195.

Braghieri, L., Eichmeyer, S., Levy, R., Mobius, M., Steinhardt, J., and Zhong, R. (2024). Level slant and polarization of news consumption on social media. *Available at SSRN 4932600*.

Broockman, D. E., Kalla, J. L., and Westwood, S. J. (2023). Does affective polarization undermine democratic norms or accountability? maybe not. *American Journal of Political Science*, 67(3):808–828.

Bursztyn, L., Egorov, G., and Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review*, 110(11):3522–48.

Capraro, V., Jagfeld, G., Klein, R., Mul, M., and de Pol, I. v. (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. *Scientific Reports*, 9(1):1–11.

Carsey, T. M. and Jackson, R. A. (2001). Misreport of vote choice in us senate and gubernatorial elections. *State Politics & Policy Quarterly*, 1(2):196–209.

Charness, G., Dimant, E., Gneezy, U., and Krupka, E. (2025). Experimental methods: Eliciting and measuring social norms. *Journal of Economic Behavior & Organization*.

Charness, G., Samek, A., and van de Ven, J. (2022). What is considered deception in experimental economics? *Experimental Economics*, 25(2):385–412.

d'Adda, G., Dufwenberg, M., Passarelli, F., and Tabellini, G. (2020). Social norms with private values: Theory and experiments. *Games and Economic Behavior*, 124:288–304.

DellaPosta, D. (2020). Pluralistic collapse: The "oil spill" model of mass opinion polarization. *American Sociological Review*, 85(3):507–536.

DellaVigna, S. and Kim, W. (2025). Policy diffusion and polarization across us states.

Dimant, E. (2023). Beyond average: A method for measuring the tightness, looseness, and polarization of social norms. *Economics Letters*.

Dimant, E. (2024). Hate trumps love: The impact of political polarization on social preferences. *Management Science*, 70(1):1–31.

Dimant, E., Gelfand, M., Hochleitner, A., and Sonderegger, S. (2025). Strategic behavior with tight, loose, and polarized norms. *Management Science*, 71(3):2245–2263.

Dimant, E. and Kimbrough, E. O. (2024). Polarization in multidisciplinary perspective. *PNAS Nexus*, 3(10):pgae425.

Dimant, E., Reinhardt, L., and Sambanis, N. (2024). Racial identity, reactions to inequality, and fairness concepts among Americans. Working paper.

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., and Ryan, J. B. (2024). *Partisan hostility and American democracy: explaining political divisions and when they matter*. University of Chicago Press.

Durkheim, (1893). *De la division du travail social: Étude sur l'organisation des sociétés supérieures*. Alcan, Paris. First edition.

Fehr, E. and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3):159–181.

Frey, B. S. and Meier, S. (2004). Social comparisons and pro-social behavior: Testing" conditional cooperation" in a field experiment. *American Economic Review*, 94(5):1717–1722.

Gächter, S., Starmer, C., and Tufano, F. (2015). Measuring the closeness of relationships: a comprehensive evaluation of the'inclusion of the other in the self'scale. *PloS one*, 10(6):e0129478.

Gelfand, M. J., Gavrilets, S., and Nunn, N. (2024). Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change. *Annual Review of Psychology*, 75(1):341–378.

Holliday, D. E., Lelkes, Y., and Sean J. Westwood, S. J. (2024). Did the 2024 election change american attitudes about democracy. Working Paper Available at SSRN: https://prlpublic.s3.us-east-1.amazonaws.com/reports/election$_2$024$_r$*eport.pdf*.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22:129–146.

Iyengar, S. and Westwood, S. J. (2015a). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3):690–707.

Iyengar, S. and Westwood, S. J. (2015b). Fear and loathing across party lines: New evidence on group polarization. *American journal of political science*, 59(3):690–707.

Jia, R. and Persson, T. (2021). Choosing ethnicity: The interplay between individual and social motives. *Journal of the European Economic Association*, 19(2):1203–1248.

Kish Bar-On, K., Dimant, E., Lelkes, Y., and Rand, D. G. (2024). Unraveling polarization: insights into individual and collective dynamics. *PNAS Nexus*, 3(10):pgae426.

Krupka, E. and Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, 30(3):307–320.

Lees, J. and Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature human behaviour*, 4(3):279–286.

Levendusky, M. S. (2009). The microfoundations of mass polarization. *Political Analysis*, 17(2):162–176.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, 111(3):831–870.

Mason, L. (2015). "i disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American journal of political science*, 59(1):128–145.

Musolff, R. and Zimmermann, F. (2025). Polarization and the 2024 u.s. presidential election. Mimeo.

Panizza, F., Dimant, E., , Druckman, J., and Kimbrough, E. O. (2026). Norm multiplicity and the enforcement of democratic principles. *Working Paper*.

Panizza, F., Dimant, E., Kimbrough, E. O., and Vostroknutov, A. (2024a). Measuring norm pluralism and perceived polarization in U.S. politics. *PNAS Nexus*, 3(10):pgae413.

Panizza, F., Dimant, E., Kimbrough, E. O., and Vostroknutov, A. (2024b). Pluralism breeds tolerance. *Working Paper*. Available at SSRN: https://dx.doi.org/10.2139/ssrn.4649792.

Qualtrics (2005). Qualtrics software, Version [January 2021] of Qualtrics. Copyright ©[2021]. Qualtrics, Provo, Utha, USA. Available at: https://www.qualtrics.com.

Tirole, J. (2023). Safe spaces: shelters or tribes? Mimeo.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, pages 24–36.

Voelkel, J. G., Stagnaro, M. N., Chu, J. Y., Pink, S. L., Mernyk, J. S., Redekopp, C., Ghezae, I., Cashman, M., Adjodah, D., Allen, L. G., et al. (2024). Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Science*, 386(6719):eadh4764.

A Theoretical Appendix

A.1 Proofs

Proof of proposition 1: Denote the allocation at which $P(a)$ is minimized (for the case of U-shaped) or maximized (for the case of hump-shaped) as $\widetilde{a}$. In the main text we focus on $\widetilde{a} = 50$, but the result can be expressed more generally. U-shaped case: Evaluated at $a = a_i^o = \theta_i - P'(a_i^o)$, expression $-a_i + \theta_i$ takes the value $P'(a_i^o)$. If $a_i^o < \widetilde{a}$, then this is < 0. Hence, $a_i^u < a_i^o$. If $a_i^u > \widetilde{a}$, then this is > 0. Hence, $a_i^u > a_i^o$. Hump-shaped case: Evaluated at $a = a_i^o = \theta_i - P'(a_i^o)$, expression $-a_i + \theta_i$ takes the value $P'(a_i^o)$. If $a_i^o > \widetilde{a}$, then this is < 0. Hence, $a_i^u < a_i^o$. If $a_i^u < \widetilde{a}$, then this is > 0. Hence, $a_i^u > a_i^o$ ■

Proof of proposition 2: Denote the allocation at which $P_0(a)$ and $P_1(a)$ are minimized as $\widetilde{a}$. In the main text we focus on $\widetilde{a} = 50$, but the result can be expressed more generally. The Decider's optimal allocation under these two expected punishment schedules, a_0^o and a_1^o, satisfy $-a_0^o + \theta_i - P_0'(a_0^o) = 0 = -a_1^o + \theta_i - P_1'(a_1^o)$. Evaluated at $a = a_1^o = \theta_i - P_1'(a_1^o)$, expression $-a + \theta_i - P_0'(a)$ becomes $P_1'(a_1^o) - P_0'(a_1^o)$. If $a_1^o < \widetilde{a}$ then this is < 0. Hence, $a_0^o < a_1^o$. If $a_1^o > \widetilde{a}$ then $P_1'(a_1^o) - P_0'(a_1^o) > 0$. Hence, $a_0^o > a_1^o$. Part (b) Note that since P_0 is everywhere steeper than P_1 except at a, we have $P_0''(a) > P_1''(a)$. The proof then follows from total differentiation of (5), which yields $\frac{da_i^P}{d\theta_i} = \frac{1}{1+P_i''(a_i^P)}$. ■

Proof of corollary 1 Straightforward.■

A.2 preregistered hypotheses

In our preregistration, the hypotheses are stated as follows:

Hypothesis 1: *The nature of the distribution from which the Observer's personal values are taken affects (i) the share of Deciders selecting x=5, (ii) the average D_i, defined as $D_i = |2x_i - 10|$, a measure of how unequal is the Decider's choice, and (iii) the variance of the Deciders' allocation decisions.*

Hypothesis 2: *If a Decider has a U-shaped (resp hump-shaped) expected punishment schedule, then being observed induces the Decider to make less (more) extreme choices. We define "more extreme" as further away from the interior minimum/maximum than the decision without Observers. Hump shaped refers to a distribution with one interior maximum, no interior minimum. U-shaped refers to a distribution with one interior minimum, no interior maximum. An interior value x_i^* (resp, x_i^{**}) is an interior minimum (maximum) of a function $f_i(x)$ if $f_i(x) > f_i(x_i^*)(f_i(x) < f_i(x_i^{**}))$ for all $x \neq x_i^*(x_i^{**})$.*

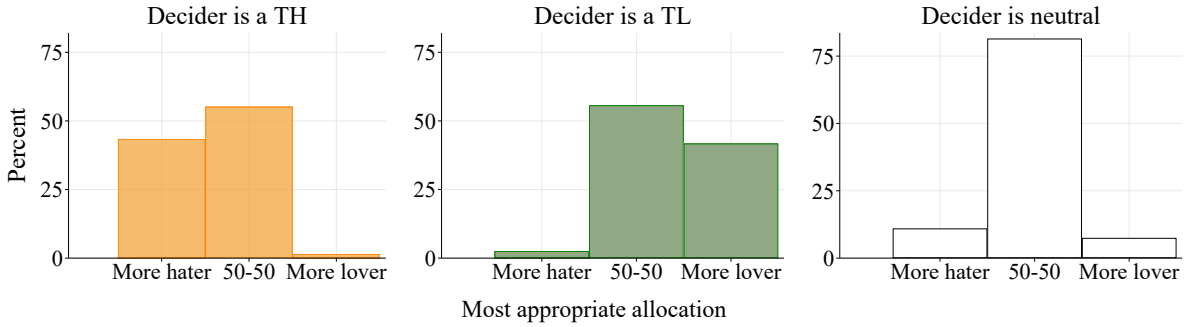
Hypothesis 3: *Focusing on Deciders with U-shaped expected punishment schedules, steeper schedules are associated with less extreme allocation decisions.*

Hypothesis 4: *Personal values and political preferences have a larger effect on the Decider's allocation choices in treatments with higher variance of actions.*

B Additional Analysis

B.1 Descriptives and baseline behavior

Figure B.1: Personal values of Deciders



B.2 Main results - robustness

Tables B.1 and B.2 show that the main specification in the paper is robust to using a tobit instead of an OLS specification and to imposing a strict exclusion criterion for comprehension. Both robustness checks confirm that the disciplining effect of social norms is weaker in more polarized environments. Both Tables report results separately for the full sample and separately for Wave 1 and Wave 2, showing overall consistent results. Table B.1 moreover shows that when using a tobit specification, the estimated marginal effects are larger than when using OLS.

Table B.1: Treatment differences: The effect of observability across environment - tobit

Dependent variable:	Implemented inequality in Part II					
	Full sample		**Wave 1**		**Wave 2**	
	(1)	(2)	(3)	(4)	(5)	(6)
Distributions (baseline=tight)						
Loose	0.10***	0.10***	0.11**	0.10*	0.10**	0.09**
	(0.03)	(0.03)	(0.05)	(0.05)	(0.04)	(0.04)
Polarized	0.22***	0.21***	0.24***	0.24***	0.21***	0.20***
	(0.04)	(0.04)	(0.06)	(0.06)	(0.05)	(0.05)
Observations	1804	1779	901	885	903	894
Average inequality (tight)	0.22	0.22	0.22	0.22	0.22	0.22
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes	No	Yes
Pseudo R^2	0.28	0.29	0.28	0.29	0.28	0.29

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.
Note. The table reports average marginal effects from tobit regressions (data is censored at 0 and 1). The dependent variable is the implemented inequality in Part II (as measured by the Gini). Distributions is a categorical variable that takes the value of 0 for the tight treatment (baseline category), 1 for loose, and 2 for polarized. The difference between loose and polarized is statistically significant in all columns (Wald tests $p \leq 0.03$). Baseline controls include PVs and the Decider's decision without an Observer. Demographic controls include age, gender, political identity, party affiliation, risk preferences, voting behavior in the last presidential elections, education, and the measured difference in closeness towards a Trump lover and hater. Columns (1) and (2) include wave fixed effects.

In addition to the implemented inequality, we look at two alternative outcome measures: the share of 50-50 choices and the variance of implemented decisions.

We find that the share of 50-50 choices significantly increases in the presence of an Observer (Wilcoxon signed rank tests for observed vs non-observed, $p < 0.001$ for all treatments). This is clearly the flip side of the documented decrease in implemented inequality that we discuss in the main paper. Similarly, the distribution of Observers matters for decisions in Part II. We find that the share of Deciders implementing zero inequality – and thus the 50-50 split - differs significantly across treatments (pairwise Wilcoxon rank sum tests: tight vs loose $p = 0.006$, tight vs polarized $p < 0.001$, loose vs polarized $p = 0.04$). This is also confirmed in Table B.3.

The presence of an Observer also decreases the overall variance of decisions for the tight and loose treatments (F-test for the equality of standard deviations for observed vs non-observed: tight $p < 0.001$, loose $p = 0.002$), but not for the polarized treatment. This general pattern holds for both Trump lovers and haters. When focusing on neutral Deciders, there are no differences across treatments (see Figure B.2). This is in line with the vast majority of neutral Deciders favoring the equal 50-50 split, independent of treatment or being observed. When looking at the variance across treatments in Part II, we find that there is significantly more variance in the polarized compared to the tight and loose environment (F-test for the equality of standard

Table B.2: Treatment differences: The effect of observability across environment - comprehension robustness

Dependent variable:	Implemented inequality (Gini) in Part II					
	Full sample		**Wave 1**		**Wave 2**	
	(1)	(2)	(3)	(4)	(5)	(6)
Distributions (baseline=tight)						
Loose	0.05***	0.05**	0.07**	0.07**	0.03	0.02
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Polarized	0.13***	0.12***	0.15***	0.15***	0.10***	0.10***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Observations	1301	1288	680	672	621	616
Average inequality (tight)	0.18	0.18	0.18	0.18	0.19	0.19
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes	No	Yes
R^2	0.40	0.41	0.39	0.40	0.43	0.45

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Results of OLS regressions. The dependent variable is the implemented inequality in Part II (as measured by the Gini). Distributions is a categorical variable that takes the value of 0 for the tight treatment (baseline category), 1 for loose, and 2 for polarized. We restrict the sample to participants who answer all three comprehension checks correctly the first time (72% of the sample). The difference between loose and polarized is statistically significant in all columns (Wald tests $p \leq 0.012$). Baseline controls include PVs and the Decider's decision without an Observer. Demographic controls include age, gender, political identity, party affiliation, risk preferences, voting behavior in the last presidential elections, education, and the measured difference in closeness towards a Trump lover and hater. Columns (1) and (2) include wave fixed effects.

Table B.3: Treatment differences: The effect of observability across environment - robustness 50-50 choices

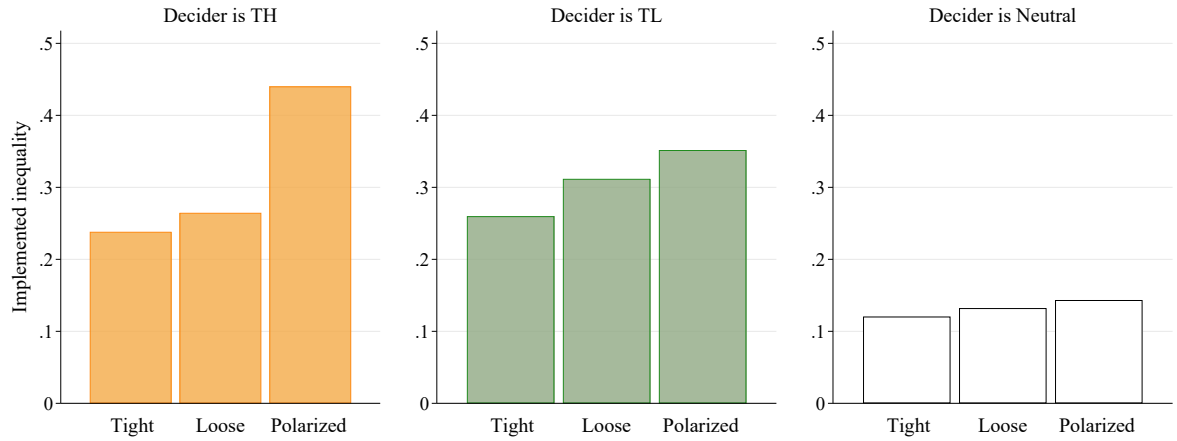
Dependent variable:	Choosing 50-50 in Part II					
	Full sample		**Wave 1**		**Wave 2**	
	(1)	(2)	(3)	(4)	(5)	(6)
Distributions (baseline=tight)						
Loose	-0.07***	-0.07***	-0.06*	-0.05*	-0.08***	-0.08***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Polarized	-0.12***	-0.11***	-0.10***	-0.11***	-0.13***	-0.13***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Constant	0.70***	0.91***	0.65***	0.83***	0.72***	0.97***
	(0.03)	(0.07)	(0.04)	(0.10)	(0.04)	(0.09)
Observations	1804	1779	901	885	903	894
Average share (tight)	0.72	0.72	0.73	0.73	0.71	0.71
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes	No	Yes
R^2	0.39	0.41	0.38	0.40	0.41	0.43

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Results of OLS regressions. The dependent variable is a binary variable, taking the value of 1 if a Decider in Part II chooses the 50-50 split, and 0 otherwise. Distributions is a categorical variable that takes the value of 0 for the tight treatment (baseline category), 1 for loose, and 2 for polarized. The difference between loose and polarized is statistically significant for the full sample (Wald tests $p \leq 0.03$) and marginally significant when looking at each wave separately (Wald tests between $p = 0.08$ and $p = 0.13$). Baseline controls include PVs and the Decider's decision without an Observer. Demographic controls include age, gender, political identity, party affiliation, risk preferences, voting behavior in the last presidential elections, education, and the measured difference in closeness towards a Trump lover and hater. Columns (1) and (2) include wave fixed effects.

deviations: tight vs polarized $p < 0.002$, loose vs polarized $p < 0.001$). The variance in the tight and loose environments is not significantly different.

Figure B.2: Implemented inequality in Part II — by political identity



B.3 Mechanism: Expected punishment

Figure B.3 breaks the expected punishment up by political identity. It shows that the pattern that Deciders anticipate the least punishment for a 50-50 split and the most for extreme allocations holds across political identities.

The aggregate punishment expectations mask substantial individual heterogeneity. While the majority of Deciders (54%) report U-shaped expected punishment, we also observe hump-shaped (12%), linear (11%), constant (7%), W-shaped (5%) and M-shaped (4%) expected punishment schedules. Figure B.4 shows examples for each type. The share of individuals with u-shaped expectations does not vary across treatments (χ^2-test, $p = 0.247$).

Figure B.3: Expected punishment by political identity

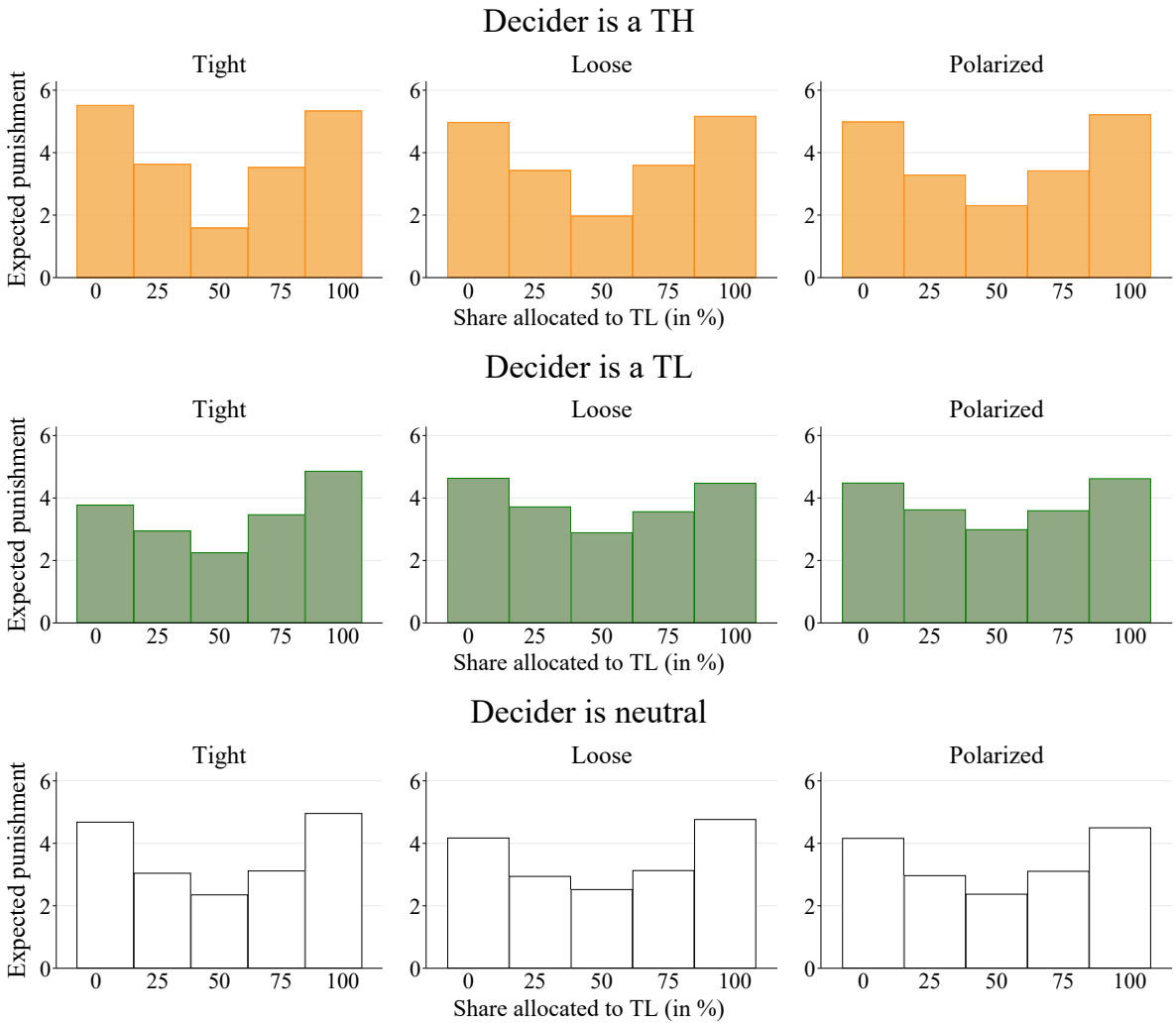


Figure B.4: Expected punishment types - Examples

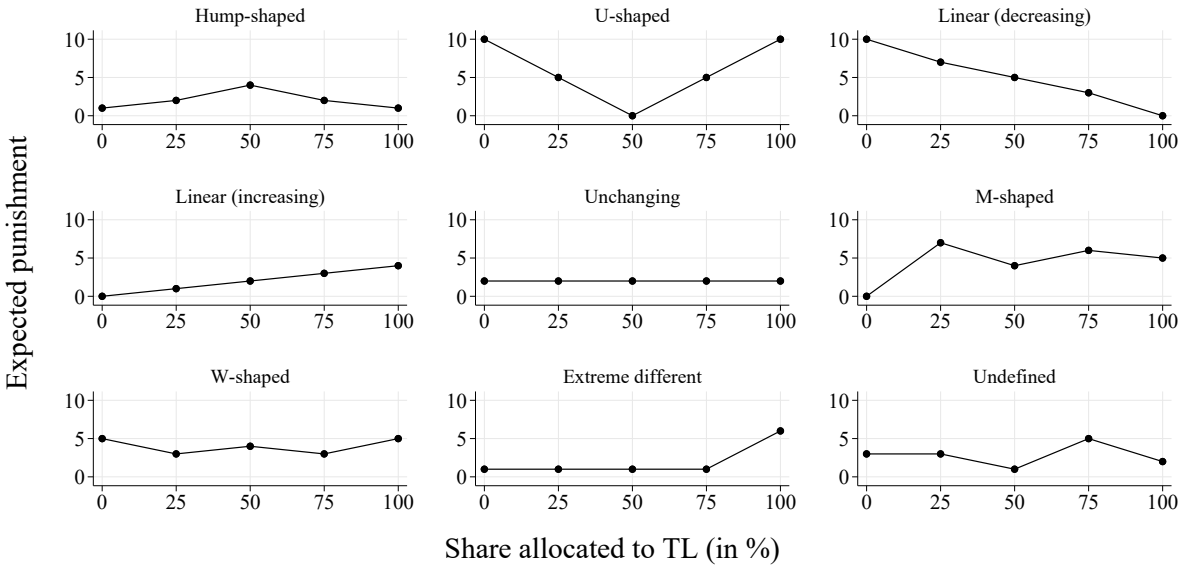


Table B.4 provides a robustness check to the analysis in the main paper. We see that the correlation between punishment expectations and implemented inequality holds when using a tobit specification. The coefficients become even slightly larger. Moreover, a significant relationship is present in both waves.

Table B.4: Punishment expectations as a correlate of implemented inequality - tobit

Dependent variable:	Implemented inequality in Part II					
	Full sample		**Wave 1**		**Wave 2**	
	(1)	(2)	(3)	(4)	(5)	(6)
Distributions (baseline=tight)						
Loose	0.10***		0.11**		0.10**	
	(0.03)		(0.05)		(0.04)	
Polarized	0.22***		0.24***		0.21***	
	(0.04)		(0.06)		(0.05)	
Steepness of punishment exp.		-0.03***		-0.03***		-0.03***
		(0.01)		(0.01)		(0.01)
Observations	1804	1804	901	901	903	903
Average inequality (tight)	0.22	0.22	0.22	0.22	0.22	0.22
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	No	No	No	No	No	No
Pseudo R^2	0.277	0.275	0.280	0.278	0.281	0.278

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.
Note. The table reports average marginal effects from tobit regressions (data is censored at 0 and 1). The dependent variable is the implemented inequality in Part II (as measured by the Gini). Distributions is a categorical variable that takes the value of 0 for the tight treatment (baseline category), 1 for loose, and 2 for polarized. The steepness of punishment expectations measures for each individual the average absolute difference between expectations for extreme versus equal allocations and varies between 0 and 10. Baseline controls include PVs and the Decider's decision without an Observer.

Table B.5 restricts the sample to participants with only U-shaped punishment expectations. Our results in the aggregate sample are robust to this restriction, confirming that steeper punishment expectations are related to a lower level of implemented inequality.

Table B.5: The effect of observability across environments and punishment expectations (restricted to people with U-shaped beliefs)

Dependent variable:	Implemented inequality (Gini) in Part II			
	(1)	(2)	(3)	(4)
Distributions (baseline=tight)				
Loose	0.05**		0.05**	
	(0.02)		(0.02)	
Polarized	0.11***		0.12***	
	(0.03)		(0.03)	
Steepness of punishment exp.		-0.01*		-0.01**
		(0.00)		(0.00)
Observations	821	821	813	813
Average inequality (tight)	0.12	0.12	0.12	0.12
Baseline controls	Yes	Yes	Yes	Yes
Demographic controls	No	No	Yes	Yes
R^2	0.32	0.30	0.33	0.31

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Results of OLS regressions. The dependent variable is the implemented inequality in Part II (as measured by the Gini). Distributions is a categorical variable that takes the value of 0 for the tight treatment (baseline category), 1 for loose, and 2 for polarized. The steepness of punishment expectations measures for each individual the average absolute difference between expectations for extreme versus equal allocations and varies between 0 and 10. The sample is restricted to Deciders with U- and hump-shaped punishment expectations. Baseline controls include PVs and the Decider's decision without an Observer. Demographic controls include age, gender, political identity, party affiliation, risk preferences, voting behavior in the last presidential elections, education, and the measured difference in closeness towards a Trump lover and hater. All specifications include wave fixed effects.

Our preregistered Hypothesis 2 states that U-shaped punishment expectations lead to less extreme choices. We define decisions as "more extreme" if the decision with Observers ($x_{i,2}$) is further away from the interior minimum (x_i^*) or maximum (x_i^{**}) than the decision without Observers ($x_{i,1}$).

$$E_i = \begin{cases} |x_i^* - x_{i,2}| - |x_i^* - x_{i,1}| = E_{i,u2} - E_{i,u1} & \text{if } f(x_i) \text{ is U-shaped} \\ |x_i^{**}i - x_{i,2}| - |x_i^{**} - x_{i,1}| = E_{i,h2} - E_{i,h1} & \text{if } f(x_i) \text{ is hump-shaped} \end{cases} \tag{8}$$

To test this hypothesis, we regress E_i on a dummy variable that takes the value of 0 if the distribution is hump-shaped and 1 if it is U-shaped (see Table B.6). As predicted, we find that the coefficient on U-shaped beliefs is negative and statistically significant for all specifications (except for column (6)). Note that against our prior, we see that the constant capturing the hump-shaped cases is also negative (even though insignificant in most specifications).

Table B.6: U-shaped punishment expectations and extreme choices

Dependent variable:	Extremeness of decisions in Part II relative to Part I					
	Full sample		**Wave 1**		**Wave 2**	
	(1)	(2)	(3)	(4)	(5)	(6)
U-shaped beliefs	-0.19***	-0.15**	-0.21**	-0.20**	-0.16**	-0.09
	(0.06)	(0.06)	(0.09)	(0.10)	(0.08)	(0.08)
Constant	-0.01	-0.29	-0.01	-0.41*	-0.03	-0.14
	(0.05)	(0.18)	(0.08)	(0.23)	(0.07)	(0.31)
Observations	1033	1019	541	533	492	486
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes	No	Yes
R^2	0.12	0.15	0.15	0.18	0.09	0.12

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Results of OLS regressions. The dependent variable is the extremeness of decisions in Part II relative to Part I as defined in Equation 8. U-shaped beliefs is a dummy variable that takes the value of 0 if the distribution of punishment expectations is hump-shaped and 1 if it is U-shaped. Baseline controls only include PVs, as the dependent variable measures extremeness relative to the Decider's decision without an Observer. Demographic controls include age, gender, political identity, party affiliation, risk preferences, voting behavior in the last presidential elections, education, and the measured closeness difference towards a Trump lover and hater.

Table B.7 uses the standard deviation of the shown treatment distribution as an instrument instead of the treatment dummies. Results are qualitatively similar to the analysis in the main paper.

Table B.7: The role of punishment expectations: instrumental variable approach - robustness: instrument = SD of treatment distribution

Dependent variables:	Steepness		Implemented inequality in Part II			
	First stage		**Reduced form**		**IV**	
	(1)	(2)	(3)	(4)	(5)	(6)
SD of treatment distribution	-0.83***	-0.79***	0.10***	0.10***		
	(0.15)	(0.15)	(0.02)	(0.02)		
Steepness of punishment exp.					-0.12***	-0.12***
					(0.03)	(0.03)
Constant	5.66***	6.87***	-0.11***	-0.26***	0.57***	0.60***
	(0.23)	(0.55)	(0.02)	(0.06)	(0.13)	(0.20)
Observations	1804	1779	1804	1779	1804	1779
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes	No	Yes
First-stage F	29.48	26.74			29.48	26.74
R^2	0.03	0.05	0.42	0.43		

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Column (1) and (2) report first stage results. The dependent variable is the steepness of punishment expectations. The steepness of punishment expectations measures for each individual the average absolute difference between expectations for extreme versus equal allocations and varies between 0 and 10. SD of treatment distribution measures the standard deviation of the Observers' distribution of PVs that was shown to Deciders in each treatment. We see that the instrument is relevant, F>10. Columns (3) and (4) present reduced form results, using the implemented inequality in Part II (as measured by the Gini) as the dependent variable. Columns (5) and (6) show the IV results, using the implemented inequality in Part II (as measured by the Gini) as the dependent variable and highlighting that beliefs causally affect implemented inequality. Baseline controls include PVs and the Decider's decision without an Observer. Demographic controls include age, gender, political identity, party affiliation, risk preferences, voting behavior in the last presidential elections, education, and the measured difference in closeness towards a Trump lover and hater. All specifications include wave fixed effects.

Table B.8 shows that our results in the main text are robust across waves. We confirm our result that PVs and closeness have a larger effect on the Decider's allocation decision in polarized than in tight environments.

Table B.8: The role of PVs and closeness across environments - by wave

Dependent variable:	Share allocated to Trump lover					
	Personal values			**Political identity**		
	(All)	(1)	(2)	(All)	(1)	(2)
Distributions (baseline=tight)						
Loose	-0.04	-0.01	-0.07	-0.00	0.02	-0.03
	(0.05)	(0.06)	(0.07)	(0.01)	(0.02)	(0.02)
Polarized	-0.17***	-0.15**	-0.18***	-0.02*	-0.03	-0.02
	(0.04)	(0.06)	(0.07)	(0.01)	(0.02)	(0.02)
PVs	0.16***	0.18***	0.15***			
	(0.02)	(0.02)	(0.02)			
Interactions						
Loose x PVs	0.02	0.01	0.03			
	(0.02)	(0.03)	(0.03)			
Polarized x PVs	0.07***	0.06**	0.08***			
	(0.02)	(0.03)	(0.03)			
Closeness				0.00	0.01	0.00
				(0.00)	(0.01)	(0.00)
Interactions						
Loose x Closeness				0.01*	0.01	0.01
				(0.00)	(0.01)	(0.01)
Polarized x Closeness				0.02***	0.02***	0.02***
				(0.00)	(0.01)	(0.01)
Constant	0.17***	0.13**	0.20***	0.51***	0.51***	0.51***
	(0.04)	(0.06)	(0.06)	(0.03)	(0.05)	(0.05)
Observations	1779	885	894	1779	885	894
Average share to TL (tight)	0.5	0.5	0.5	0.5	0.5	0.5
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.48	0.52	0.44	0.33	0.36	0.30

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Results of OLS regressions. The dependent variable is the amount the Decider allocates to the Trump lover. Distributions is a categorical variable that takes the value of 0 for the tight treatment (baseline category), 1 for loose, and 2 for polarized. PVs is a discrete variable that indicates what the Decider thinks should be allocated to the Trump lover and can take values between 0 and 1. Closeness refers to the reported closeness towards a Trump hater relative to a Trump lover and can take values between -6 and 6. The difference between loose x PVs/closeness and polarized x PVs/closeness is statistically significant in all columns (Wald tests, $p \leq 0.06$) except for column 5 ($p = 0.14$). The baseline control is the Decider's decision without an Observer. Demographic controls include age, gender, political identity, party affiliation, risk preferences, voting behavior in the last presidential elections, and education. Columns (1) and (2) include wave fixed effects.

B.4 Isolating the effect of punishment

In December 2025 we ran a pre-registered follow-up study to isolate whether punishment expectations are the key driver in explaining our experimental results. The additional data collection followed the same procedures as the main data collection. We recruited in total 900 Deciders on Prolific who did not previously participate in our study. 50% of participants self-identified as Trump lovers, 50% as Trump haters.

The follow-up experiment consists of three treatments. In line with the power analysis for the main experiment, we recruited 300 Deciders per treatment:

T1 A *control treatment*, where Deciders take an allocation decision between a Trump lover and a Trump hater without being observed. This treatment is akin to Part I in the main data collection.

T2 A *distribution treatment*, where Deciders see the distribution of Observers but take their decision in private.

T3 A *punishment treatment*, where Deciders see the distribution, but are then actually observed and potentially punished for their decision. This is akin to Part II in the main data collection.

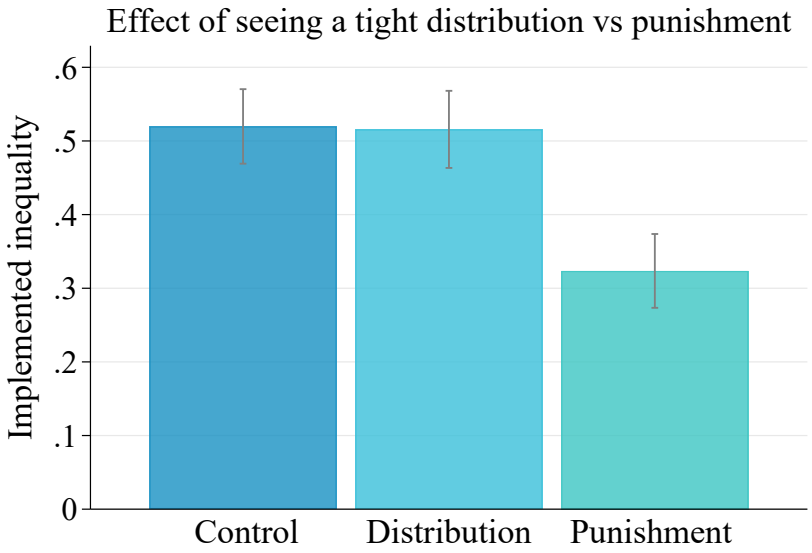
To make the information about Observers meaningful, we told Deciders in both T2 and T3 that some of them will be observed when taking their decision. They were then shown the distribution of potential Observers, knowing that if they are observed, their Observer will be drawn from this distribution. Then half of the Deciders were randomly assigned to being observed (T3) while the other half was randomly assigned to taking their decision in private (T2). Up to this point instructions were thus exactly identical in both conditions.

Since we see the strongest reduction in implemented inequality between Part I and II in our main study for the tight distribution we focus on this treatment for the follow-up data collection. All Deciders in T2 and T3 see the tight distribution as visualized in Section 3.2.

If we are concerned that the pure experience of seeing a distribution of how other people think about the allocation decision can affect behavior, we should expect a significant difference in the implemented inequality between T1 and T2. This could be for instance the case if Deciders are unsure about what to do and take the information about what others find appropriate as guidance for their own behavior. In line with our results in the main data collection this would imply a significant reduction in implemented inequality in T2 compared to T1. If our results were fully driven by the informational value of the observed distribution, we would then expect no difference between T2 and T3. If both the distribution itself and punishment expectations matter, we would expect T2 to have lower implemented inequality than T1 but higher inequality than T3.

By contrast, if participants only change their behavior in Part II because of punishment expectations, we should expect no difference between T1 and T2, but then a drop in the implemented inequality for T3 compared to both other treatments. As Figure B.5 shows our result support this interpretation. Behavior in T1 and T2 is literally identical, while we see a strong and statistically significant reduction in the implemented inequality when introducing the possibility of punishment.

Figure B.5: Average punishment by allocation decision



This conclusion is confirmed in Table B.9, where we regress the implemented inequality on treatment indicators. Our results thus provide strong evidence for the validity of the exclusion restriction, supporting our interpretation that punishment expectations are the key driver of our results in the main experiment.

Table B.9: The effect of seeing a distribution versus being observed

Dependent variable:	Implemented inequality	
	(1)	(2)
Treatment (baseline=control)		
Distribution	-0.004	0.004
	(0.037)	(0.035)
Punishment	-0.196***	-0.184***
	(0.036)	(0.034)
Constant	0.520***	0.073
	(0.026)	(0.134)
Observations	879	872
Average inequality (control)	0.52	0.52
Demographic controls	No	Yes
R^2	0.04	0.18

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.
Note. Results of OLS regressions. The dependent variable is the implemented inequality (as measured by the Gini) as the dependent variable. Treatment is a categorical variable that takes the value of 0 for the control treatment (baseline category), 1 for the distribution, and 2 for the punishment treatment. Demographic controls include age, gender, political identity, party affiliation, voting behavior in the last presidential elections, education, and the measured difference in closeness towards a Trump lover and hater.

B.5 Punishment behavior

Table B.10 regresses punishment decisions on all possible allocation decisions. The table highlights that Observers punish the 50-50 split the least, or any other allocation more. While this is true on aggregate, Figure B.6 shows that both Trump lovers and haters punish, allocating everything to their in-group similarly to the 50-50 split. By contrast, they exert strong punishment if a Decider allocates everything to the out-group. Putting both groups together then results in the overall U-shape of punishments for the full sample.

Table B.10: Punishment as a function of the allocation decision

Dependent variable:	Punishment decision					
	All		**Wave 1**		**Wave 2**	
	(1)	(2)	(3)	(4)	(5)	(6)
Allocations (baseline=50%)						
0% to TL	1.26***	1.26***	1.21***	1.21***	1.30***	1.30***
	(0.17)	(0.17)	(0.23)	(0.23)	(0.24)	(0.24)
25% to TL	0.68***	0.68***	0.56***	0.56***	0.81***	0.81***
	(0.12)	(0.12)	(0.16)	(0.16)	(0.17)	(0.17)
75% to TL	0.56***	0.56***	0.79***	0.79***	0.34**	0.34**
	(0.11)	(0.11)	(0.17)	(0.17)	(0.14)	(0.15)
100% to TL	1.25***	1.25***	1.64***	1.64***	0.86***	0.86***
	(0.17)	(0.17)	(0.26)	(0.26)	(0.23)	(0.23)
Constant	1.30***	1.11*	1.99***	2.28***	1.43***	0.19
	(0.14)	(0.64)	(0.16)	(0.79)	(0.15)	(1.19)
Observations	3005	3005	1500	1500	1505	1505
Baseline punishment (50% to TL)	1.83	1.83	1.99	1.99	1.67	1.67
N clusters	601	601	300	300	301	301
Demographic controls	No	Yes	No	Yes	No	Yes
R^2	0.04	0.07	0.03	0.10	0.03	0.06

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the individual level.
Note. Results of OLS regressions. The dependent variable is the amount the Observer takes away from the Decider (punishment) for each possible allocation decision. Demographic controls include PVs, age, gender, political identity, party affiliation, voting behavior in the last presidential elections, education, norm pluralism, and the measured difference in closeness towards a Trump lover and hater. Columns (1) and (2) include wave fixed effects.

Figure B.6: Average punishment by allocation decision

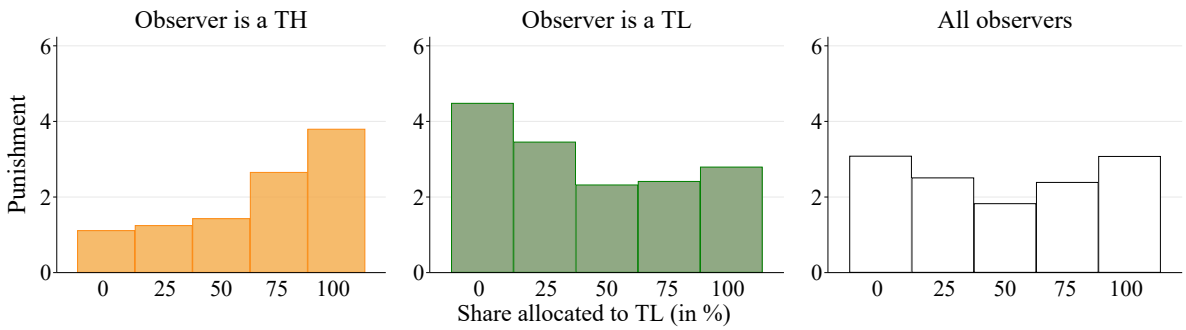


Table B.11 looks at the relationship between the distance of the allocation decision to the Ob-

server's own PV and the implemented punishment (see Figure B.7 on PVs). As the table shows, Observers punish more the further the Decider's decision lies from their personal values. The distance squared is insignificant in all specifications, suggesting on average a linear relationship between distance from PVs and punishment.

Table B.11: Punishment as a function of the distance between the Observer's PVs and the allocation decision

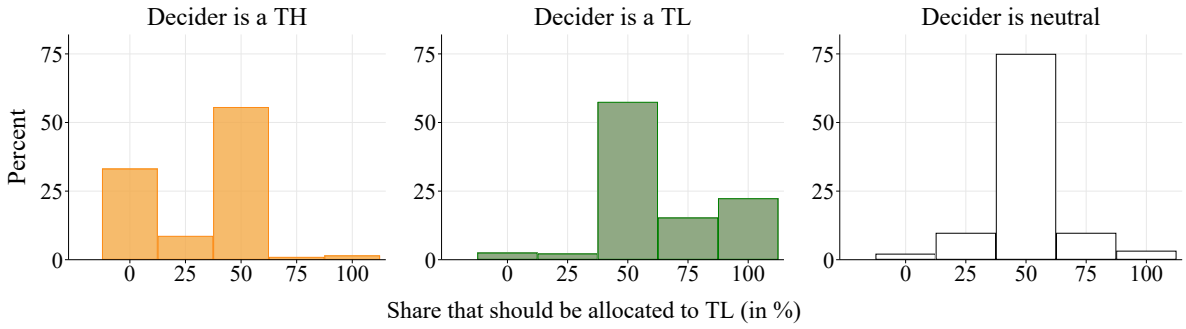
Dependent variable:	Punishment decision					
	All		**Wave 1**		**Wave 2**	
	(1)	(2)	(3)	(4)	(5)	(6)
Distance between allocation and PV	2.93***	2.88***	2.67***	2.66***	3.17***	3.08***
	(0.53)	(0.54)	(0.70)	(0.71)	(0.81)	(0.82)
Distance squared	0.55	0.64	0.59	0.61	0.54	0.70
	(0.64)	(0.65)	(0.82)	(0.82)	(1.01)	(1.02)
Constant	1.12***	0.89	1.23***	-0.04	1.50***	2.08***
	(0.15)	(0.63)	(0.16)	(1.18)	(0.17)	(0.77)
Observations	3005	3005	1505	1505	1500	1500
Average punishment	2.58	2.58	2.33	2.33	2.83	2.83
N clusters	601	601	301	301	300	300
Demographic controls	No	Yes	No	Yes	No	Yes
R^2	0.08	0.12	0.07	0.10	0.09	0.14

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the individual level.
Note. Results of OLS regressions. The dependent variable is the amount the Observer takes away from the Decider (punishment). *Distance between allocation and PV* measures the absolute distance between the allocation chosen by the Decider and the preferred allocation of the Observer (PV). Demographic controls include PVs, age, gender, political identity, party affiliation, voting behavior in the last presidential elections, education, norm pluralism, and the measured difference in closeness towards a Trump lover and hater. Columns (1) and (2) include wave fixed effects.

When looking at the shape of the punishment schedule, there is some interesting heterogeneity at the individual level. We classify Observers' punishment schedules as concave, convex, or linear relative to their own PV. To do so, we compare changes in punishment across decisions relative to the Observer's PV. In particular, we measure changes by taking the difference between the punishment for allocations that are further from the Decider's PVs and those who are closer. If punishments increase more steeply closer to the PV, we label the schedule concave; if they increase more steeply further from the PV, we label it convex; and if the slopes are equal, we classify it as linear. As an example if an Observer has a PV of giving everything to the Trump lover, they would be classified as linear (concave/convex) if the change in punishment when comparing allocating everything to the Trump lover and 75% to the Trump lover is the same as (larger/smaller than) the change in punishment when comparing 25% to the Trump lover and 0% to the Trump lover.

While we find that the majority of Observers (55%) increase their punishment of allocation decisions linearly with the distance to their own PVs, some Observers show a concave (22%) or a convex (23%) punishment schedule. While overall, this pattern is similar for Observers with different political identities, we find that relative to Trump haters, Trump lovers are 6.7% more likely to have concave (Wilcoxon rank sum test, $p = 0.08$) and 8.4% less likely to have linear punishment schedules ($p = 0.07$).

Figure B.7: Personal values of Observers



B.6 Comparison pre- and post-elections

In our setting partisan Deciders have to assess whether favoring their own political in-group puts them at the risk of social punishment. An interesting question is thus to explore the relative importance of political preferences and punishment expectations.

Table B.12 reports results from regressing the implemented inequality in Part II on political preferences — measured by the difference in reported closeness to a Trump lover and hater — and the steepness of punishment expectations. The table shows that both significantly predict allocation decisions. We then conduct a Shorrocks-Shapley decomposition based on R^2s and find that the relative importance of political preferences in the full sample is 72%, while the estimate for the relative importance of punishment expectations is 28%. Compared to Wave 1, the relative importance of punishment expectations in Wave 2 is substantially larger – it increases from only 18% to 43%. In line with the lower importance of political preferences after the elections, we find that participants report feeling closer to Trump haters in Wave 2 compared to Wave 1 (Wilcoxon rank sum test, $p = 0.02$), which is mainly driven by Trump lovers. This may highlight a strong salience of partisan divides in the run-up to the elections.

Table B.12: Relative importance of political preferences vs punishment expectations - by wave

Dependent variable:	Implemented inequality in Part II		
	All	Wave 1	Wave 2
Political preferences	0.04***	0.05***	0.03***
	(0.00)	(0.01)	(0.01)
Steepness of punishment exp.	-0.02***	-0.02***	-0.02***
	(0.00)	(0.00)	(0.00)
Constant	0.23***	0.20***	0.27***
	(0.03)	(0.03)	(0.03)
Observations	1804	901	903
R^2	0.07	0.09	0.06
Relative contributions (Shorrocks-Shapley decomposition):			
Political preferences	0.72	0.82	0.57
Punishment expectations	0.28	0.18	0.43

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Results of OLS regressions. The dependent variable is the implemented inequality in Part II (as measured by the Gini). Shorrocks-Shapley decomposition is based on the R^2 in the respective OLS regression and estimates the relative importance of each regressor. Political preferences are measured by the difference in reported closeness between a Trump lover and a hater. The steepness of punishment expectations measures for each individual the average absolute difference between expectations for extreme versus equal allocations and varies between 0 and 10. Column (1) includes wave fixed effects.

Table B.13 conducts the same analysis separated by treatments instead of waves. In line with Result 4, we find that political preferences have a higher relative importance in polarized and loose compared to tight environments.

Table B.13: Relative importance of political preferences vs punishment expectations - by treatment

Dependent variable:	Implemented inequality in Part II			
	All	Tight	Loose	Polarized
Political preferences	0.04***	0.04***	0.04***	0.05***
	(0.00)	(0.01)	(0.01)	(0.01)
Steepness of punishment exp.	-0.02***	-0.03***	-0.02***	-0.01**
	(0.00)	(0.00)	(0.01)	(0.01)
Constant	0.23***	0.24***	0.21***	0.23***
	(0.03)	(0.06)	(0.05)	(0.06)
Observations	1804	599	602	603
R^2	0.07	0.08	0.07	0.07
Relative contributions (Shorrocks-Shapley decomposition):				
Political preferences	0.72	0.49	0.76	0.93
Punishment expectations	0.28	0.51	0.24	0.07

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.
Note. Results of OLS regressions. The dependent variable is the implemented inequality in Part II (as measured by the Gini). Shorrocks-Shapley decomposition is based on the R^2 in the respective OLS regression and estimates the relative importance of each regressor. Political preferences are measured by the difference in reported closeness between a Trump lover and hater. All specifications include wave fixed effects.

Table B.14 regresses the implemented inequality in Part II on treatment indicators, a wave indicator, and their interaction. As the table shows, neither the Wave 2 indicator nor any of the interactions is statistically significant. In both waves, polarized environments lead to significantly more implemented inequality compared to tight environments.

Tables B.15 – B.17 explore the role of emotions in Wave 2. Table B.15 and B.16 focus on Deciders. The table shows that the more Trump lovers feel empowered through the elections and the more satisfied they are with their party's campaign, the more inequality they implement. Trump haters, by contrast, implement more inequality the more they feel afraid. Table B.16 shows that being more afraid also translates into higher punishment expectations for Deciders.

Finally, for Observers Table B.17 shows that Trump lovers who are more satisfied with their party's campaign tend to impose harsher punishments.

Table B.14: Treatment effects across data collection waves

Dependent variable:	Implemented inequality in Part II	
	(1)	(2)
Distributions (baseline=tight)		
Loose	0.04	0.03
	(0.02)	(0.02)
Polarized	0.10***	0.10***
	(0.03)	(0.03)
Wave 2	0.02	0.01
	(0.02)	(0.02)
Interactions		
Loose x Wave 2	0.01	0.01
	(0.03)	(0.03)
Polarized x Wave 2	0.01	0.00
	(0.04)	(0.04)
Constant	-0.03*	-0.18***
	(0.02)	(0.06)
Observations	1804	1779
Average inequality (tight)	0.22	0.22
Baseline controls	Yes	Yes
Demographic controls	No	Yes
R^2	0.42	0.43

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.
Note. Results of OLS regressions. The dependent variable is the implemented inequality in Part II (as measured by the Gini). Distributions is a categorical variable that takes the value of 0 for the tight treatment (baseline category), 1 for loose, and 2 for polarized. Wave 2 is a binary variable that takes the value 1 if data was collected in the second wave and 0 otherwise. Demographic controls include age, gender, risk preferences, party affiliation, voting behavior in the last presidential elections, education, and the measured difference in closeness towards a Trump lover and hater.

Table B.15: The role of emotions for Deciders' behavior

Dependent variables:	Implemented inequality (Gini) in Part I (Wave 2 only)					
	All		**THs**		**TLs**	
	(1)	(2)	(3)	(4)	(5)	(6)
Power & status	-0.05	-0.04	-0.06	-0.00	0.08	0.04
	(0.03)	(0.03)	(0.06)	(0.07)	(0.05)	(0.05)
Fear	0.08***	0.06***	0.11***	0.10***	-0.00	-0.00
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Satisfied with campaign	0.06***	0.03*	0.03	0.02	0.11***	0.08**
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Observations	903	894	325	323	363	359
Average inequality	0.39	0.39	0.54	0.54	0.39	0.39
Demographic controls	No	Yes	No	Yes	No	Yes
R^2	0.04	0.16	0.05	0.13	0.07	0.11

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.
Note. Results of OLS regressions. The dependent variable is the implemented inequality in Part I (as measured by the Gini). Demographic controls include age, gender, risk preferences, party affiliation, voting behavior in the last presidential elections, education, and the measured difference in closeness towards a Trump lover and hater. TH refers to Trump haters, TL to Trump lovers.

Table B.16: The role of emotions for Deciders' punishment expectations

Dependent variables:	Average expected punishment (Wave 2 only)					
	All		**THs**		**TLs**	
	(1)	(2)	(3)	(4)	(5)	(6)
Distributions (baseline=tight)						
Loose	0.09	0.11	0.01	-0.00	0.34	0.39
	(0.16)	(0.16)	(0.27)	(0.27)	(0.26)	(0.26)
Polarized	-0.05	-0.03	-0.18	-0.13	0.40	0.41
	(0.16)	(0.16)	(0.25)	(0.25)	(0.26)	(0.26)
Power & status	0.06	-0.01	-0.04	-0.20	0.08	0.12
	(0.14)	(0.16)	(0.28)	(0.29)	(0.27)	(0.27)
Fear	0.20**	0.22**	0.16	0.19	0.25	0.27
	(0.09)	(0.09)	(0.16)	(0.16)	(0.16)	(0.17)
Satisfied with campaign	-0.02	0.02	-0.08	-0.06	0.14	0.26*
	(0.07)	(0.07)	(0.11)	(0.11)	(0.14)	(0.15)
Observations	903	894	325	323	363	359
Average punishment expectation	3.70	3.70	3.90	3.90	3.63	3.63
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes	No	Yes
R^2	0.010	0.036	0.012	0.057	0.021	0.061

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Results of OLS regressions. The dependent variable is the average expected punishment by Deciders. Distributions is a categorical variable that takes the value of 0 for the tight treatment (baseline category), 1 for loose, and 2 for polarized. Demographic controls include age, gender, risk preferences, party affiliation, voting behavior in the last presidential elections, education, and the measured difference in closeness towards a Trump lover and hater. TH refers to Trump haters, TL to Trump lovers.

Table B.17: The role of emotions for Observers' punishment behavior

Dependent variable:	Average punishment decision (Wave 2 only)					
	All		**THs**		**TLs**	
	(1)	(2)	(3)	(4)	(5)	(6)
Power & status	0.33	0.20	0.34	0.74	-0.15	-0.37
	(0.26)	(0.30)	(0.55)	(0.59)	(0.43)	(0.49)
Fear	-0.12	-0.12	0.44	0.19	-0.24	-0.11
	(0.19)	(0.19)	(0.39)	(0.39)	(0.28)	(0.27)
Satisfied with campaign	0.30**	0.16	0.18	0.08	0.64**	0.65*
	(0.14)	(0.18)	(0.23)	(0.27)	(0.29)	(0.34)
Observations	300	300	104	104	151	151
Average punishment	2.83	2.83	2.15	2.15	3.51	3.51
Demographic controls	No	Yes	No	Yes	No	Yes
R^2	0.07	0.13	0.02	0.16	0.04	0.13

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Results of OLS regressions. The dependent variable is the average amount the Observer takes away from the Decider (punishment). Demographic controls include PVs, age, gender, voting behavior in the last presidential elections, party affiliation, education, norm pluralism, and the measured difference in closeness towards a Trump lover and hater. TH refers to Trump haters, TL to Trump lovers.

C Instructions

In this section we present instructions for our experimental studies. The instructions are largely identical across Wave 1 and 2, with small differences in the post-experimental survey regarding questions about the elections. We indicate in the instructions if a question was only present in one of the two waves. Comments starting with *"Note"* are annotations for readers and were not shown to participants.

C.1 Decider study

Consent

Project Title: Decision Study

Purpose of the Study: The purpose of this research project is to understand how people act in various social situations.

Procedures: The procedures involve answering a set of short questionnaires. The procedure will take approximately 10 minutes to complete.

Confidentiality: Your participation in this study is voluntary and confidential. We will do our best to keep your personal information confidential. Your name will not be recorded or associated with any of your answers. Any potential loss of confidentiality will be minimized by storing data in password-protected computer files and locked filing cabinets and storage areas. Only the researcher will have access to the data. If we write a report or article about this research project, your identity will be protected to the maximum extent possible.

Compensation: You will receive $1.30 for your participation in this study plus an additional bonus of up to $1.60, depending on your decisions and the decisions of other participants. You will be responsible for any taxes assessed on the compensation. In studies like ours, there are sometimes bots that compromise the quality of our data. The study therefore, includes several CAPTCHA that need to be solved correctly for participants to be able to start the study and be compensated.

Right to Withdraw and Questions: Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

By checking the "I agree" box below, you indicate that you are at least 18 years of age and you voluntarily agree to participate in this research study.

- I agree to participate

- I do NOT agree to participate

Welcome

Thank you very much for participating in this study! This study consists of two parts and a questionnaire. **In both parts you will face a situation in which you will be matched with two real other people.** Your task is to decide how those two people should be paid. After the study, we will randomly select one participant and their decision will be implemented. This means that your decision can have **real consequences** and determine what each of them earns.

On the next page we will describe this situation to you in more detail.

Before you proceed, please answer the sports test. The test is simple, when asked for your favorite sport you must enter the word clear in the text box below.

Based on the text you read above, what favorite sport have you been asked to enter in the text box below?

_____

ID Please insert your **Prolific ID** here. _____

Before we begin the survey, we would like to ask you a few questions to understand your political preferences.

What do you feel towards the person above?

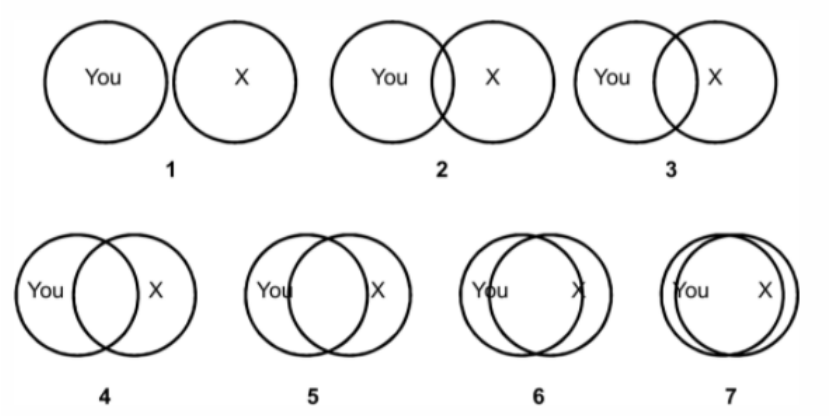
- Extreme hate
- Moderate hate
- Indifferent
- Moderate love
- Extreme love

Please explain your choice in a few sentences: _____

Note: The next question is only shown if "Indifferent" is chosen above:
We understand you do not have strong feelings regarding Trump. But if you had to choose, would you consider yourself closer to being a

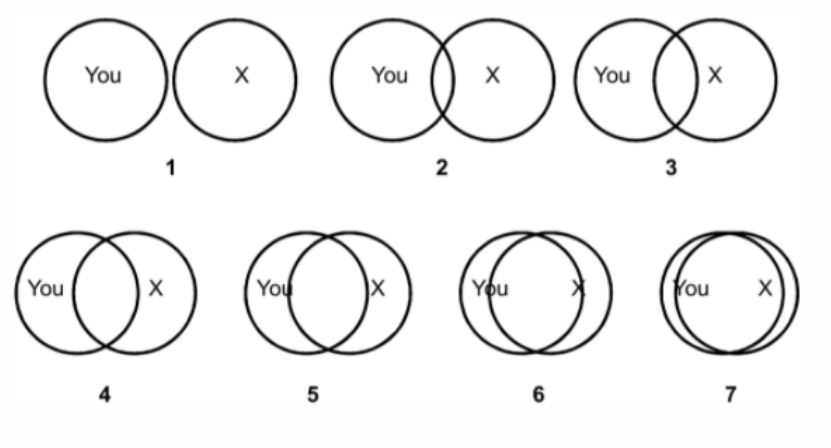
- Trump lover
- Trump hater

Note: The order of the next two questions is randomized.
Think about a person who indicated they love Donald Trump on the previous scale.
Please select the pair of circles that best describes your closeness with this person. The circle with X represents that person.



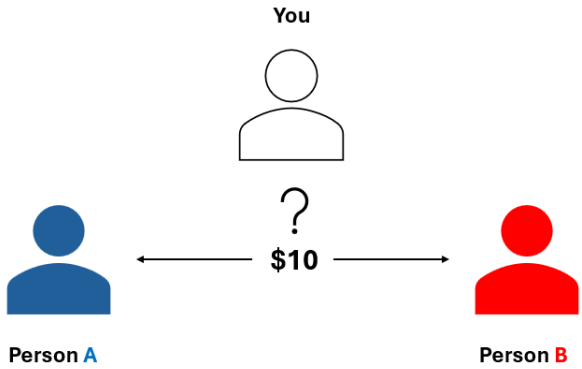
Thank you for answering this question. Next, we will ask you about a person with the opposite political views.

Think about a person who indicated they hate Donald Trump on the previous scale.
Please select the pair of circles that best describes your closeness with this person. The circle with X represents that person.



Part 1
We now ask you to make a choice that has consequences for a real-life situation. We recruited two individuals, **person A and person B**, and asked them about their opinion on Trump using

the scale you just saw. **Person A is a declared Trump Hater. Person B is a declared Trump Lover.** They were also told that they could receive an extra bonus that is determined by a third person. This money is independent of the compensation for participating in the survey and not linked to any work assignment.



You are this third person. You have $10 and can freely decide how you want to distribute them between the two people.
Your decision

You can now decide how you want to allocate the **$10** between **person A (Trump Hater) and person B (Trump Lover).** Please indicate your decision below:
The amount allocated to the **Trump Hater** is indicated in **blue** and the amount allocated to the **Trump lover** is indicated in **red**.

- **$10** for **A**, **$0** for **B**
- **$7.50** for **A**, **$2.50** for **B**
- **$5** for **A**, **$5** for **B**
- **$2.50** for **A**, **$7.50** for **B**
- **$0** for **A**, **$10** for **B**

Perceptions

In the following, we ask you a few questions about your perception of this allocation situation. According to **your own opinion** and independent of the opinion of others, what is the most appropriate allocation between person **A (Trump Hater) and B (Trump Lover)**?

Appropriate here means what you personally consider to be "correct" or "moral".

- **$10** for **A**, **$0** for **B**
- **$7.50** for **A**, **$2.50** for **B**

- **$5** for **A**, **$5** for **B**
- **$2.50** for **A**, **$7.50** for **B**
- **$0** for **A**, **$10** for **B**

Note: The order of the following two questions is randomized.
What do you think most people defining themselves as a **Trump Hater** consider as the most appropriate allocation between **person A (Trump Hater) and B (Trump Lover)**?

If your guess is correct and matches the most common answer among **Trump Haters** in our study, you will receive an **additional bonus of $0.20**.

Appropriate here means what they personally consider to be "correct" or "moral".

- **$10** for **A**, **$0** for **B**
- **$7.50** for **A**, **$2.50** for **B**
- **$5** for **A**, **$5** for **B**
- **$2.50** for **A**, **$7.50** for **B**
- **$0** for **A**, **$10** for **B**

What do you think most people defining themselves as a **Trump Lover** consider as the most appropriate allocation between **person A (Trump Hater) and B (Trump Lover)**?

If your guess is correct and matches the most common answer among **Trump Lovers** in our study, you will receive an **additional bonus of $0.20**.

Appropriate here means what they personally consider to be "correct" or "moral".

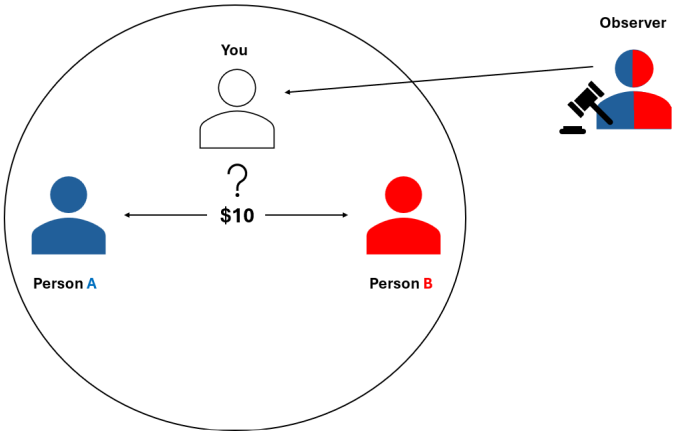
- **$10** for **A**, **$0** for **B**
- **$7.50** for **A**, **$2.50** for **B**
- **$5** for **A**, **$5** for **B**
- **$2.50** for **A**, **$7.50** for **B**
- **$0** for **A**, **$10** for **B**

Part 2

In Part 2 we ask you to make the same decision for a new pair of people. Again, **person A** is a **Trump Hater**, while **person B** is a **Trump Lover**. However, this time another person is observing your choice. We refer to that person as the **Observer**.

In addition to your show-up fee, you have been assigned an endowment of **100 points** to participate in this study.

The **Observer can take away up to 100 points** from you based on your allocation decision between person **A** and **B**. The less the Observer approves of your choice, the more will be taken away. To take away 10 points from you, the Observer has to pay 3.



At the end of the experiment, the points you have will determine your bonus with 100 points = $1.

With 50% probability, the Observer will be a Trump Hater.
With 50% probability, the Observer will be a Trump Lover.

This means that you are equally likely to be observed by either a **Trump Hater** or **Trump Lover**.

Below we ask you some questions about the **Observer**. Recall that the **Observer** is another person participating in this experiment.

Note: Participants can only move on once they get the control questions right. 1) Will the Observer know how you allocate the $10 between person **A** and **B**?

- Yes
- No

2) What can the **Observer** do?

- Nothing, they just see your decision
- See your decision and take away up to 100 points from your bonus Give you up to 100 points as an additional bonus

3) What is the **Observer's** political orientation?

- Trump Hater

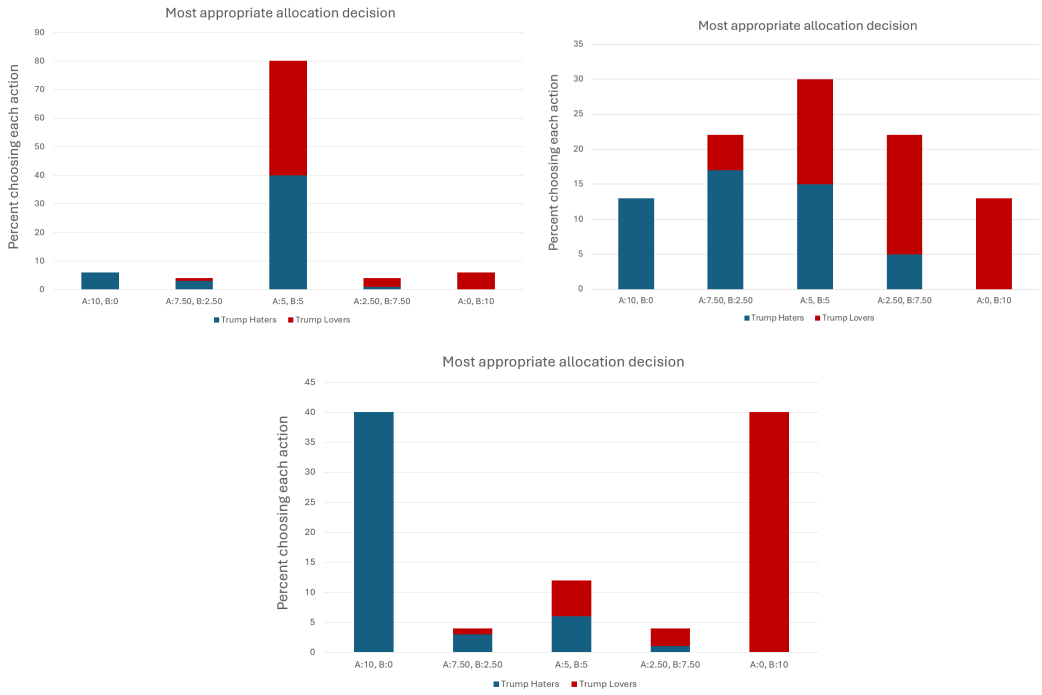
- Trump Lover
- Equally likely to be a Trump Hater or Lover

We asked a group made up of both Trump Lovers and Trump Haters to tell us which allocation between person A and B they consider to be the most appropriate. From their answers we constructed different sub-groups. The graph below shows the answers in one randomly selected sub-group.

In the picture, a higher column means that a higher share of individuals in the group think this is the most appropriate allocation. The **red** and **blue** portions of a column show the political views of those who say that allocation is the most appropriate.

Your Observer will be randomly drawn from this group.

Note: Participants randomly see one of the three graphs depending on whether they are in the tight, loose, or polarized condition.



1) What does the distribution shown above tell you?

- It shows which decision is the **most approved** one by a group of Observers.
- It shows which decision is the **most disapproved** one by a group of Observers.

2) What has the Observer who can punish you for your decision to do with the distribution?

- The Observer has the same opinion as most people in the graph
- The Observer is drawn randomly from all people shown in the graph

3) Do you think the Observer will punish you more for

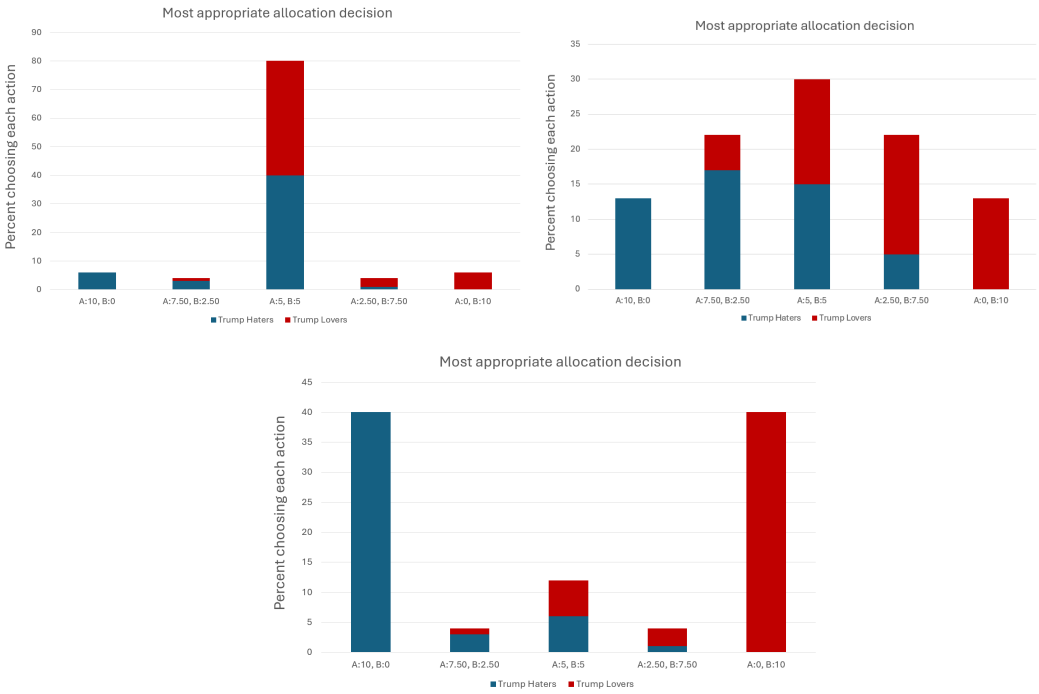
- Decisions they approve of
- Decisions they disapprove of

Your decision

Your task is again to decide how to allocate the $10 between person **A (Trump Hater)** and person **B (Trump Lover)**.

Recall that your **Observer** can take away up to 100 points from your bonus depending on how much they dislike your decision. Your Observer is a member of a group with the following perceptions of appropriateness.

Note: Participants randomly see one of the three graphs depending on whether they are in the tight, loose, or polarized condition.



Please indicate below how you want to allocate the $10 between person **A (Trump Hater)** and **B (Trump Lover)**:

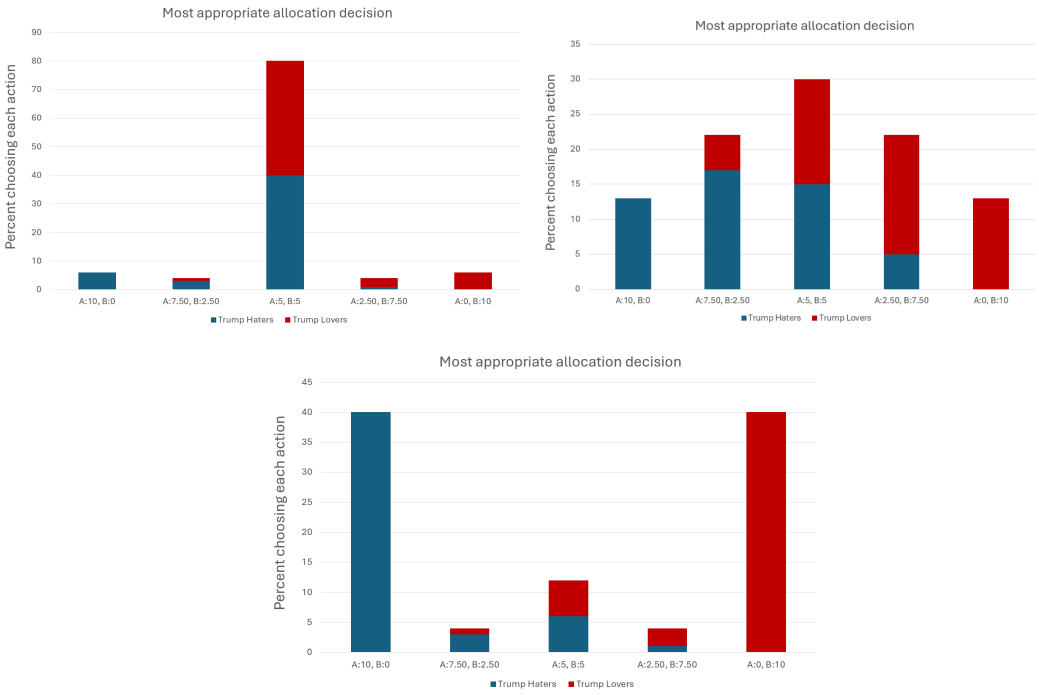
The amount allocated to the **Trump Hater** is indicated in **blue** and the amount allocated to the **Trump Lover** is indicated in **red**.

- **$10** for **A**, **$0** for **B**
- **$7.50** for **A**, **$2.50** for **B**
- **$5** for **A**, **$5** for **B**

- **$2.50** for **A**, **$7.50** for **B**
- **$0** for **A**, **$10** for **B**

Your beliefs

Think about **10 Observers** drawn from the distribution below. The graph shows again which allocation Observers see as **most appropriate.**
Note: Participants randomly see one of the three graphs depending on whether they are in the tight, loose, or polarized condition.



How many points do you expect them to take away from you on average for different allocations?

For instance, if you believe that all Observers would punish 5-5 maximally, the average would be 100. If you think noone would punish it, it would be 0. If you think that 5 Observers punish it maximally, while 5 do not punish, the average would be 50.

We will randomly choose one of the allocation scenarios. If your guess is correct for that scenario and matches the average punishment decision of 10 randomly drawn Observers in our study, you will receive an **additional bonus of $0.20**.

For each allocation between person **A (Trump Hater)** and **B (Trump Lover)**, please indicate below how many points you believe the Observers would take away from you (rounded to the closest 10):

$10 for **A**, **$0** for **B**	0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
$7.50 for **A**, **$2.50** for **B**	0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
$5 for **A**, **$5** for **B**	0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
$2.50 for **A**, **$7.50** for **B**	0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
$0 for **A**, **$10** for **B**	0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100

Note: Participants choose their answer (0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100) through a drop down menue.

Questionnaire

1) What is your age? _____

2) Which gender do you identify with?

- Female
- Male
- Other

What is the highest level of schooling you completed?

- No formal qualifications
- Secondary school
- University/ college degree
- Prefer not to say

4) Please tell me, in general, how willing or unwilling you are to take risks on a scale from 0 to 10. _____

5) Which political party do you identify most with?

- Republicans
- Neither
- Democrats

Note: This question was only asked in Wave 1.

6) Whom did you vote for in the last general elections?

- Trump
- Did not vote
- Biden

Note: This question was only asked in Wave 1.

7) Whom do you intend to vote for in the 2024 general elections?

- Definitely Trump
- Probably Trump
- Both equally likely
- Probably Harris
- Definitely Harris

Note: This question was only asked in Wave 1.
8) Whom do you think is going to win the 2024 general elections? *Wave 1 only*

- Trump
- Will not vote
- Harris

Note: This question was only asked in Wave 2.
6) Whom did you vote for in the last general elections?

- Trump
- Did not vote
- Harris

Note: This question was only asked in Wave 2.
7) How satisfied are you with the campaigning strategy of the party you voted for?

- Very satisfied
- Satisfied
- Neither satisfied nor disappointed
- Disappointed
- Very disappointed

Note: This question was only asked in Wave 2. The order of the next two questions was randomized.
8) We'd like you to rate how you feel towards some groups on a scale of 0 to 100. Zero means very unfavorable and 100 means very favorable. Fifty means you do not feel favorable or unfavorable. How would you rate your feeling toward **Democrats**?_____

Note: This question was only asked in Wave 2.
9) We'd like you to rate how you feel towards some groups on a scale of 0 to 100. Zero means very unfavorable and 100 means very favorable. Fifty means you do not feel favorable or unfavorable. How would you rate your feeling toward **Republicans**?_____

Note: This question was only asked in Wave 2.
10) Do you agree or disagree that you can expect fair treatment from government authorities?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly disagree

Note: This question was only asked in Wave 2.
11) Do you agree or disagree with the following: America is heading toward the end of democracy, where free and fair elections will no longer occur.

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly disagree

Note: This question was only asked in Wave 2.
12) Compared to before the election, I feel...

- more powerful as an individual
- the same
- less powerful as an individual
- more secure in my position in society
- the same
- less secure in my position in society

Note: This question was only asked in Wave 2.
13) Compared to other groups, my group's position in society has...

- Improved
- Stayed the same
- Gotten worse

Note: This question was only asked in Wave 2.
14) My group's standing in society relative to other groups has...

- Improved
- Stayed the same

- Gotten worse

Note: This question was only asked in Wave 2.
15) My group's status relative to other groups has...

- Improved
- Stayed the same
- Gotten worse

9)/16) How many surveys on political issues have you participated in in the last 3 months? Please enter a number._____

10)/17) What do you think is this survey about? _____

Thanks a lot for participating in this survey! If you have any feedback for us you can write it here:_____

C.2 Observer study

Consent

Project Title: Decision Study

Purpose of the Study: The purpose of this research project is to understand how people act in various social situations.

Procedures: The procedures involve answering a set of short questionnaires. The procedure will take approximately 15 minutes to complete.

Confidentiality: Your participation in this study is voluntary and confidential. We will do our best to keep your personal information confidential. Your name will not be recorded or associated with any of your answers. Any potential loss of confidentiality will be minimized by storing data in password-protected computer files and locked filing cabinets and storage areas. Only the researcher will have access to the data. If we write a report or article about this research project, your identity will be protected to the maximum extent possible.

Compensation: You will receive $1.90 for your participation in this study plus an additional bonus of up to $1.80 depending on your decisions and the decisions of other participants. You will be responsible for any taxes assessed on the compensation. In studies like ours, there are sometimes bots that compromise the quality of our data. The study therefore includes several captchas that need to be solved correctly for participants to be able to start the study and be compensated.

Right to Withdraw and Questions: Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

By checking the "I agree" box below, you indicate that you are at least 18 years of age and you voluntarily agree to participate in this research study.

- I agree to participate
- I do NOT agree to participate

Welcome

Thank you very much for participating in this study! In addition to your show-up fee, you have been assigned an endowment of **100 points** to participate in this study. Points will change depending on your decision and those of other participants. At the end of the experiment, **the points you have will determine your bonus with 100 points = \$1. In this study you will observe another participant who has been matched with two real people.** The task of the other participant is to decide how to allocate a bonus between the two people. We will ask your opinion on this decision and give you the chance to act upon it.

On the next page we will describe this situation to you in more detail.

Before you proceed, please answer the sports test. The test is simple, when asked for your favorite sport you must enter the word clear in the text box below.

Based on the text you read above, what favorite sport have you been asked to enter in the text box below?

_____

ID Please insert your **Prolific ID** here. _____

Before we begin the survey, we would like to ask you a few questions to understand your political preferences.

What do you feel towards the person above?

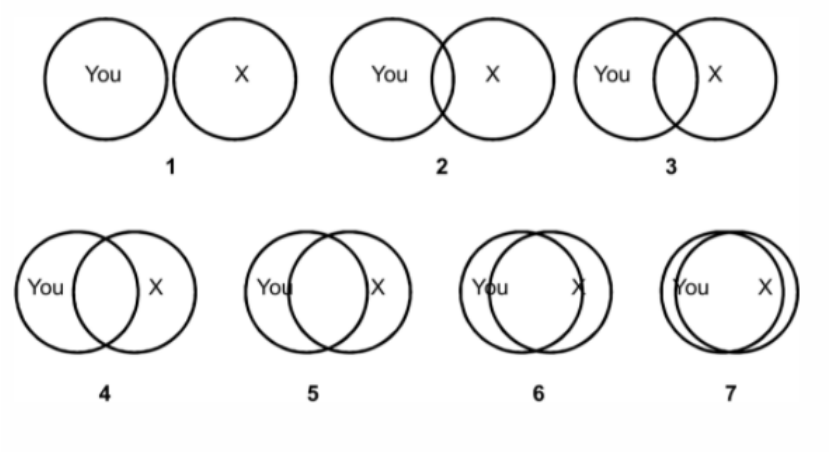
- Extreme hate
- Moderate hate
- Indifferent
- Moderate love
- Extreme love

Please explain your choice in a few sentences: _____

Note: The next question is only shown if "Indifferent" is chosen above:
We understand you do not have strong feelings regarding Trump. But if you had to choose, would you consider yourself closer to being a

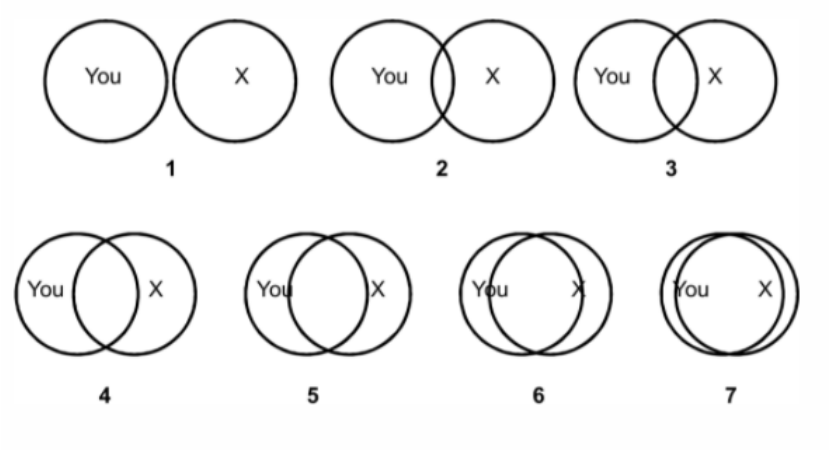
- Trump lover
- Trump hater

Note: The order of the next two questions is randomized.
Think about a person who indicated they love Donald Trump on the previous scale.
Please select the pair of circles that best describes your closeness with this person. The circle with X represents that person.



Thank you for answering this question. Next, we will ask you about a person with the opposite political views.

Think about a person who indicated they hate Donald Trump on the previous scale.
Please select the pair of circles that best describes your closeness with this person. The circle with X represents that person.



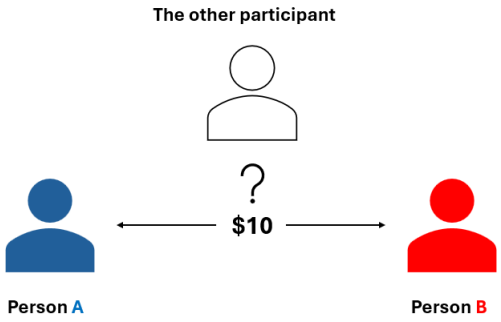
The setting

We recruited two individuals, **person A and person B**, and asked them their opinion on Trump using the scale you just saw. **Person A is a declared Trump Hater. Person B is a declared Trump Lover.**

They were also told that they could receive an extra bonus that is determined by another participant. This money is independent of the compensation for participating in the survey and not linked to any work assignment.

The other participant's decision

The other participant was explained this situation, given **$10** and could freely decide how to allocate the $10 between person **A (Trump Hater)** and person **B (Trump Lover)**.

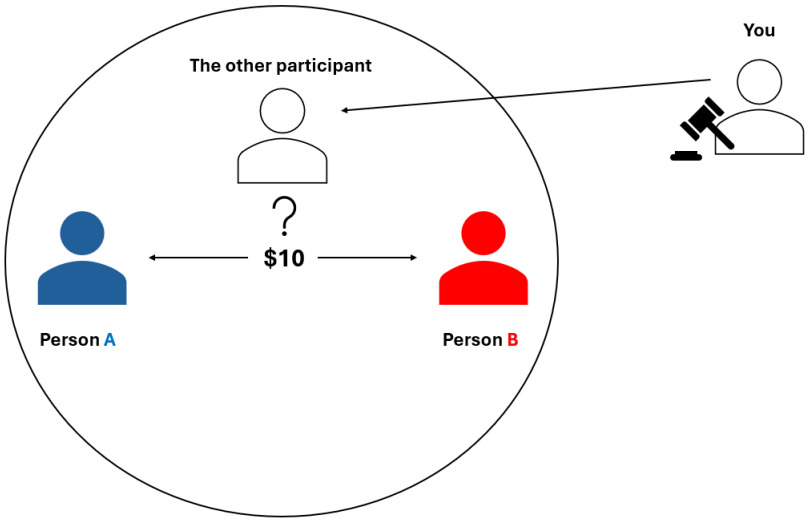


The possible allocation options were:

- **$10** for **A**, **$0** for **B**
- **$7.50** for **A**, **$2.50** for **B**
- **$5** for **A**, **$5** for **B**
- **$2.50** for **A**, **$7.50** for **B**
- **$0** for **A**, **$10** for **B**

Your task

The other participant was allocated a bonus of 100 points for making the allocation decision on top of a show-up fee. You can **take away any of the bonus points** (in multiples of 10) if you wish to.



You have to pay in order to take money away from the other participant. **To take away 10 points you have to pay 3 points from your own endowment**. Thus, to take away the maximum amount of 100 points, you have to pay 30 points. If you decide not to take away any amount, you have to pay 0 points.

Below we will ask you a few questions about your task.

1) What can you do as an Observer?

- Nothing, you just observe the allocation decision
- You observe the allocation decision and can take away up to 100 bonus points from the other participant
- You can give up to 100 bonus points to the other participant

2) What do you have to do in order to take away bonus points from the other participant?

- Anonymously decide on a number between 0 and 100
- Send a message to the other participant explaining your choice

3) How much do you need to pay to take away 10 bonus points from the other participant?

- 10 of your own bonus points
- 3 of your own bonus points

Your decision

You can condition your decision on the allocation decision of the other participant between person **A (Trump Hater)** and **B (Trump Lover)**. If the other participant chose allocation x, your decision for scenario x will determine how much money is taken from the other participant.

To determine your conditional choice, **please tell us how much you want to take away from the other participant if they chose:**
Note: Participants choose their answer (0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100) through a drop down menue.

$10 for **A**, **$0** for **B**	0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
$7.50 for **A**, **$2.50** for **B**	0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
$5 for **A**, **$5** for **B**	0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
$2.50 for **A**, **$7.50** for **B**	0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
$0 for **A**, **$10** for **B**	0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100

Perceptions

In the following, we ask you a few questions about your perception of this allocation situation. According to **your own opinion** and independent of the opinion of others, what is the most appropriate allocation between person **A (Trump Hater) and B (Trump Lover)**?

Appropriate here means what you personally consider to be "correct" or "moral".

- **$10** for **A**, **$0** for **B**
- **$7.50** for **A**, **$2.50** for **B**
- **$5** for **A**, **$5** for **B**
- **$2.50** for **A**, **$7.50** for **B**
- **$0** for **A**, **$10** for **B**

Note: The order of the following two questions is randomized.
What do you think most people defining themselves as a **Trump Hater** consider as the most appropriate allocation between **person A (Trump Hater) and B (Trump Lover)**?

If your guess is correct and matches the most common answer among **Trump Haters** in our study, you will receive an **additional bonus of $0.20**.

Appropriate here means what they personally consider to be "correct" or "moral".

- **$10** for **A**, **$0** for **B**
- **$7.50** for **A**, **$2.50** for **B**
- **$5** for **A**, **$5** for **B**

- **$2.50** for **A**, **$7.50** for **B**
- **$0** for **A**, **$10** for **B**

What do you think most people defining themselves as a **Trump Lover** consider as the most appropriate allocation between **person A (Trump Hater) and B (Trump Lover)**?

If your guess is correct and matches the most common answer among **Trump Lovers** in our study, you will receive an **additional bonus of $0.20**.

Appropriate here means what they personally consider to be "correct" or "moral".

- **$10** for **A**, **$0** for **B**
- **$7.50** for **A**, **$2.50** for **B**
- **$5** for **A**, **$5** for **B**
- **$2.50** for **A**, **$7.50** for **B**
- **$0** for **A**, **$10** for **B**

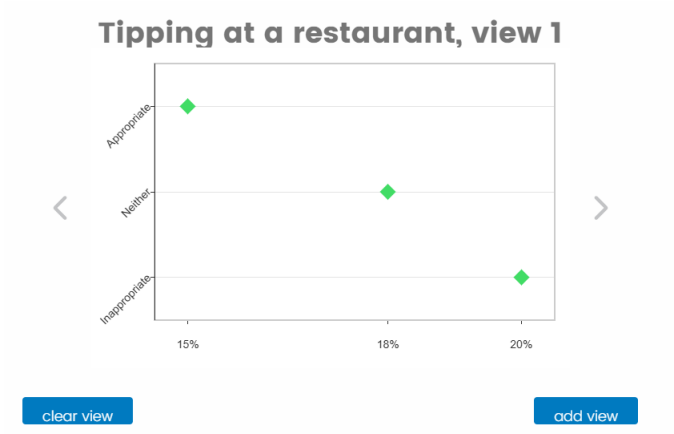
Different views

Before, we asked you about the **most appropriate allocation** according to a person defining themselves as a **Trump Hater** or **Trump Lover**.

People can however, also have different views about the other allocation options. You can represent someone's view with a drawing like the one below. To show you how this works, we first give you an **example of a different decision scenario**.

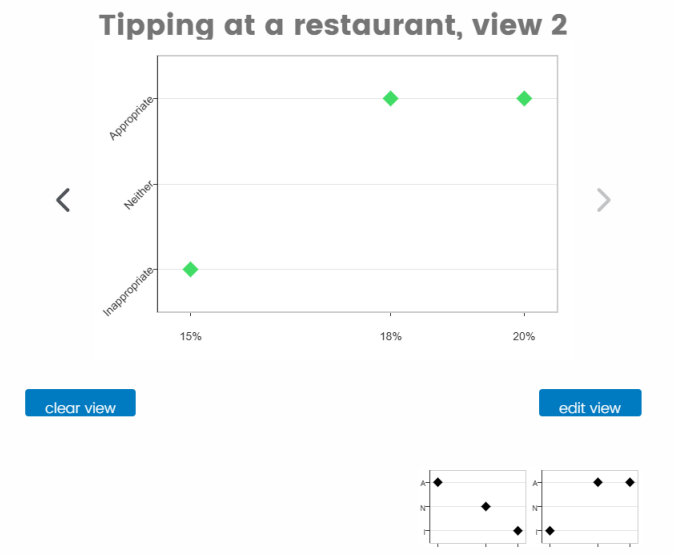
Imagine you eat out in a restaurant and have the option to either tip 15%, 18% or 20%. Now, some people might say that 15% is an appropriate amount to tip, and might consider other options as less appropriate, for example 18% as neither appropriate nor inappropriate, and 20% as inappropriate. Please click on the chart below to rate 15% as Appropriate, 18% as Neither, and 20% as Inappropriate. When you have drawn the view correctly, the drawing will turn **green**. Then, click **add view** to store this view, it will appear under the chart.

Note: This is a screenshot showing the elicitation tool if participants already answered the question correctly.



Of course different people may have different views. Now that you've drawn this view, you can add others. For example, other people might say that 15% is Inappropriate, and 18% and 20% are Appropriate. Please **add** this **second view** in the same way as the first one.

Note: This is a screenshot showing the elicitation tool if participants already answered the question correctly.



You cannot add any more views for this example. Please click to continue to the next page.

Now let's get back to the scenario where somebody has to allocate money between person **A** and **B**. We will first ask you to draw all the different views you think **Trump Lovers** or **Trump Haters** have. Then we will ask you to guess how many (out of 100 total) hold the first view, the second view, and so on.

Bonus

If your answers match what the others say you can earn an additional bonus of $0.20.

If you draw the same views that other respondents draw and your guess about how many people hold each view is the same as the average guess of other respondents then your chance of earning the bonus is 100 out of 100 (i.e. you'll earn the bonus for sure). The more your response differs from the average response made by others, the less likely you are to earn the bonus. In other words, **this task is designed to guess what other participants will respond: the closer you are, the better!**

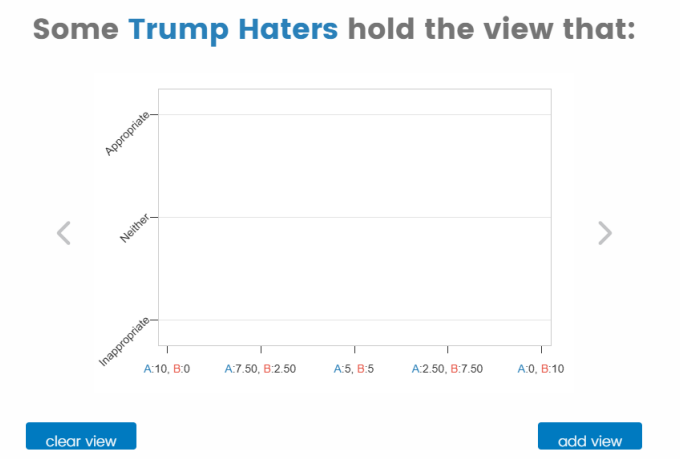
If you want to have a closer look at how your earnings will be calculated click **here**.

*Note: If participants click on **here** they see the following text:*

Reasons you are less likely to earn the bonus: You drew a view that was not drawn by anyone else. For example, if you guess that 10 people hold a view that no one else drew, then your chances of winning are reduced by 10 out of 100. You failed to draw a view that other respondents drew. For example, if you don't draw a view that others guessed 5 people out 100 hold, then your chances of winning are reduced by 5/100 scaled by the share of others who drew this view. Your guess about how many people hold a view is too high or too low compared to the average guess. For example, if you drew a view that others also drew, and you guessed that 20 people out of 100 hold that view, but other people guess that 40 out of 100 hold that view, then your chances of winning are reduced by 20 out of 100.

Note: We randomize whether we ask participants first about Trump Lovers or Haters. How many different views do you think can be found among **Trump Haters** on the allocation decision between **person A (Trump Hater)** and **B (Trump Lover)?** What are they?

You can make between **1 and 10 drawings** to represent different views that you think exist. All actions need to have a rating.



Please indicate **how many out of 100 Trump Haters** you think hold each view.

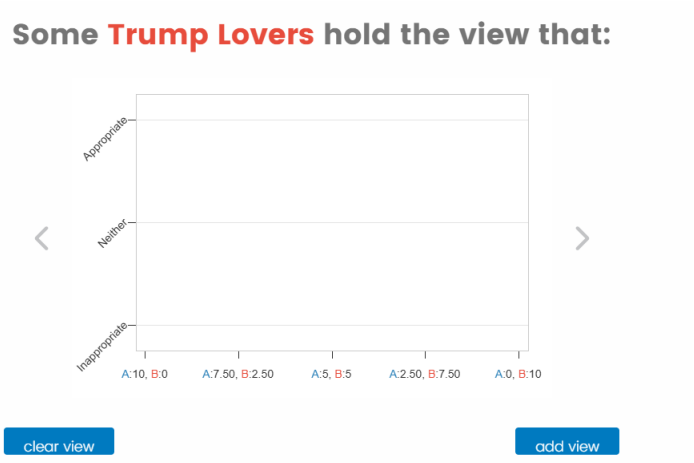
They must sum to 100.

Note: This is a screenshot showing the elicitation tool taking two hypothetical drawings as an example.

Thank you for drawing the views. Next, we will ask you about a person with the opposite political views. Again, if your answers match what the others say you can earn **an additional bonus of $0.20.**

How many different views do you think can be found among **Trump Lovers** on the allocation decision between **person A (Trump Hater)** and **B (Trump Lover)?** What are they?

You can make between **1 and 10 drawings** to represent different views that you think exist. All actions need to have a rating.



Please indicate **how many out of 100 Trump Lovers** you think hold each view.
They must sum to 100.

Note: This is a screenshot showing the elicitation tool taking two hypothetical drawings as an example.

Questionnaire

Note: The questionnaire is identical to the one in the Decider study.

C.3 Recipient study

Consent

Project Title: Decision Study

Purpose of the Study: The purpose of this research project is to understand how people act in various social situations.

Procedures: The procedures involve answering a short survey that takes about 3 minutes. Afterwards, we will ask another participant to decide whether you receive an additional bonus for this task.

Confidentiality: Your participation in this study is voluntary and confidential. We will do our best to keep your personal information confidential. Your name will not be recorded or associated with any of your answers. Any potential loss of confidentiality will be minimized by storing data in password-protected computer files and locked filing cabinets and storage areas. Only the researcher will have access to the data. If we write a report or article about this research project, your identity will be protected to the maximum extent possible.

Compensation: You will receive $0.50 for your participation in this study plus a bonus between $0 and $10 depending on the decision of the other participant. You will be responsible for any taxes assessed on the compensation. In studies like ours, there are sometimes bots that compromise the quality of our data. The study therefore includes several captchas that need to be solved correctly for participants to be able to start the study and be compensated.

Right to Withdraw and Questions: Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

By checking the "I agree" box below, you indicate that you are at least 18 years of age and you voluntarily agree to participate in this research study.

- I agree to participate
- I do NOT agree to participate

Welcome

Thank you very much for participating in this study! This study consists of a short questionnaire on your political views. Upon completion you will receive $0.50 for your participation.

We will ask another participant to divide $10 between you and another person. Depending on the decision of this other participant you will receive an additional bonus between $0 and $10.

Before you proceed, please answer the sports test. The test is simple, when asked for your favorite sport you must enter the word clear in the text box below.

Based on the text you read above, what favorite sport have you been asked to enter in the text box below?

_____

ID Please insert your **Prolific ID** here. _____

First, we would like to ask you a few questions to understand your political preferences.

What do you feel towards the person above?

- Extreme hate
- Moderate hate
- Indifferent
- Moderate love
- Extreme love

Please explain your choice in a few sentences: _____

Note: The next question is only shown if "Indifferent" is chosen above:
We understand you do not have strong feelings regarding Trump. But if you had to choose, would you consider yourself closer to being a

- Trump lover
- Trump hater

Thanks a lot for participating in this survey! If you have any feedback for us you can write it here: _____

Table C.1: The effect of seeing a distribution versus being punished

Dependent variable:	Implemented inequality	
	(1)	(2)
Treatment (baseline=control)		
Distribution	-0.004	0.004
	(0.037)	(0.035)
Punishment	-0.196***	-0.184***
	(0.036)	(0.034)
Constant	0.520***	0.073
	(0.026)	(0.134)
Observations	879	872
Average inequality (control)	0.52	0.52
Demographic controls	No	Yes
R^2	0.04	0.18

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note. Results of OLS regressions. The dependent variable is the inequality implemented by Deciders (as measured by the Gini), the case with observability. Treatment is a categorical variable that takes the value of 0 for the control (no distribution, no punishment), 1 for the treatment where participants see a distribution and 2 for the treatment with punishment. Demographic controls include age, gender, political identity, party affiliation, risk preferences, voting behavior in the last presidential elections, education and the measured difference in closeness towards a Trump lover and hater.