Pluralism Breeds Tolerance

Folco Panizza[a], Eugen Dimant[b,c], Erik O. Kimbrough[d], Alexander Vostroknutov[e]

[a] *IMT School for Advanced Studies Lucca*
[b] *University of Pennsylvania, Center for Social Norms and Behavioral Dynamics*
[c] *CESifo*
[d] *Smith Institute for Political Economy and Philosophy, Chapman University*
[e] *Maastricht University*

Abstract

This study introduces the Norm-Drawing Task, a novel approach to measure pluralism, or the coexistence of multiple normative beliefs in a given context. By combining established methods, we identify heterogeneous normative beliefs in well-known economic games, challenging the typical assumption of a single prevailing norm. Moreover, we are able to link norm multiplicity to actual behavior. In a well-powered and pre-registered experiment, we observe that participants who perceive multiple norms are more tolerant and punish norm violations less frequently and less severely than those who perceive a singular norm. In a second experiment, we show that this effect is causal. Comparing two elicitation protocols that allow participants to report multiple norms versus requiring them to report a single norm, we find that punishment is lower in the former than in the latter: pluralism breeds tolerance. The implications of our study are broad, indicating that societal structures and policy decisions could be influenced by the underlying multiplicity of norms. Moreover, the Norm-Drawing Task, for which we provide a ready-made software implementation, offers a new avenue for exploring important societal issues like the perception of minority groups and the dynamics of polarization.

Keywords: Distribution, Norm Elicitation, Social Norms, Norm Uncertainty, Tight and Loose Norms

JEL: C9, D01, D9

*This research was kindly funded by International Foundation for Research in Experimental Economics (grant no. 31-1-19); pre-registration: socialscienceregistry.org/trials/12012; software implementation: osf.io/7hc9v
Email addresses: folco.panizza@imtlucca.it (Folco Panizza), edimant@sas.upenn.edu (Eugen Dimant), ekimbrou@chapman.edu (Erik O. Kimbrough), a.vostroknutov@maastrichtuniversity.nl (Alexander Vostroknutov)

This version: November 22, 2024

1. Introduction

There is a broad consensus among social scientists that norms permeate our lives, facilitating cooperation and coordination, distinguishing groups from one another, and shaping our beliefs and attitudes (see e.g., Sherif, 1936; Cialdini and Trost, 1998; Bicchieri, 2006; Henrich, 2017; Bursztyn et al., 2020). Norms are conceptualized as mutually reinforcing patterns of beliefs and behavior, and they are often divided into an injunctive aspect, which refers to what ought to be done, and a descriptive aspect, which describes what is actually done by people in a group. These two elements of a norm are essential because the most popular accounts of norm-driven behavior assume that people are motivated to conform to norms that they believe others will follow and that they believe others expect them to follow in kind. In these accounts, changes to either aspect of beliefs can have important consequences for behavior.

To better understand the role of norms in shaping behavior, social scientists have thus recently worked to develop methods for measuring them in all their aspects. While measuring behavior is straightforward, measuring beliefs is more difficult – all the more so when these beliefs are about non-verifiable objects like normative judgments regarding what ought to be done. Nevertheless, techniques for measuring shared normative beliefs have been fruitfully employed to improve our understanding of individual decision-making, strategic decision-making, and political expression and behavior, to name just a few examples (Krupka and Weber, 2013; Bicchieri and Xiao, 2009; Barr et al., 2018; Dimant, 2019; Bicchieri et al., 2022; Pickup et al., 2021, 2023). Participants read a description of a set of possible actions and are asked to identify either the normatively best action or to estimate the relative appropriateness of each action, under incentives to match others' responses. Knowing which actions are seen as normatively appealing and which are seen as normatively unappealing helps researchers understand the social forces that shape decisions.

In some cases, such as the widely studied Dictator Game, beliefs about the relative appropriateness of the different possible allocations have been shown to be strikingly consistent across a wide range of participant pools and countries. On average, respondents seem to agree on the equal outcome being the most appropriate, with appropriateness decaying as we move away from equality. Moreover, average normative beliefs are also an important predictor of how dictators' decisions vary with context (Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016, 2018; Bicchieri et al., 2022; Aycinena et al., 2022a,b).

Crucially, however, existing norm measurement methods often implicitly assume that there is only one norm, or that identifying the most common norm is sufficient to capture the social expectations and behaviors within a group. While this is a legitimate assumption under certain conditions, in many circumstances measuring a single norm may obfuscate the reality of *pluralism*: the coexistence of many normative beliefs in the same environment that can possibly contradict each other.

One case in which there may be multiple norms is when those norms are tied to distinctive identity groups (Akerlof and Kranton, 2000; Chang et al., 2019; Krupka et al., 2022; Groenendyk et al., 2023). For example, during the COVID-19 pandemic, attitudes towards mask-wearing were closely related to political identity (Lang et al., 2021). Some people believed that the norm was to wear a mask in public places, whereas others thought that the norm was to keep the pre-COVID status quo. This is a typical example of norm multiplicity where it was clear to everyone that there were two norms followed by different groups of people. This situation created a lot of confusion and normative disagreement and did not promulgate healthy behaviors (Gelfand et al., 2021a,b; Dimant et al., 2022a).

Even in the abstract setting of the Dictator Game, questions have been raised about whether the stable *average* pattern observed in norm elicitation measures actually masks several underlying heterogeneous types (Kimbrough et al., 2022). For example, whereas norm measurements consistently identify a norm of equality in the game, a minority of people may recognize a norm of generosity instead.[1] Most relevantly, both theory and evidence have been accumulating that multiple norms can coexist even in the absence of sharp identity boundaries, that various events can trigger transitions between them, and that their co-existence can have important consequences for behavior (e.g., Centola et al., 2018; Fromell et al., 2021; Dimant and Gesche, 2023; Dimant et al., 2023; Merguei et al., 2022). These consequences can be dire: some have argued that norm multiplicity can create normative disagreements or deterioration of norms that can hurt cooperation or even lead to a conflict or war. But they can also be salutary: mutual recognition of differences of viewpoint can breed tolerance.

In light of these issues, it is essential to 1) develop methods for identifying norm multiplicity, in order to fully understand the normative landscape against which individuals evaluate and adjust their social behavior, and 2) begin to augment theories of norm-driven behavior to explicitly incorporate the effects of norm multiplicity on decision-making. In this paper, we make progress towards both of these goals.

First, we propose a new Norm-Drawing Task that allows us to explicitly elicit beliefs about the nature and relative popularity of multiple norms in choice problems. In a two-step procedure, we first elicit a collection of norms (instead of one) using an interface inspired by Crosetto and De Haan (2023), in which participants draw each norm by tracing a distinct path through the action space assigning relative appropriateness to each action, and then we

[1] Kimbrough et al. (2022) use latent class models to identify 5 underlying normative belief types of which the average pattern in the Dictator Game is composed; they find that though the average normative belief is strikingly similar to that seen in other settings, there is some evidence of heterogeneity in the underlying composition of types across two distinct cultural settings. As the authors point out, given that norms were elicited using the coordination game method due to Krupka and Weber which incentivizes people to report the most common normative belief in the reference group, the fact that participants report heterogeneous beliefs in their study can only arise if either participants make mistakes or there is genuine uncertainty about which of many beliefs is actually most common in the population.

apply the Belief Elicitation by Superimposition Approach or BESA (Fragiadakis et al., 2019) to the resulting collection of norms to reveal participants' beliefs about how commonly held is each norm in the reference group. Combining the two procedures generates an incentive-compatible method that not only elicits multiple norms in a given situation but also lets participants determine endogenously how many such norms they think there are. The Norm-Drawing Task generates a new type of data in which we observe the set of norms that each participant believes apply in a given situation. To illustrate how these beliefs can influence behavior, we propose a simple theoretical argument built on the ideas developed in Kimbrough and Vostroknutov (2023a,b) and Merguei et al. (2022) that relate the multiplicity of norms to the punishment of norm violators. Specifically, we argue that norm multiplicity should make the punishment of norm violations smaller and less frequent compared to the case of single-norm beliefs. This is because, in multiple-norm environments, the violations of which norms should be punished are unclear. Theoretically, weaker punishment can arise either because punishers who are aware of multiple norms will punish according to the "cheapest" norm for punishment, or because the "expected" norm that averages the normative beliefs of the population will generally be "looser" than any single norm and so violations will be seen as meriting less punishment. As noted above, this implied reduction in punishment can be seen as a mechanism by which pluralism breeds tolerance.

Our theoretical argument is substantiated by two experiments. In our first experiment, we test the Norm-Drawing Task by eliciting beliefs in a 2×2 design with one dimension corresponding to the type of game (a Dictator Game or a three-player Allocation Game similar to those in Engelmann and Strobel, 2004) and another dimension corresponding to presence or absence of a passive charity player, on the assumption that a charity will be seen as more deserving of resources than another survey respondent (Eckel et al., 2023). At the same time, we collect data on punishment choices within the same participants to connect their normative beliefs to their behavior and to test if beliefs about norm multiplicity are associated with reduced punishment.

We hypothesized that there would be fewer norms reported in the charity versions of both games and fewer norms reported in the Dictator Game than in the Allocation game. Thus, we expected that punishment would also be lower in the Dictator game and in the charity versions of both games. However, we find that in both the Dictator Game and the Allocation Game participants draw 4 to 5 norms on average, and this is unaffected by the introduction of the charity recipient. Thus, in our first study, we were unable to generate exogenous variation in perceptions of multiplicity. In hindsight, this is perhaps unsurprising given recent evidence of norm multiplicity in the Dictator Game and the observation that different people have different attitudes toward charitable giving (DellaVigna et al., 2012). Nevertheless, we find that the punishment patterns indeed fit our theory: participants who draw only one norm in the Norm-Drawing Task punish significantly more harshly and also

more often than those who draw multiple norms.

In a second experiment, we set out to assess whether this observation is a mere selection effect or whether pluralism *causally* increases tolerance. Thus, we attempt to shock perceptions of norm multiplicity more directly, by comparing the punishment behavior of participants who reported normative beliefs in our Norm-Drawing Task—where they can report pluralistic beliefs—to the punishment behavior of participants who report normative beliefs in the Krupka-Weber task where they can only report a single norm. We hypothesized that punishment would be lower for participants exposed to the Norm-Drawing task than for those exposed to the Krupka-Weber task. Our data from the second experiment both replicate the correlations identified in the first experiment and provide strong support for a causal effect of pluralism on tolerance. We find that participants in the Norm-Drawing treatment are ∼20pp *less* likely to punish a third-party than their counterparts in the Krupka-Weber treatment.

Our results can have far-reaching implications. For example, Gelfand et al. (2011) observe that human societies can be divided into two general classes: tight and loose societies that are different in terms of harshness of punishment of norm violations. This has consequences for the quality of institutions, welfare, and many other important economic indicators (Gelfand et al., 2024). We suggest that it will be helpful to distinguish differences in tightness and looseness that arise from norm multiplicity from more basic differences in the stridency with which normative beliefs are held: tight societies have a single, widely shared norm with correspondingly harsh enforcement, which leads to a more structured society with orderly rules of behavior; loose societies may simply hold their norms less stridently or they may reflect coexistence of a multiplicity of norms, either of which leads to less enforcement and more tolerance of difference. Looseness derived from multiplicity has the potential to yield very different dynamics over time, with opinion converging around a single normative viewpoint at one extreme or with gradual deterioration of norms due to lack of enforcement (Henrich, 2017), potentially resulting in more chaotic institutions. We suggest that our method could be used as a complement to existing measures of tightness and looseness to help further understand the mechanisms that link this cultural dimension to behavior.

The Norm-Drawing Task can also be used for other purposes. For example, detecting pluralistic ignorance (Bicchieri, 2006; Smerdon et al., 2020) or polarization (Iyengar et al., 2019; Bursztyn et al., 2020; Levy, 2021; Dimant, 2023b; Panizza et al., 2024), as well as monitoring the dynamics of norms (Gelfand et al., 2024), including minority views that otherwise would go undetected. More broadly, the task allows us to reconstruct the normative landscape that participants believe they live in: the number and kind of norms that they express in certain contexts can signal what sort of attitudes and heterogeneity in beliefs they expect to deal with in reality and what consequences for behavior this might have.

The paper is structured as follows. In Section 2, we review the existing norm elicitation techniques and argue why they cannot adequately capture norm multiplicity. In Section 3, we present the Norm-Drawing Task and propose the theoretical argument about the data it generates. In Section 4, we describe our experimental design and report our correlational findings. In Section 5, we report the results of a second experiment that provides direct causal evidence linking punishment and multiplicity. Section 6, interprets our findings, proposes the future directions of research, and concludes.

2. Why Existing Norm Elicitation Techniques Cannot Adequately Capture Multiplicity

Several approaches to eliciting normative beliefs exist in the literature (Charness et al., 2021). If the appropriateness of actions can be inferred from observing the behavior of others (descriptive norms), norms can be measured by incentivizing people to predict behavior. For instance, Bicchieri and Xiao (2009) conducted an experiment where participants were asked to predict how many others would divide money approximately equally in a Dictator Game at the start of the study. Participants received a fixed payment for correct predictions, although incentives could also be based on various mechanisms, such as scoring rules or the interval method. However, there is not always a one-to-one mapping between norms and behavior, as one may split resources equally in a Dictator Game because e.g. the norm prescribes either generosity or equality.

Thus, researchers have developed methods to directly measure injunctive normative beliefs, which are understood as shared, second-order beliefs about what is or is not appropriate. Injunctive beliefs cannot be independently and objectively verified; thus, eliciting these beliefs is trickier. One approach, due to Bicchieri and Xiao (2009), adopts a two-step process for eliciting beliefs that combines a non-incentivized self-report of what the participant personally believes is normatively best in step one with incentives to guess the personal beliefs reported by others in step two. For the Dictator Game, participants are asked whether dictators should split the money equally, and then they are incentivized to accurately estimate how many other participants answered "Yes" to the first question. This approach works if participants report their personal beliefs honestly, but given that personal beliefs are internally held and unverifiable, there is no assurance that respondents will reply honestly, especially concerning sensitive topics.

An elicitation method that is incentive-compatible, yet does not rely on objectively verifiable information was introduced by Krupka and Weber (2013). This method elicits social norms by incentivizing participants to report shared beliefs via coordination games. Specifically, participants are asked to predict how others would rate an action in terms of its social appropriateness (e.g., inappropriate, neutral, appropriate). To win a prize, participants must match their guess about the appropriateness rating to the most common

guess made by others. They are paid only if their guess matches the modal guess. This method works if, to solve the coordination problem, participants rely on common knowledge of shared beliefs about what is socially (in)appropriate as a focal point.

A significant limitation of both the two-step and coordination-game approaches is that they incentivize participants to identify the single most common normative belief, which assumes away multiplicity and uncertainty by design.[2] This could lead to false consensus or an incomplete picture of the norms that, in fact, shape peoples' choices. For example, in contexts where multiple norms are present, respondents may find it difficult to agree on what actions are appropriate, potentially leading to imprecise norm estimates. As we argue, it is not obvious to expect a single norm in many environments, nor that norms will not change, for instance, because minority views take hold. Even in contexts where apparently only one norm exists, diverging perspectives might lurk beneath the surface and influence behavior.

One natural possibility is to adapt these methods to measure uncertainty about the relative appropriateness of each action. Several methods exist that could be adapted to measure a distribution of injunctive norms (e.g., Fragiadakis et al., 2019; Peeters and Wolk, 2019; Dimant et al., 2022b); we present possible uses of these methods in Appendix A.1. However, while eliciting distributions over the appropriateness of actions can reveal information about uncertainty, it does not allow one to see the underlying norms out of which the distribution is composed. Think, for instance, about the actions of blowing one's nose and of sniffing in public: if beliefs are estimated over each action separately, there are multiple sets of norms that could in principle generate the same uncertainty. Figure 1 illustrates. Consider a world in which there are two norms as drawn in panel (a), each of which is shared by 50% of the population vs. a world with two norms as depicted in panel (b), each of which is shared by 50% of the population. Existing measures of normative uncertainty would yield panel (c) in both cases and could not distinguish which pattern of underlying normative belief generated the measure.

In other words, such approaches cannot tell us which norms (mappings from actions to relative social appropriateness) are being mixed, since the distributions just sum them all together. The key conceptual issue is that a norm is not just characterized by the most appropriate action but rather a set of beliefs about the *relative* appropriateness of all possible actions. Without knowing how each point in the distribution of beliefs about one action is connected to beliefs about other available actions, a method that measures distributions cannot reveal what kind of norm(s) exist in that context.

[2] A notable exception is when norms are known to be linked to distinct reference groups; then, eliciting norms separately for each reference group can reveal multiplicity. However, identifying reference groups is not always possible, and it is not necessarily true that all members of a group share the same norms (see also Panizza et al., 2024).

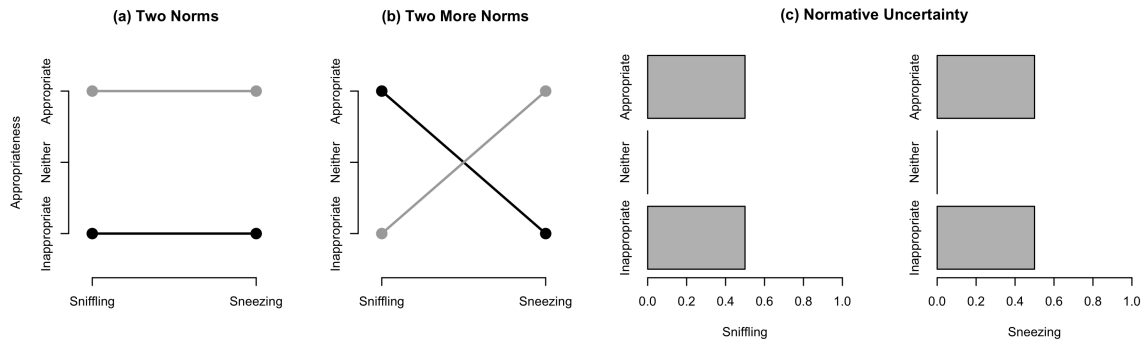

Figure 1: Normative Uncertainty and Multiplicity. Panels (a) and (b) depict two different patterns of normative belief, which, if each held by 1/2 of a population, would generate the same pattern of normative uncertainty depicted in panel (c).

Thus, what is needed is a measure that can capture beliefs about the full distribution of injunctive norms, i.e., if a single injunctive norm reflects second-order beliefs about the relative appropriateness of all actions in a context, then to capture multiplicity, we need to be able to measure third-order beliefs about the distribution of such second-order beliefs. Here, we introduce a new method for measuring such beliefs that uses the Belief Elicitation by Superimposition Approach (BESA for short, Fragiadakis et al., 2019; Aycinena et al., 2022b) to generalize the coordination game method due to Krupka and Weber (2013) from second-order to third-order beliefs.

3. The Norm-Drawing Task

As noted above, we conceive of a norm as an evaluation of the relative appropriateness of a set of feasible actions. That is, a norm can represented as a single path through a figure like that depicted Figure 1a. The objective of the Norm-Drawing task is thus to incentive compatibly reveal the set of all such norms that participants perceive to be held by members of a reference group. If there is a single, unique norm, then participants should have incentive to only draw that norm, but if there are multiple norms, participants should have incentive to draw all of them and accurately estimate the share of the population holding each set of beliefs.

The elicitation method follows a two-step procedure: first, participants define the set of norms that exist in the scenario at hand (we refer to these as "views" in the instructions), then they guess how these norms are distributed in the population. We will describe the second step first for illustrative purposes.

To guess the frequency of different norms in the population, participants allocate 100 tokens across different options that correspond to different norms drawn in the first step. For example, if there are two norms about the appropriate way deal with allergy symptoms in public (e.g. sniff, don't sneeze and neither sniff nor sneeze), a participant allocates

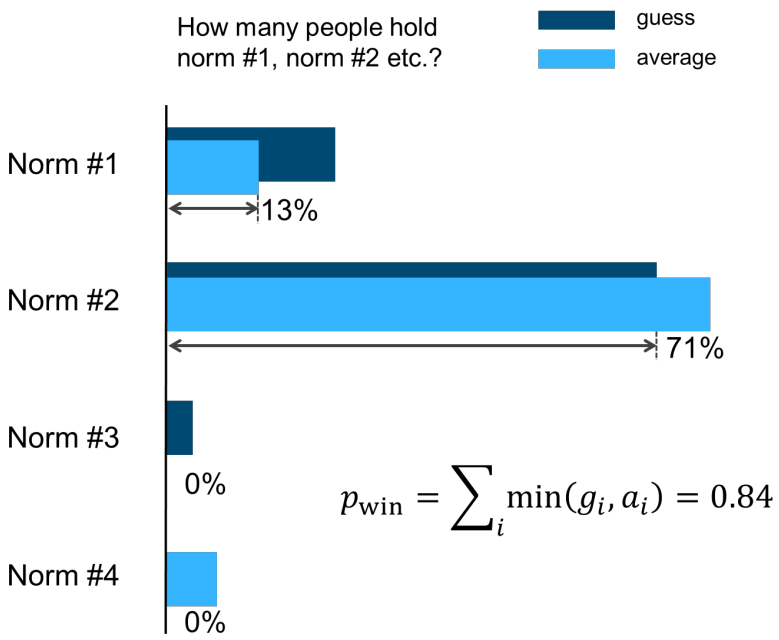

Figure 2: An example of token distribution. Participants make their guesses about the set of norms, then their distribution is compared to the average distribution guessed by other participants. Their probability of winning the prize for this task is equal to the sum of the minimum between their guess and the average guess for each norm. If the participant draws a norm nobody else drew, or if the participant fails to draw a norm that others drew, the minimum for that norm is equal to 0. In this example, the sum (and probability of winning) equals 84%.

tokens to guess how many people (out of 100) hold each of the two norms. To make this procedure incentive-compatible we use BESA described above (Fragiadakis et al., 2019). After participants allocate tokens among the options, their allocation is superimposed on the average distribution. They are paid according to a lottery that pays $\$\pi$ with a probability that is proportional to the overlap between the two (Figure 2) and \$0 otherwise. For n norms, we compute the overlap as $\sum_{i=1}^{n} \min\{p_i, E[p_i]\}$, where p_i is the number of tokens allocated to norm i and $E[p_i]$ is the average number of tokens assigned to norm i by all participants in the reference group.

The set of norms to which tokens can be allocated in the second step is determined in the first step. Here, participants draw the norms that they believe exist in the population. Participants do this by drawing a path of appropriateness ratings through the entire action space – one such path corresponding to each norm (see Figure 3). Participants rate each action's appropriateness on a 3-item Likert scale (appropriate, neither appropriate nor inappropriate, inappropriate).[3]

[3]We ran a series of calibration pilots and they suggested that a higher number of options would reduce participants' ability to coordinate in the specific scenarios tested. Note that the number of options can be adapted depending on the scenario and research questions. Aycinena et al. (2022a) show that including a neutral option increases the predictive informativeness of norms elicited via the Krupka and Weber (2013) method, and so we follow that procedure here.

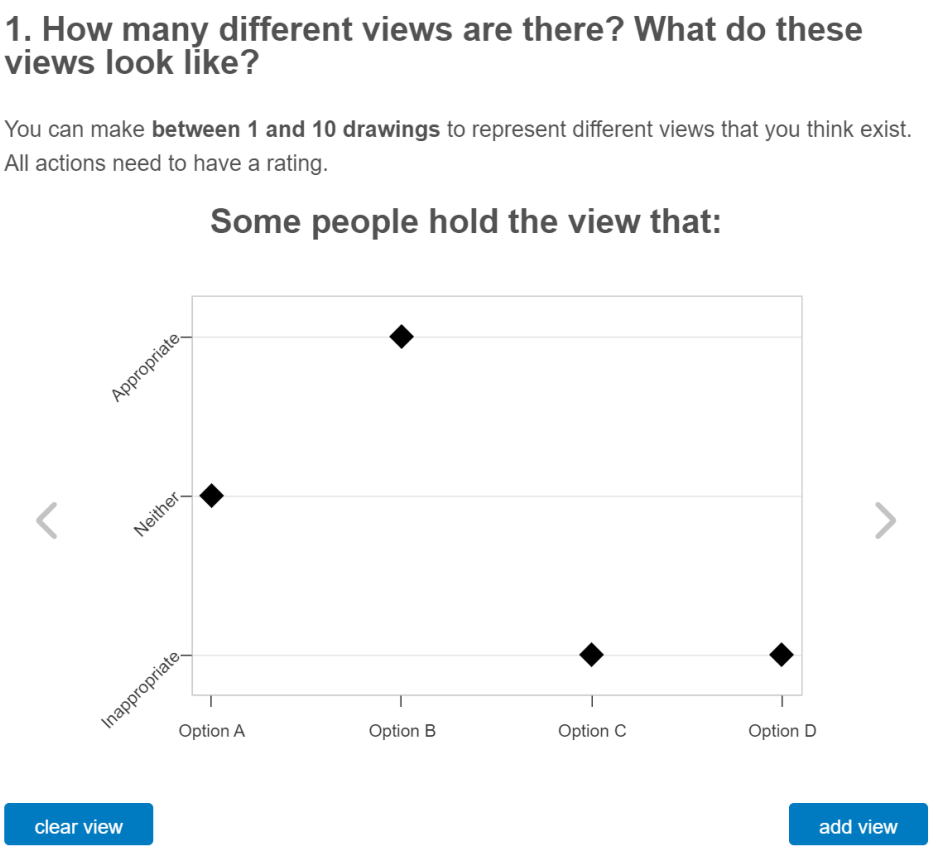

Figure 3: Graphical interface for drawing norms. Participants start with a blank drawing and can click or drag the dots to rate the appropriateness of each action. Participants can also navigate through the different drawings, delete drawings, or add up to ten different drawings. Actions are connected by lines if on a continuous scale, otherwise unconnected.

Participants can draw between one and ten distinct norms using the graphical interface. Crucially, the BESA incentive scheme encourages participants to draw exactly the norms that they think exist in the population, no more and no less. This is because both drawing a norm that no one else has drawn or failing to draw a norm that others have drawn can only reduce the likelihood of receiving the payment in step two. Thus, our Norm-Drawing Task is incentive-compatible and allows us to elicit the shape of each individual norm as well as the frequency of each norm in the population.

3.1. Theoretical predictions

Given these new data that were not available in the previous elicitation methods, we propose a brief theoretical argument on what can be expected in terms of behavior in our experiment when participants perceive a multiplicity of norms as compared to only one norm. Suppose that we elicit normative beliefs about some set of actions A (e.g., $A = \{\text{sniffle}, \text{sneeze}\}$).

Then in the first step, our task produces n norms represented by functions $\eta_i : A \to R$ for $i = 1, .., n$ where $R = \{\text{inappropriate, neutral, appropriate}\}$ is the scale of appropriateness chosen for the experiment. In the second step, participants choose numbers $p_i \in [0, 1]$ that represent the proportions of the population that hold norm η_i. We impose the condition $\sum_{i=1}^{n} p_i = 1$.

With these preliminaries, we can now think about the behavioral implications of the perceived multiplicity of norms. Specifically, we consider the punishment of norm violations that can be meted out by a participant who thinks there are many norms that apply to a given situation as compared to a participant who thinks there is only one norm. To do that, we follow Kimbrough and Vostroknutov (2023a) and suggest that the amount of punishment for violating some norm η_i is proportional to the resentment that the chosen action evokes. We model the resentment from choosing action $a \in A$ as the difference in social appropriateness between the action $a_i^* \in A$ that gives η_i-maximal appropriateness (what people often call "the norm" or the most appropriate action according to η_i) and the appropriateness of a. Thus, the resentment at a given η_i is given by $r(a|\eta_i) = \eta_i(a_i^*) - \eta_i(a)$. This number can be interpreted as the severity of norm violation when a was chosen, given that the normatively best action a_i^* was available.

Now, we compare the resentments coming from a participant who believes that there are many norms with a participant who believes there is only one. To model the participant with many norms, we follow Merguei et al. (2022) and consider an "expected" norm function

$$\mu(a) = \sum_{i=1}^{n} p_i \eta_i(a)$$

for all $a \in A$. The norm function μ can be thought of as the expected norm function that evokes resentment in the participant who perceives norms η_i with $i = 1, .., n$. The question now is: does μ evoke more or less resentment than each individual η_i? To answer this, suppose that $m^* \in A$ maximizes μ, or that m^* is the most appropriate action according to μ. Then

$$r(a|\mu) = \sum_{i=1}^{n} p_i(\eta_i(m^*) - \eta_i(a)) \leq \sum_{i=1}^{n} p_i r(a|\eta_i).$$

The last inequality follows from the fact that m^* is in general not the same as individual a_i^* that maximizes η_i. Moreover, this inequality implies that the expected resentment after a given μ (from the participant who perceives multiple norms) is no larger than the expected resentment after a in the population consisting of proportions p_i of agents who follow explicitly η_i. This already demonstrates that we should expect less punishment on average from participants who believe multiple norms exist than from participants who only believe in one norm.

An additional argument that participants with multiple norms punish less than partic-

ipants with one norm comes from the results of Merguei et al. (2022). The authors show experimentally that their participants, when presented information about the normative beliefs of two participants η_1 and η_2 in the Dictator Game after observing an offer a, choose to punish using the norm (η_1 or η_2) that prescribes cheaper punishment (generates less resentment $r(a|\eta_i)$). Thus, it is also possible that the participants who perceive multiple norms compute resentment as

$$r(a|\mu) = \min_{i=1..n} r(a|\eta_i).$$

In this case, punishment will be (weakly) smaller than any punishment prescribed by a single norm η_i.

We can conclude that theoretically, we should expect less punishment from the participants who report multiple norms than from those with only one. That is, pluralism breeds tolerance. We test this conjecture in two experiments below.

4. Experiment 1: Testing the relation between pluralism and tolerance

4.1. Methods

4.1.1. Experimental Treatments

We use the Norm-Drawing Task to elicit perceptions of multiplicity in two scenarios: a Dictator Game and an Allocation Game adapted from Engelmann and Strobel (2004). In the Dictator Game scenario, the dictator splits a pie of size 4 (experimental currency points; 1 point = \$0.50) with another unknown person. There are five possible allocations, from keeping all four points to giving all four points. In the Allocation Game scenario, a person has to choose between four possible allocations of money among themselves and two other people: a selfish option, an equitable option, a maximin option, and an efficient option (see Table 1).

Table 1: Distribution of experimental currency in the Allocation Game.

Allocation	selfish	equitable	maximin	efficient
Person A (chooser)	6	2	4	5
Person B/charity	0	2	3	5
Person C	0	2	3	1
sum of payoffs	6	6	10	11

We chose to test our task in Dictator Game and Allocation Game for the following reasons. First, both of them have been extensively studied in experimental economics, and we have a rather good understanding about what to expect in terms of behavior. We also know from previous experiments (e.g., Veselỳ, 2015; Erkut et al., 2015; Kimbrough and Vostroknutov,

2018; Aycinena et al., 2022b; König-Kersting, 2024; Kimbrough et al., 2024) how norms in the Dictator Game, as measured in the task by Krupka and Weber, should look. This presents us with a partial benchmark to compare our results with since, to our knowledge, no one previously measured normative beliefs in Allocation Game. Second, the two games differ in terms of the number of intuitively plausible norms that we found in the literature. While in the Dictator Game there is a broad consensus that equal split is the focal, most appropriate outcome, the Allocation Game is notorious for generating heterogeneous behavior where some people favor efficiency and some maximin norms (Engelmann and Strobel, 2004; Baader and Vostroknutov, 2017). Thus, we set out to test our task in these games to see if the intuitive presence of multiple norms in Allocation Game and a single norm in Dictator Game translates into similar observations in the Norm-Drawing Task.

As an additional test of the Norm-Drawing Task, we examine whether responses in the two scenarios depend on who the recipients are. For this reason, we include an alternative version of the two games with one recipient being replaced by a charity. The charity organization, Helen Keller International, was selected based on the rankings by GiveWell, a non-profit that helps donors figure out how to maximize their impact in terms of lives saved per dollar donated. In addition, Helen Keller International operates both in the U.S. and worldwide, so it can plausibly appeal to both conservative and liberal-leaning individuals (Pizziol et al., 2023). In the charity version of the Dictator Game, the dictator chooses how to distribute currency between themself and the charity. In the charity version of the Allocation Game, the decision-maker allocates points between themself, the charity, and a second experimental participant (Person C in Table 1).

4.1.2. Experimental Design

In the experiment, participants complete one of the four treatments: 2 scenarios (Dictator Game or Allocation Game) $\times$ 2 types of recipients (standard or charity). The full outline of the four questionnaires is available on the online repository for this experiment (osf.io/yh6gd).

In all treatments, participants first complete the Norm-Drawing Task for the specific scenario (Dictator or Allocation Game). The prize is set to 6 experimental currency points. After the Norm-Drawing Task, participants complete a third-party punishment game regarding the same scenario: participants are endowed with 4 points (in the Dictator Game scenario) or 6 points (in the Allocation Game scenario) and can choose to spend any amount to reduce the earnings of the dictator/chooser by an equal amount. Participants respond using a strategy method: they decide how much, if at all, to punish each possible action in the game described in the scenario. They are then paired, ex post, with another participant who made an actual decision in the game, and their punishment strategy is implemented based on that player's decision. Participants could decide to punish more than the player's earnings (e.g., 4 points even if the player only gets 2 points); in this case, the player received

0 points but was still informed about the size of the punishment.

In the third part of the experiment, participants acted as dictators/choosers to provide a match for other participants who were choosing punishment. However, in order to prevent the previous tasks from influencing their responses in the game, participants made choices in the other game, keeping the presence or absence of the charity fixed. For example, participants, who drew norms in the standard Dictator Game scenario, made choices in the standard Allocation Game, and vice versa. Participants were aware of the payment scheme, and thus knew that their actions could be punished.

4.1.3. Hypotheses

The experimental design, hypotheses, and analyses were pre-registered on the website of the American Economic Association (RCT ID AEARCTR-0012012). Any amendment to the original protocol is presented next to the related analysis with an explanation.[4] Multiple comparisons are corrected using the false discovery rate method (Benjamini and Hochberg, 1995). Square brackets indicate 95% confidence intervals. All tests are conducted in R (R Core Team, 2022). We formulate four hypotheses.

H1: non-randomness of responses. In the Norm-Drawing Task, participants will coordinate on one or multiple norms in a non-random manner: that is, the total number of norms guessed and the agreement between participants cannot be explained by participants simply drawing norms at random. Assuming the null hypothesis that participants draw norms at random, the relative frequency of each possible norm will be n/N, where n is the average number of norms guessed by participants, ranging between 1 and the maximum possible (i.e., 10), and $N = r^a$ (i.e., all possible norms that can be guessed), where r denotes the number of possible ratings (e.g., appropriate/neither/inappropriate, $r = 3$), and a denotes the number of possible actions (5 in the Dictator Game, 4 in the Allocation Game). To test for randomness, we compare observed frequencies to this discrete uniform distribution using a chi-squared test. We repeat the test for each experimental treatment.

H2: norm multiplicity. In the Norm-Drawing Task, some participants will draw multiple norms and some will draw only one norm. This will be the result of a mixture of H2A) participants acknowledging only one norm (e.g., the one they follow, as in the case of false consensus) and H2B) participants also considering other norms that differ in terms of what is most appropriate. To test for the existence of these different patterns of norm reporting we will look at descriptive statistics. For H2A we will count the number of participants who guess a single norm; for H2B we will count the number of participants who guess multiple

[4]The original Hypothesis 4, comparison with findings in Engelmann and Strobel (2004), is presented in Appendix B.1 as it is an auxiliary hypothesis that does not contribute to testing the main theoretical predictions of the method.

norms with mutually exclusive most appropriate actions.

H3: more punishment with unique norms. Participants who guess the existence of only one norm will punish non-appropriate actions (actions rated "inappropriate" or "neither appropriate nor inappropriate") more frequently (H3A) and with a higher magnitude than other participants (H3B). These predictions come from the theoretical argument at the end of Section 3. To test H3A, we use a mixed-effects logistic regression where the dependent variable is binary punishment (1 if action is punished, 0 otherwise), and a dummy variable indicating whether the participant guessed a single norm (1 if yes, 0 otherwise) as an independent variable. Participant ID is included as a random intercept. To test H3B, we use a mixed-effects linear regression where the dependent variable is punishment points allocated (from 0 to maximum). All other variables are the same as above.[5]

H4: charity as coordination device. The number of guessed norms will, on average, be smaller in the charity scenarios than in the standard scenarios. This should be the case because we anticipated that providing information about the effectiveness of Helen Keller International would induce strong agreement that it is a deserving recipient (strongly activating the view that one should give to charities). We report a Poisson regression with the number of guessed norms as the dependent variable and scenario and charity presence as predictors. We then test whether, for each scenario, the number of norms is higher in versions that do not include a charity as an agent (i.e., we expect more variability).

4.1.4. Sample Size

We sought to recruit 800 participants, 200 per experimental treatment. Although the sample size was computed based on budget constraints, we estimated the minimum detectable effect (MDE) size for our main hypothesis H1. Assuming that all participants guess the maximum possible number of norms (10 in the experiment), given all possible norms that can be guessed in the scenario with the most actions (N = 243), and a sample of 200 participants per test, then our MDE would be $w = 0.20$ with $\alpha = 5\%$ and a power $(1 - \beta)$ of 95%.

We ended up recruiting 820 U.S. residents on Prolific, a survey platform specialized in online experiments. The mean age was 42 ($SD = 14$), 49% were female, 49% were male, and 2% indicated other or preferred not to state their gender. Participants were approximately balanced across treatments (standard Dictator Game $N = 197$, charity Dictator Game $N = 226$, standard Allocation Game $N = 202$, charity Allocation Game $N = 195$) and so were age and gender.

[5]The test is repeated using several model configurations, including the formulation originally included in the preregistration, which all yield consistently significant results (Appendix B.2). Results are robust to only considering participants who drew a single norm with a single most appropriate action (i.e., who do not consider norms with multiple most appropriate ratings), with or without two dummy variables for the scenario and the presence of the charity, including or excluding punishment rating and its interaction with the dummy as independent variables.

4.2. Results

4.2.1. Descriptive Statistics

Figures 4 and 5 plot histograms of the number of norms drawn in the Norm-Drawing Task. Participants drew on average 4 to 5 norms ($SD = 3$). This suggests that the multiplicity of norms is a rather common phenomenon even when we consider very standard, typical allocation tasks like the Dictator and Allocation Games.

Figures C.9 to C.12 (see Appendix C) present the five most weighted norms and the weighted average of all norms in each treatment. Notably, while some norms were given considerable weight, the average number of tokens placed on any given norm was at most 41/100 (charity Allocation Game); so according to most participants, no single norm is held by the majority of people.

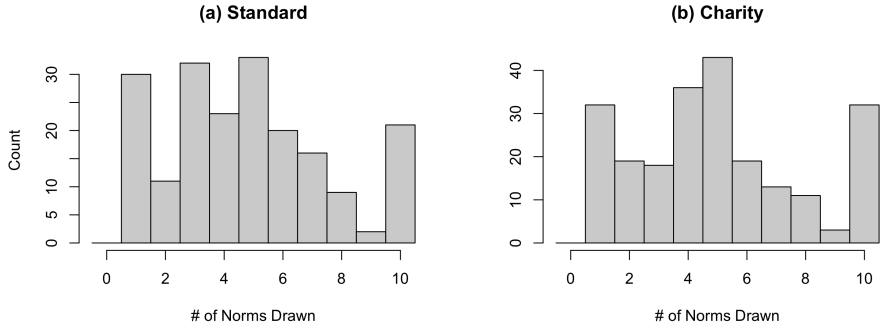

Figure 4: Histogram of Number of Norms Drawn by participant, Dictator Game.

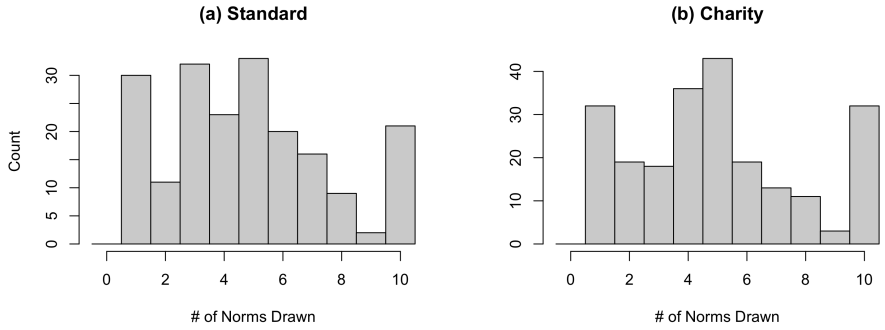

Figure 5: Histogram of Number of Norms Drawn by participant, Allocation Game.

These data suggest that existing norm elicitation procedures like the Krupka-Weber task would have concealed the variance in responses. It is interesting to mention, though, that the standard Krupka-Weber task seems to work in the sense that it produces average normative ratings in the Dictator Game that are very similar to our aggregates (compare, for example, Figure 2 in Kimbrough and Vostroknutov, 2018, to the bottom right graph on

Figure C.9). Thus, we can conclude that the "single-norm" Krupka-Weber task might be suitable for situations where only average normative beliefs are the object of interest.

A descriptive summary of the punishment and decision tasks is presented in tables 2 and 3.

Treatment	Most Frequently Punished Action	M	SD
Dictator Game (Standard)	Keep 3, Give 1	1.22	1.40
Dictator Game (Charity)	Keep 3, Give 1	1.06	1.32
Allocation Game (Standard)	Efficient option	1.50	1.81
Allocation Game (Charity)	Efficient option	1.43	1.80

Table 2: Third-Party Punishment Task. The modal response in the task was not to punish.

Scenario	Mean Given (SD)	Modal Response	Pr(Give > 2)
Dictator Game (Standard)	2.7 (0.91)	Keep 2, Give 2	1.5%
Dictator Game (Charity)	2.6 (1.10)	Keep 2, Give 2	10.7%
		Modal Response	**Pr(Max Efficiency)**
Allocation Game (Standard)		Equitable option	3.5%
Allocation Game (Charity)		Maximin option	16.4%

Table 3: Decision tasks.

4.2.2. H1: Non-Randomness of Responses

There are $3^5 = 243$ possible norm drawings for the Dictator Game scenario, and participants drew 130 unique norms in the standard version and 121 unique norms in the charity version, around one half of all possible combinations. In the Allocation Game scenario, $3^4 = 81$ drawings were possible, and participants drew 77 unique norms in the standard version and 74 unique norms in the charity version. In all four treatments, the distribution of beliefs was not uniform according to the pre-registered Chi-squared test (all $p < .001$). This result is robust even when excluding those norms that were not drawn by any participant: the frequency with which participants drew norms was still not uniform, as some norms were drawn overwhelmingly more often than others (all $p < .001$). This finding is also reflected in the number of tokens placed by participants on each norm: the ten norms with the highest average number of tokens account for between 54% (standard Dictator Game) and 65% (charity Allocation Game) of the entire distribution of norms in each treatment.

Finding 1: *Consistent with H1, participants' reported norms in the Norm-Drawing Task do not follow a random distribution.*

4.2.3. H2: Multiplicity of Norms

For Hypothesis 2, we look at the frequency of H2A participants who drew a single norm, and H2B participants who drew multiple norms with mutually exclusive appropriate actions.

In all treatments, more than half of the participants fall into one of these two categories (Table 4).

Table 4: Proportion of participants drawing a unique norm or multiple norms with mutually exclusive appropriate actions. The 'Others' category includes remaining participants, those who drew two or more norms having at least one appropriate action in common.

Scenario	Recipient	Unique	Multiple	Others
Dictator Game	Standard	15%	40%	45%
Dictator Game	Charity	14%	37%	49%
Allocation Game	Standard	17%	37%	46%
Allocation Game	Charity	23%	38%	39%

These data confirm that some individuals report only one norm, as the question is posed in existing norm elicitation tasks, but this group represents a minority of the sample. However, it is possible that some participants drew multiple versions of a single norm with different levels of tightness, i.e., all drawings could agree on the appropriate actions, but differ in the severity of their ratings of non-appropriate actions. Even allowing for this possibility, the proportion of participants with a unique norm increases by no more than 2%. Therefore, since the majority of respondents draw multiple norms, it is possible that those who report only one norm are subject to false consensus or pluralistic ignorance. As a second result, we observe that a consistent portion of the sample reports multiple, conflicting norms. This, in turn, may indicate that these participants take conflicting views into account when making their decision.

Finding 2: *Consistent with both H2A and H2B, the majority of participants reported only one norm or multiple contrasting norms, suggesting that there is a multiplicity of views within each scenario.*

4.2.4. H3: More Punishment with Unique Norm

Next, we test whether the probability and magnitude of punishment of inappropriate actions is higher for those participants who reported only one norm than in the rest of the sample. Both logistic and linear regressions predicting punishment suggest that these participants are more likely to punish ($\beta = 1.07$ $[0.48, 1.66]$, $z = 3.551$, $p < 0.001$) and punish more intensively ($\beta = 0.41$ $[0.18, 0.65]$, $t(963.4) = 3.419$, $p < 0.001$) than other participants (Figure 6). This result is robust to considering only the subset of participants who drew a single norm that featured a single most appropriate action (punishment frequency: $\beta = 0.96$ $[0.05, 1.86]$, $z = 2.072$, $p = 0.038$; punishment amounts: $\beta = 0.43$ $[0.05, 0.81]$, $t(822.4) = 2.195$, $p = 0.028$; Appendix D). If we look at the weight placed on each participant's most heavily weighted norm (as measured by the norm on which the highest number of tokens placed) instead of categorizing participants based on the number of norms drawn, the results

remain consistent: the more weight placed on a single norm, the higher the likelihood and magnitude of punishments (Appendix B.3).

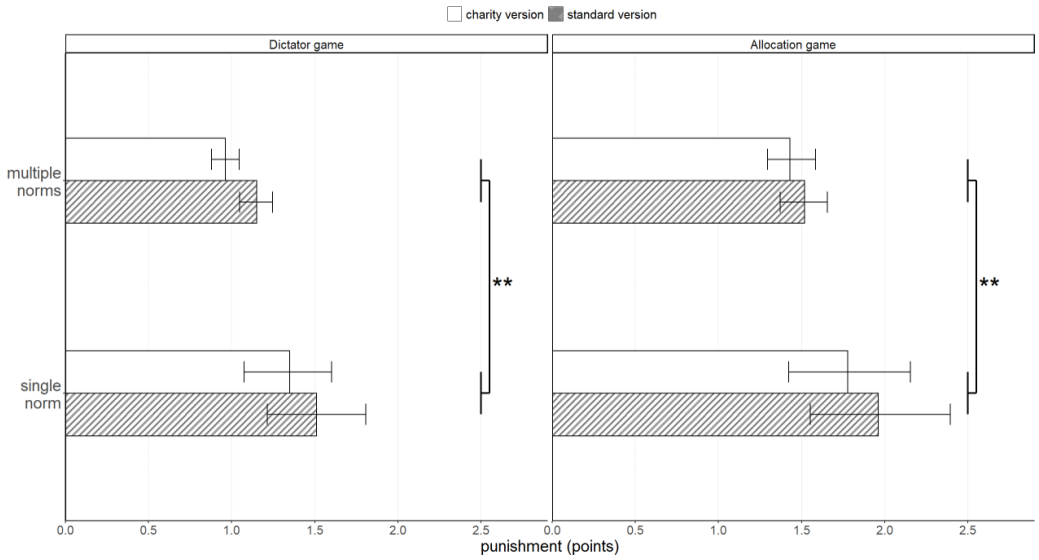

Figure 6: Average punishment, by scenario. Error bars show bootstrapped estimates for the 95% confidence interval around the mean, but are not clustered on the participant level, so should be considered illustrative only. The vertical bars illustrate the significance tests between multiple norm and single norm participants. **: $p < .05$

Finding 3: *Consistent with H3, participants who reported only one norm punished norm violations more frequently and more harshly.*

4.2.5. H4: Fewer Norms in the Presence of a Charity

This hypothesis predicted that the presence of the charity would have acted as a normative coordination device and thus would have brought an average reduction in the number of norms drawn. However, the regression coefficient on the Charity dummy in the pre-registered Poisson regression turned out to be non-significant ($\beta = 0.00$ $[-0.07, 0.06]$, $z = -0.062$, $p = 0.951$). Indeed, as anticipated in the descriptive statistics, the average number of norms drawn ranges from 4 to 5 in all treatments. This suggests that people do not have more uniform normative beliefs about giving to charity than about giving to other human participants.

Finding 4a: *Contrary to H4, we find no significant difference in the number of norms drawn between the standard and charity versions of each scenario.*

Despite this, Figures C.9 to C.12 in Appendix C show that the norms receiving the most tokens differ considerably across treatments. For instance, the standard Dictator Game treatment yields norms that favor equality (give 2 keep 2), and the charity Dictator Game treatment yields norms that favor generosity (Keep 0, Give 4). Thus, as an exploratory

analysis, we compare the types of norms between the standard and charity versions of each game. We employ a Dirichlet regression for compositional data. To simplify computations, we group together norms that share the same most appropriate actions (i.e., looser or tighter versions of the same norm). Several norms differ in both scenarios. In Dictator Game scenarios, the results show that the average number of tokens placed on "Keep 2, Give 2" norms is smaller in the charity version than in the standard version (6% vs. 31%, $\beta = -0.87$ $[-1.05, -0.68]$, $z = -9.214$, $p < .001$), but conversely the charity version yields a higher share of tokens placed on "Give 3 or 4" (15% vs. 5%, $\beta = 0.44$ $[0.25, 0.62]$, $z = 4.577$, $p < .001$) and "Keep 0, Give 4" norms (8% vs. 2%, $\beta = 0.33$ $[0.14, 0.52]$, $z = 3.463$, $p = .006$).

These results suggest that people use a different set of norms when dealing with charity as compared to other human beings, i.e. that norms are context-dependent. It seems that in the standard Dictator Game, participants mostly consider equality norms, but are unsure about how strong they are. Specifically, Norms 4 and 5 in Figure C.9 have the same most appropriate outcome but differ in how inappropriate the other outcomes are (stricter Norm 5 generates more punishment). In the charity Dictator Game in Figure C.10, we see only generosity norms (giving everything is the most appropriate outcome). Norms 1, 3, and 4 seem to share the same normative principle, but again are different in terms of tightness (Norm 1 is the tightest and generates the most punishment). Thus, participants agree that generosity is the norm in charity Dictator Game but are unsure about how tight the norm is.

In the Allocation Game scenario, the charity version yields a smaller share of equitable norms (15% vs 26%, $\beta = -0.34$ $[-0.54, -0.15]$, $z = -3.474$, $p = .008$) or "either equitable or maximin" norms (17% vs 26%, $\beta = -0.27$ $[-0.46, -0.08]$, $z = -2.754$, $p = .047$) compared to the standard version. This suggests again, as in the Dictator Game, that charity recipients evoke efficiency norms to a larger degree than recipients who are other experimental participants.[6]

Finding 4b: *Participants report different norms in the standard and charity versions of both games. Equality norms preponderate in the standard treatments. Efficiency norms dominate in the charity Allocation Game) and generosity norms dominate in the charity Dictator Game.*

[6] As a word of caution, we would like to bring the reader's attention to the fact that even though we do find the differences in the types of norms in standard and charity treatments, these types might not be the same in different settings. What we mean is that our participants are aware that their choices in Dictator Game and Allocation Game can be punished. This changes the behavior as compared to the scenario without punishment. For example, Engelmann and Strobel (2004) find that efficiency and maximin norms are most prevalent in the design without punishment, whereas we find the equality norm as the most frequent. The difference might come from the presence of punishment after the game. See also Appendix B.1. Since this is constant across treatments, it is not a problem for our analyses.

5. Experiment 2: Assessing the causal relationship between pluralism and tolerance

In the first experiment, we have shown how the Norm-Drawing Task is able to capture multiple co-existing norms, how much these norms are shared (or differ) among respondents, and that multiplicity is associated with reduced punishment, a proxy for tolerance. However, the evidence gathered so far cannot causally link multiplicity with tolerance; that is, we cannot claim that reporting multiple views is the reason why less punishment is observed, since both responses in the Norm-Drawing Task and in the Third-Party Punishment Task may be associated with some other, unrelated factor. In order to test whether pluralism actually begets tolerance, we designed a second experiment.

5.1. Experimental Design

We recruited a second sample of participants to complete the charity version of the Dictator Game. Participants were randomly assigned to three treatments. One treatment replicated the scenario from experiment 1, with the Norm-Drawing Task first followed by the third-party punishment game (Norm-Drawing treatment). The second treatment replaced the Norm-Drawing Task with the Krupka-Weber task (Krupka and Weber, 2013), which elicits the single most normative view rather than multiple views (Krupka-Weber treatment). Lastly, we included a control treatment in which participants completed the third-party punishment game first, followed by the Norm-Drawing Task, so that punishment could be directly measured before any explicit mention of norms.

5.2. Hypotheses

The experimental design, hypotheses, and analyses were pre-registered on the Open Science Framework (osf.io/74u9c). We made one amendment to the original protocol due to a wording error.[7]

In addition to replicating the original findings relating perceptions of multiplicity to tolerance, this second experiment is designed to test the hypothesis that different norm elicitation methods would cause different punishment behaviors. Specifically, we hypothesize that there will be less punishment when third-party punishment by those who punish after completing the Norm-Drawing Task, which focuses on the plurality of views, than by those who punish after completing the Krupka-Weber task, which elicits a single norm perceived to be the most prevalent in the group. That is, the design allows us to test whether focusing on multiplicity increases tolerance.

[7]The unamended results, which are nevertheless consistent with those presented in the main text, are presented in Supplementary Analysis Appendix E.1.

5.3. Materials and methods

The experiment closely followed the procedure for the first experiment, with some minor changes: the graphical user interface was improved based on feedback from the original experiment, such as a different placement of buttons in the Norm-Drawing Task and a different arrangement of questions in the third-party punishment game. In addition, we reiterated in the punishment instructions that the money will indeed be donated to Helen Keller International. The implementation of the Krupka-Weber task followed the formulation of the questions in the original paper, but the response method was analogous to the Norm-Drawing Task, using a graphical interface to draw a single norm.

For the analyses, we adopt the standard 5% significance level to test against the null hypotheses. In case of a non-significant result in comparing treatments, we conduct equivalence testing using the TOST procedure (Lakens, 2017) with a difference in punishment of ± 0.5 points. Post-hoc tests and multiple analyses are corrected for multiple comparisons using a Benjamini-Hochberg procedure. Confidence intervals represent the 95% confidence level.

5.4. Sample Size

Based on budget considerations and on aims to replicate the findings in the first experiment, we aimed to collect a sample of 600 participants. We conducted two power analyses to measure the ability of the experiment to replicate the original findings (increased frequency and increased amount of third-party punishment for participants reporting only one norm in the Norm-Drawing Task) using the 'mixedpower' R package (Kumle et al., 2021). Given that these two analyses apply to only two of the three experimental treatments (excluding the Krupka-Weber treatment), we computed power based on $N = 600 \times 2/3 = 400$. Based on the original effect sizes and given an alpha of 5%, the results revealed that with this sample, the power for the punishment rate finding is $\approx 100\%$ whereas the power for the punishment magnitude finding is $\approx 63\%$.

5.5. Results

5.5.1. Sample characteristics

We recruited a sample of 600 participants via Prolific (Krupka-Weber: $N = 216$; Norm-Drawing: $N = 196$; control: $N = 188$) representative of the U.S. population in terms of gender, age, and ethnicity. 49.5% of participants was female (48.5% male, 2% other), the average age was 45.7 ($SD = 15.8$), and 73.3% was college-educated.

5.5.2. Norm elicitation responses

In the Krupka-Weber task, the median rating for each action was that it is socially inappropriate to keep everything, neither appropriate nor inappropriate to keep 3 points and give 1 point, and socially appropriate to give any higher amount. The three most common views in

the Norm-Drawing Task are that all actions are appropriate (10.5%), that all actions are neither appropriate nor inappropriate (8%), and the median norm reported in Krupka-Weber (6.8%). We also tested whether participants in the Norm-Drawing and control treatments reported different norms by means of a Dirichlet regression. All comparisons failed to reach the 5% significance level.

5.5.3. Replication of punishment findings

15.6% of participants drew only one norm in the Norm-Drawing Task, in line with the 17% who did so in the first experiment. As above, we examine the difference in punishment frequency between this subgroup and the rest of the sample using a logistic mixed-effects model (Table 5). In this model, punishment (0 = no punishment; 1 = punishment) was the predicted variable, with a dummy indicating whether the participant reported only one view (1) or not (0) as a predictor, and a random intercept for each participant. To test for differences in punishment amounts, we use a linear mixed-effects regression with the same parameters, but using the punishment amount instead of a dummy as the predicted variable. The results confirm the original findings: likelihood of punishment is 22pp higher for participants reporting only one view (one view: 67% [64.5%,69.6%], other participants: 45% [43.1%,47.1%], $z = 6.370$, $p < .001$), and overall punishment is also larger for this group (+0.73 points on a 0-4 scale, $t(410.8) = 4.548$, $p < .001$).

	Punishment Frequency		**Punishment Amount**	
	(1)	*(2)*	*(1)*	*(2)*
α	−0.543*	−0.034	1.004***	1.056***
	(0.269)	(0.375)	(0.059)	(0.084)
one view	3.971***	3.554***	0.725***	0.677**
	(0.788)	(1.003)	(0.159)	(0.221)
Norm-Drawing		−0.980		−0.102
		(0.531)		(0.118)
one view × Norm-Drawing		0.798		0.094
		(1.407)		(0.320)

Table 5: Punishment frequency and amount, with and without task presentation order. The first two models are logistic regressions (coefficients are log odds), and the second two are linear regressions. Models marked as *(1)* are the pre-registered models, whereas models marked as *(2)* are the models with treatment information. Standard errors in parentheses. The intercept α represents punishment for participants in the control group who reported multiple views. *: $p < .050$; **: $p < .010$; ***: $p < .001$.

We additionally test whether there are differences between the control and Norm-Drawing treatments by introducing a dummy variable for treatment and its interaction with the one-view dummy. The result is significant regardless of the treatment, and with a similar magnitude in both tests (all p-values for the interaction term $> .571$). This suggests that even if the Norm-Drawing Task is presented after the punishment decision, there is still a strong relation between the number of views reported and punishment. Furthermore, supplementary exploratory analyses show that results are robust to the use of other measures (Supplementary analyses Appendix E.1, Appendix E.2, and Appendix E.3), and that increased

punishment is mostly driven by an increase in the probability of punishment rather than an increase in the individual amount punished (Supplementary analysis Appendix E.4). In short, we cleanly replicate the results of experiment 1.

Finding 5: *In line with experiment 1, subjects who report only one norm in the Norm-Drawing task punish both more frequently and more intensely, regardless of whether the Norm-Drawing Task task is presented before or after the punishment decision.*

5.5.4. Influence of the norm elicitation method on punishment

We compared the punishment frequency and magnitude across the three experimental conditions using logistic and linear mixed effects regressions with experimental treatment as predictor variable and participant ID as a random intercept (Figure 7). The results reveal that punishment frequency is highest in the Krupka-Weber treatment (63.7% [60.7%,66.7%]), compared to both the Norm-Drawing treatment (45.2% [42.6%,47.8%], +18.5pp, $z = 4.012$, $p < .001$) and control treatment (54.3% [51.7%,57%], +9.4pp, $z = 2.097$, $p = .043$). Punishment frequency was also significantly lower in the Norm-Drawing treatment treatment compared to the control treatment (-9.2pp, $z = -2.027$, $p = .043$). When looking at punishment amounts, Krupka-Weber participants punished significantly more than Norm-Drawing participants (+0.297 [+0.028,+0.565] on a 0 to 4 scale, $t(590) = 2.653$, $p = .025$), but we found no significant difference when comparing treatments to the control group (all $p > 0.129$).

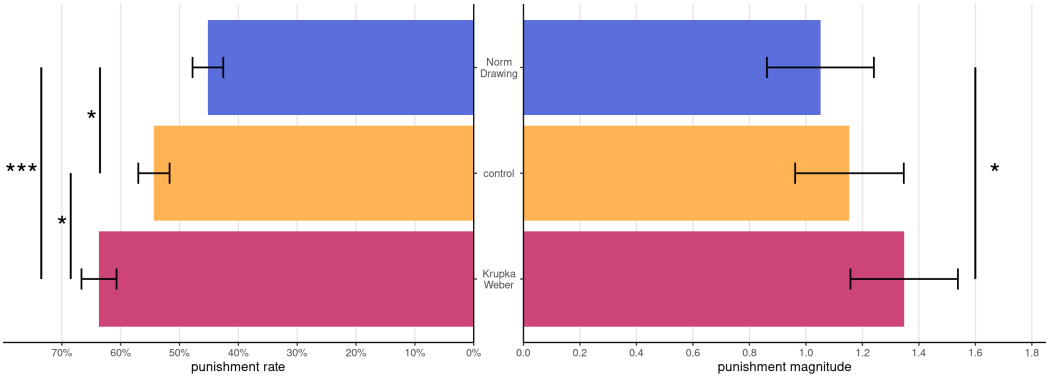

Figure 7: *: $p < .050$; **: $p < .010$; ***: $p < .001$.

As preregistered, we conducted equivalence tests to determine if the magnitude of punishment was practically equivalent across treatments. The tests revealed that the average punishment in both treatments was within half a point of the control group's average punishment. A more stringent, non-preregistered test using a 1/3-point interval indicated that the equivalence held for the Norm-Drawing treatment (both boundaries $p < .020$), but not for the Krupka-Weber treatment (upper boundary $p = .108$). These results suggest that punishment amounts in the control group are practically equivalent to those in the other treatments, with the equivalence being more robust for the Norm-Drawing treatment.

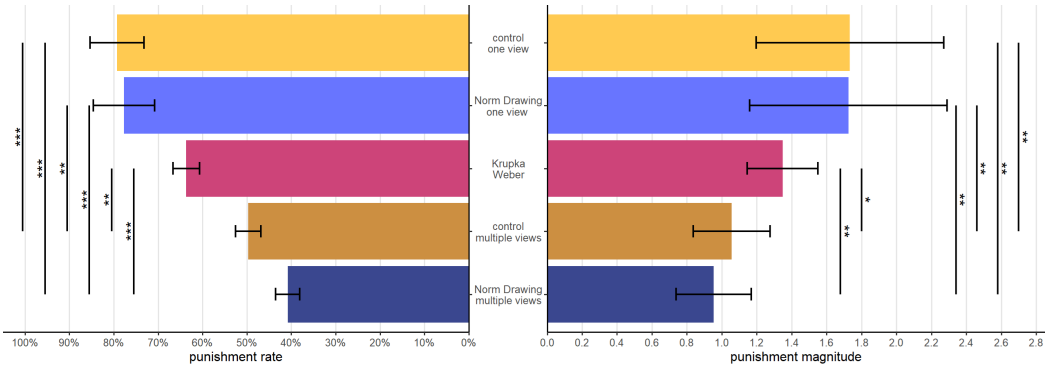

Figure 8: *: $p < .050$; **: $p < .010$; ***: $p < .001$.

Given that a considerable portion of participants report only one view in the Norm-Drawing Task, we can test whether these participants behave similarly to participants in the Krupka-Weber treatment. To test this, we repeat the above analyses, this time subdividing participants in the control and Norm-Drawing treatment based on whether they reported one or multiple views. Pairwise comparisons reveal that all groups reporting multiple views punish less harshly and less frequently than all groups reporting only one view, including participants in the Krupka-Weber treatment (Figure 8). This finding suggests that regardless of whether by choice (Norm-Drawing task) or by design (Krupka-Weber task), participants who report one norm tend to be more punitive even if – arguably – some participants who are forced to report the most common norm in the Krupka-Weber treatment would recognize the co-existence of multiple views if they were asked to do so in the Norm-Drawing Task.

Finding 6: *Allowing subjects to report multiple norms (via the Norm-Drawing Task) causes an increase in tolerance. The relation between multiplicity and tolerance is observed regardless of whether reporting more or fewer norms is revealed by choice or enforced by design.*

6. Discussion

In this study, we present the Norm-Drawing Task, a new norm elicitation task that aims to directly measure norm multiplicity: how many norms co-exist in a given choice context; how different the beliefs among various people are with regard to this multiplicity; and how multiplicity influences (punishment) behavior. We tested the task in two experimental games using two different sets of participants, varying the choice context by either including a charitable organization as a recipient or not.

Results reveal a significant multiplicity of norms in all scenarios. We also find that a stable proportion of participants, around 15%, believe that only one norm applies in each scenario, which suggests that some people have overly simplified views of the normative landscape. Finally—and consistently with our theoretical argument—we find that these "single-norm" participants punish more than their "multiple-norm" counterparts (see

discussion below). This suggests that our method is not only able to capture the three above-mentioned desiderata of understanding norm multiplicity, but also sheds some new light on the necessity to take this multiplicity into account given that we find on average 4 to 5 different norms perceived by our average participant in all scenarios.

Our results suggest that participants do not respond randomly in the task and that most of them indeed perceive a multiplicity of different normative views. The fact that we found multiplicity in very standard games studied for a long time in experimental economics suggests that the possibility for norm multiplicity should be taken into account in all situations, be they experimental or applied. Thus, our Norm-Drawing task opens new avenues for research into this important issue. For example, using our method it is possible to observe instances of false consensus or pluralistic ignorance happening among participants who draw only one norm when instead there is a multiplicity of views in the population.[8] This method also opens up the possibility of discovering intermediate levels of knowledge where participants report multiple norms, but still ignore some or most of them, and their beliefs can be linked to their behavior. In addition to identifying biases in recognizing diverse perspectives on what constitutes appropriate behavior in a given context, this method also makes it possible to determine whether a norm is perceived as more or less strict by different respondents, adding to the usefulness of this method over existing approaches.

In addition to enhancing the toolbox, our task suggests new possibilities for connecting norm-driven behavior to important societal problems like, for example, violence. Our results suggest that people who perceive only one norm in an environment where there are – in fact – many, might be susceptible to normative disagreement with others who perceive multiple norms. This is simply because they will not tolerate any form of behavior different from what their solely perceived norm dictates. Thus, the proportion of people who perceive only one norm can serve as an indicator of the potential for normative disagreement or violence.

Our first experiment also reveals how participants who report only one norm tend to punish more and more frequently than other participants, suggesting that the presence of multiple norms may provide leeway for people in the scenario to act according to their preferences. Our second experiment confirms this finding and further offers causal evidence that seeing the world through the lens of multiple norms reduced punishment and increases tolerance: being forced to report a single norm causes an increase in punishment frequency and magnitude compared to a control group directly completing the punishment task and compared to participants who are allowed to draw multiple norms. This suggests that making one norm salient influences punishment preferences. These findings align with findings in experimental and neuro-economics on moral wiggle room (Dana et al., 2007) or moral

[8]It is possible to disentangle the two phenomena, but this would require knowledge of personal beliefs since false consensus implies that the norm is one's own, while pluralistic ignorance does not necessarily entail endorsement of such a norm.

opportunism (van Baar et al., 2019; Merguei et al., 2022). While it is perhaps not surprising that the perceived presence of multiple norms acts as a deterrent to punishing norm violators, it has not been trivial to capture this phenomenon without experimental manipulation (Dimant and Gesche, 2023). Here, we offer a way to study this phenomenon by simply observing the differences in beliefs between respondents.

Our finding that punishment decreases in multi-norm environments (as compared to single-norm environments perceived by some of our participants) also suggests that the multiplicity of norms can promote more tolerance of different views within a group. At the same time, it is also possible that a reduction in punishment may cause norms to deteriorate over time. Less punishment resulting from perception of multiplicity can make norm violations more socially tolerable and thus lead to an erosion of norms. However, it is also possible for multiplicity to increase punishment: although we do not observe it directly in the experiment, multiplicity could lead to the emergence of conflicting norms, such that they are endorsed by distinct subgroups of the same population. This, in turn, could lead to an increase in punishments for violations of either norm. These prospects suggest another reason why it might be important to check for the multiplicity of norms in applied research.

Our analyses concerning the effect of introducing a charity as a recipient also provide insight into how perceived norms may change with contextual features of the scenario, such as the agents involved. Although we expected the charity to act as a coordinating device, reducing the number of norms drawn and placing emphasis on norms of generosity towards the charity, our tests for the difference turned out to be non-significant. It is possible that adding a charity might have added an additional layer of complexity to the scenario, rather than pruning the space of norms. In support of this possibility, our tests suggest that the content of norms changes considerably between the two versions of each scenario even though their cardinality does not.

Based on our experimental findings, we propose several applications of the Norm-Drawing task. First, because participants can report the existence of multiple norms, the method allows researchers and practitioners to study the entirety of the normative landscape in which people believe they are acting. This property is useful in environments that are perceived as polarized: individuals may believe that there are multiple conflicting views and act in consideration of those views. Studying perceived polarization can be helpful in understanding how it affects behavior, for instance in online environments where the perception of extreme views is amplified by the architecture of social media platforms (Bail, 2022). Conversely, given the ability of the method to identify cases of false consensus and pluralistic ignorance, this method may be useful in conjunction with interventions targeting these biases: the Norm-Drawing task could first be used to determine whether a plurality of norms exists and, if so, identify those respondents who report only one or a fraction of the norms actually reported by most participants.

In addition, the Norm-Drawing task could be used to measure norm "tightness" (Gelfand et al., 2011) in contexts where a single norm is established, such as traffic laws. For example, researchers might include in the task only actions that deviate from the expected rules of behavior (e.g., jaywalking) and ask participants to predict the population's views about these behaviors. In addition to modifying the set of actions studied, researchers may change the number of ratings possible: if for instance researchers are not interested in this tightness-looseness dimension, they could simply ask whether actions are appropriate or inappropriate, and disregard how norm deviations are perceived. This a benefit of our flexible graphical interface, which could simply ask participants to mark only the actions they consider appropriate for each view. Such an approach would be particularly useful for reducing complexity in scenarios where there are multiple actions and therefore a combinatorial explosion of possible norms.

References

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Artinger, F., Exadaktylos, F., Koppel, H., and Sääksvuori, L. (2010). Applying quadratic scoring rule transparently in multiple choice settings: A note. Working paper.

Aycinena, D., Bogliacino, F., and Kimbrough, E. O. (2022a). Measuring norms: Assessing the Krupka-Weber elicitation method. Mimeo.

Aycinena, D., Bogliacino, F., and Kimbrough, E. O. (2022b). Measuring norms using the BESA method. Mimeo.

Baader, M. and Vostroknutov, A. (2017). Interaction of reasoning ability and distributional preferences in a social dilemma. *Journal of Economic Behavior & Organization*, 142:79–91.

Bail, C. (2022). *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.

Barr, A., Lane, T., and Nosenzo, D. (2018). On the social inappropriateness of discrimination. *Journal of Public Economics*, 164:153–164.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C., Dimant, E., Gächter, S., and Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, 132:59–72.

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.

Borel, E. (1921). La théorie du jeu et les équations intégralesa noyau symétrique. *Comptes rendus de l'Académie des Sciences*, 173(1304-1308):58.

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

Centola, D., Becker, J., Brackbill, D., and Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119.

Chang, D., Chen, R., and Krupka, E. (2019). Rhetoric matters: A social norms explanation for the anomaly of framing. *Games and Economic Behavior*, 116:158–178.

Charness, G., Gneezy, U., and Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization*, 189:234–256.

Cialdini, R. B. and Trost, M. R. (1998). Social influence: Social norms, conformity and compliance.

Crosetto, P. and De Haan, T. (2023). Comparing input interfaces to elicit belief distributions. *Judgment and Decision Making*, 18:e27.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

DellaVigna, S., List, J. A., and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The quarterly journal of economics*, 127(1):1–56.

Dimant, E. (2019). Contagion of pro-and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*, 73:66–88.

Dimant, E. (2023a). Beyond average: A method for measuring the tightness, looseness, and polarization of social norms. *Economics Letters*, 233:111417.

Dimant, E. (2023b). Hate trumps love: The impact of political polarization on social preferences. *Management Science*.

Dimant, E., Clemente, E. G., Pieper, D., Dreber, A., Gelfand, M., and 9, B. S. U. C. H. M. . H. A. . T. P. (2022a). Politicizing mask-wearing: predicting the success of behavioral interventions among republicans and democrats in the us. *Scientific Reports*, 12(1):7575.

Dimant, E., Galeotti, F., and Villeval, M. C. (2023). Motivated information acquisition and social norm formation. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.4525398.

Dimant, E., Gelfand, M., Hochleitner, A., and Sonderegger, S. (2022b). Strategic behavior with tight, loose, and polarized norms. Working Paper Available at SSRN: https://bit.ly/3ryY3Pc.

Dimant, E. and Gesche, T. (2023). Nudging enforcers: How norm perceptions and motives for lying shape sanctions. *PNAS Nexus*, 2(7):pgad224.

Eckel, C. C., Hoover, H. G., Krupka, E. L., Sinha, N., and Wilson, R. K. (2023). Using social norms to explain giving behavior. *Experimental Economics*, 26(5):1115–1141.

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review*, 94(4):857–869.

Erkut, H., Nosenzo, D., and Sefton, M. (2015). Identifying social norms using coordination games: Spectators vs. stakeholders. *Economics Letters*, 130:28–31.

Fragiadakis, D. E., Kovaliukaite, A., and Rojo Arjona, D. (2019). the belief elicitation by superimposition approach. Working paper.

Fromell, H., Nosenzo, D., Owens, T., and Tufano, F. (2021). One size does not fit all: Plurality of social norms and saving behavior in Kenya. *Journal of Economic Behavior & Organization*, 192:73–91.

Gelfand, M., Li, R., Stamkou, E., Pieper, D., Denison, E., Fernandez, J., Choi, V. K., Chatman, J., Jackson, J. C., and Dimant, E. (2021a). Persuading republicans and democrats to comply with mask wearing: An intervention tournament.

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.

Gelfand, M. J., Jackson, J. C., Pan, X., Nau, D., Pieper, D., Denison, E., Dagher, M., Van Lange, P. A., Chiu, C.-Y., and Wang, M. (2021b). The relationship between cultural tightness–looseness and covid-19 cases and deaths: a global analysis. *The Lancet Planetary Health*, 5(3):e135–e144.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., et al. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, 332(6033):1100–1104.

Groenendyk, E., Kimbrough, E. O., and Pickup, M. (2023). How norms shape the nature of belief systems in mass publics. *American Journal of Political Science*, 67(3):623–638.

Henrich, J. (2017). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

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.

Kimbrough, E. and Vostroknutov, A. (2018). A portable method of eliciting respect for social norms. *Economics Letters*, 168:147–150.

Kimbrough, E. and Vostroknutov, A. (2023a). Resentment and punishment. mimeo, Chapman University and Maastricht University.

Kimbrough, E. and Vostroknutov, A. (2023b). A theory of injunctive norms. mimeo, Chapman University and Maastricht University.

Kimbrough, E. O., Krupka, E. L., Kumar, R., Murray, J., Ramalingam, A., Sánchez-Franco, S., Sarmiento, O. L., Kee, F., and Hunter, R. (2022). On the stability of norms and norm-following propensity: A cross-cultural panel study with adolescents. SSRN Working paper 4025407.

Kimbrough, E. O., Krupka, E. L., Kumar, R., Murray, J. M., Ramalingam, A., Sánchez-Franco, S., Sarmiento, O. L., Kee, F., and Hunter, R. F. (2024). On the stability of norms and norm-following propensity: A cross-cultural panel study with adolescents. *Experimental Economics*, pages 1–28.

Kimbrough, E. O. and Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3):608–638.

König-Kersting, C. (2024). On the robustness of social norm elicitation. *Journal of the Economic Science Association*, pages 1–13.

Krupka, E. L., Weber, R., Crosno, R. T., and Hoover, H. (2022). "when in rome": Identifying social norms usingcoordination games. *Judgment and Decision Making*, 17(2):263–283.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

Kumle, L., Võ, M. L.-H., and Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in r. *Behavior research methods*, 53(6):2528–2543.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362.

Lang, J., Erickson, W. W., and Jing-Schmidt, Z. (2021). # maskon!# maskoff! digital polarization of mask-wearing in the united states during covid-19. *PloS one*, 16(4):e0250817.

Levy, R. (2021). Social media, news consumption, and polarization: evidence from a field experiment. *American Economic Review*, 111(3):831–70.

Merguei, N., Strobel, M., and Vostroknutov, A. (2022). Moral opportunism as a consequence of decision making under uncertainty. *Journal of Economic Behavior & Organization*, 197:624–642.

Offerman, T., Sonnemans, J., Van de Kuilen, G., and Wakker, P. P. (2009). A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *The Review of Economic Studies*, 76(4):1461–1489.

Panizza, F., Dimant, E., Kimbrough, E. O., and Vostroknutov, A. (2024). Measuring norm pluralism and perceived polarization in us politics. *PNAS Nexus*.

Peeters, R. and Wolk, L. (2019). Elicitation of expectations using Colonel Blotto. *Experimental Economics*, 22(1):268–288.

Pickup, M., Kimbrough, E. O., and de Rooij, E. A. (2021). Expressive politics as (costly) norm following. *Political behavior*, pages 1–21.

Pickup, M., Kimbrough, E. O., and de Rooij, E. A. (2023). Crossing the line: Evidence for the categorization theory of spatial voting. *British Journal of Political Science*, page 1–11.

Pizziol, V., Demaj, X., Di Paolo, R., and Capraro, V. (2023). Political ideology and generosity around the globe. *Proceedings of the National Academy of Sciences*, 120(15):e2219676120.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roberson, B. (2006). The colonel blotto game. *Economic Theory*, 29(1):1–24.

Sherif, M. (1936). The psychology of social norms.

Smerdon, D., Offerman, T., and Gneezy, U. (2020). 'everybody's doing it': on the persistence of bad social norms. *Experimental Economics*, 23:392–420.

van Baar, J. M., Chang, L. J., and Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature communications*, 10(1):1483.

Veselỳ, Š. (2015). Elicitation of normative and fairness judgments: Do incentives matter? *Judgment and Decision making*, 10(2):191–197.

Appendix A. Supplementary Materials

Appendix A.1. Eliciting distributions and uncertainty

In this section, we present several incentive-compatible methods for eliciting distributions of beliefs, that could be adapted to measure norms. One such approach is the quadratic scoring rule (QSR) that is frequently used to incentivize the elicitation of various kinds of beliefs (Artinger et al., 2010). With regard to norms, Dimant et al. (2022b); Dimant (2023a) uses QSR to estimate beliefs about descriptive norms in a one-shot Public Good Game. Participants are paired with each other before the game and have to guess how the opponent will play by assigning probabilities to each possible choice of contribution. Then, the actual contribution of the opponent is revealed, and the participants are paid based on the probability they have assigned to the true choice according to QSR.

As with any scoring rule method, QSR works best when predictions are compared to an objective benchmark, which makes it a suitable elicitation method for descriptive norms. However, when the benchmark is not objectively verifiable—which is the case with the injunctive norms—QSR does not provide an incentive-compatible mechanism.[9] Such a situation may not be desirable. In such cases, other methods are used that overcome this problem at the expense of simplicity.

Another approach employs a variant of the Colonel Blotto game to elicit beliefs (Peeters and Wolk, 2019). The Blotto game was first formalized by Borel (1921) and developed in Roberson (2006). Here, each option is conceived as a "battlefield" where players place their "forces" (i.e., tokens). Players try to predict which battlefield will have the largest number of forces deployed since a fixed prize will be assigned to participants according to the proportion of tokens on that battlefield (the more tokens the higher the probability of the prize). The best strategy in this game is to weigh each battlefield according to its probability of being the one with the most forces and to place one's tokens on that battlefield.

Colonel Blotto could easily be adapted to elicit distributions of normative beliefs by defining each appropriateness rating as a battlefield. While this approach allows for eliciting a unique norm, its winner-take-all incentive structure risks concealing the presence of multiple norms. As an example, think of an action that is considered very appropriate by a majority of the population and very inappropriate by a minority. Since in the Colonel Blotto game, only one battlefield will be chosen for payment, the best strategy for participants is to focus on how most players will place their tokens. This makes participants think about the norm endorsed by the majority, and disregard the tokens placed by the minority. The elicitation procedure thus risks not reflecting participants' actual distribution of normative beliefs by not giving them incentives to reveal their beliefs about potential minorities.

Another method for eliciting uncertain beliefs is the BESA method presented in the main text (Fragiadakis et al., 2019). This is a scoring-rule method that elicits belief distributions while preserving incentive compatibility.

Aycinena et al. (2022b) show how to adapt this method to measure uncertain normative beliefs by extending the two-step method developed by Bicchieri and Xiao (2009); in their design the histogram produced by participants is compared to the empirical distribution of personal normative beliefs reported in step 1.

Here we note that a further extension of this approach can make it analogous to the coordination game method introduced by Krupka and Weber (2013). In this variant of BESA, participants are instructed to assign tokens across the appropriateness ratings for each action according to how they believe others will place their own tokens. Matching others' guesses about the distribution increases the chances of payment: once all players have placed their tokens, an average distribution is computed across all participants. The closer the participant's distribution is to the average, the higher the likelihood of winning a fixed sum of money.

This variant of BESA can be used to estimate the distribution of beliefs over single actions. Consider blowing one's nose in public: respondents may be asked to estimate the share of individuals who think that nose-blowing is appropriate, and the share of those who think is inappropriate, then to estimate the share who think it is appropriate to sniff, and so on. However, as we note in the main text, this cannot solve the problem of distinguishing multiple norms from one another.

[9] What's more, QSR is only incentive compatible under risk neutrality (Offerman et al., 2009).

Appendix B. Supplementary Analyses

Appendix B.1. Original Hypothesis 4

We report below the original hypothesis 4 as formulated in the pre-registered report, and the related analyses.

Hypothesis: In the standard Allocation Game, the number of participants guessing a norm with equality rated as the only appropriate action will be smaller than A) the number of participants guessing a norm with the maximin option rated as the only appropriate action, B) the number of participants guessing a norm with efficiency rated as the only appropriate action. To test H4A and H4B we use pairwise post-hoc tests following a Chi-squared test comparing the number of participants drawing each norm type. This test is analogous to the pre-registered test but replaces it because the original model was not identifiable.

Results: In their seminal paper, Dirk Engelmann and Martin Strobel found that equity concerns did not predict choices in most participants. Contrary to this prediction, participants in the standard Allocation Game treatment demonstrated a strong preference for the equitable option: this action was preferred by 45% of participants. Moreover, the number of participants drawing an equity norm (i.e. a norm with only the equitable option rated as appropriate, $N = 106$) was more frequently drawn than an efficiency norm ($N = 31$) or a maximin norm ($N = 49$, post-hoc Chi-squared test comparisons, all $p < .001$).

Contrary to H4A and H4B, participants drawing an equity norm outnumber those drawing an efficient or maximin norm.

Thus, contrary to expectations, we did not find the same results from the literature suggesting that equity concerns are not the primary normative driver of participants' decisions. Instead, the equity norm was the most frequently drawn compared to the other ones surveyed (maximin, efficiency, selfishness), and by a wide margin. We suggest that the focus on equity may be driven by the particular sample on which we tested our method: prolific participants. Indeed, it is possible that participants in this particular participant pool share a sense of reciprocity to the point that compensation should not be asymmetric, even when participants do not know each other directly. Another factor that may have influenced the observed behavior is the structure of the game compared to the original version: the game in the experiment presents four options that separate the different social preferences taken into account, whereas the original study used a series of games with three options. It is possible that the different payoff matrix and the use of multiple games might have influenced the preferences of players. Running the experiment on different samples may reveal whether norm prevalence depends on the reference group, or on differences between our game and those used in previous studies.

Appendix B.2. Pre-registered test for Hypothesis 3

The original test for H3A and H3B included as independent variables the participant's minimum appropriateness rating for the given action across all guessed norms (inappropriate, neither, appropriate), and the interaction between this variable and the presence of a single norm with a unique appropriate action. This test split results between "inappropriate" and "neither appropriate nor inappropriate" ratings, whereas the hypothesis does not distinguish between the two cases. We report for completeness the results of the tests here, which are in line with the tests reported in the main text. Results confirm the finding that reporting a unique norm with a unique appropriate action in the Norm-Drawing Task yields higher likelihood to punish (H3A: $\beta = 1.15$ $[0.23, 2.07]$, $z = 2.449$, $p = 0.014$) and to greater punishments (H3B: $\beta = 0.51$ $[0.12, 0.89]$, $z = 2.551$, $p = 0.011$). As an exploratory test, we also examine the effect of our main independent variable divided by the rating. This analysis suggests that the increased likelihood and amount punished is mainly driven by actions rated as "neither appropriate nor inappropriate" (H3A: $\beta = 1.93$ $[0.83, 3.02]$, $z = 3.452$, $p_{fdr} = 0.001$; H3B: $\beta = 0.63$ $[0.17, 1.09]$, $z = 2.680$, $p_{fdr} = 0.015$). The contrasts for "inappropriate ratings is in the same direction but does not reach significance (H3A: $\beta = 0.37$ $[-0.65, 1.38]$, $z = 0.707$, $p_{fdr} = 0.480$; H3B: $\beta = 0.38$ $[0.17, 1.09]$, $z = 1.686$, $p_{fdr} = 0.092$). These results suggest that participants who consider only one action to be appropriate are less tolerant of gray areas, such as actions that are neither generally inappropriate nor appropriate.

Appendix B.3. Hypothesis 3 using the number of tokens

As an additional robustness test for H3, we tested whether the rate and magnitude of punishments were proportional to the number of tokens placed on the norm that received the most tokens. To do so, we ran the same regressions as in H3. As in the original regressions, we included only actions that were rated as "inappropriate" or "neither appropriate nor inappropriate," in this case according to the norm that received the most tokens. Instead of the original dummy variable, we included the number of tokens placed on that norm. In the case of ties between multiple norms, participants were excluded from this exploratory analysis. The results again are in line with the ones reported in the main text: the higher the number of tokens placed, the higher the likelihood of punishing (H3A: $\beta = .010$ $[.002, .019]$, $z = 2.899$, $p = 0.017$) and the amount punished (H3B: $\beta = .005$ $[.001, .009]$, $t(593.7) = 2.899$, $p = 0.004$). According to this last model, for every 10 more tokens placed on the norm with the most tokens, punishments increase by an average of 0.05 points.

Appendix C. Additional Figures

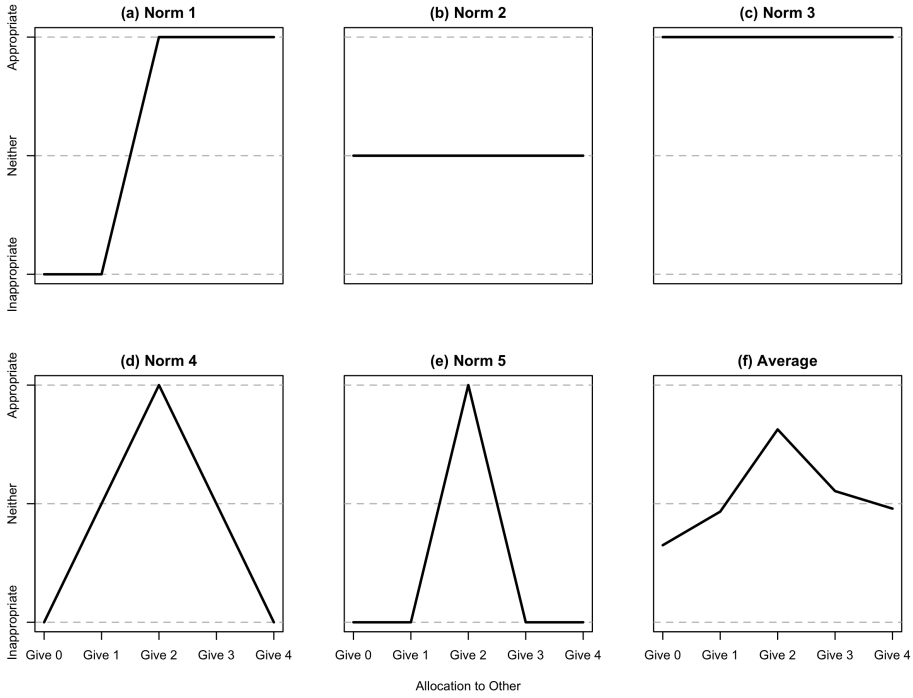

Figure C.9: Top 5 Most Commonly Drawn Norms, and the Belief-Weighted Average Norm, Standard Dictator Game.

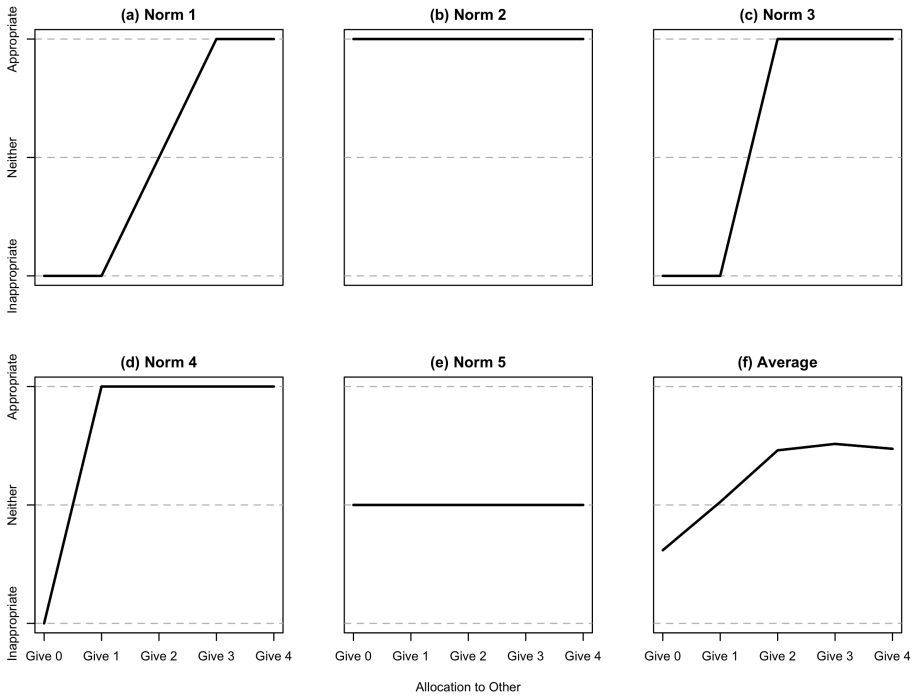

Figure C.10: Top 5 Most Commonly Drawn Norms, and the Belief-Weighted Average Norm, Charity Dictator Game.

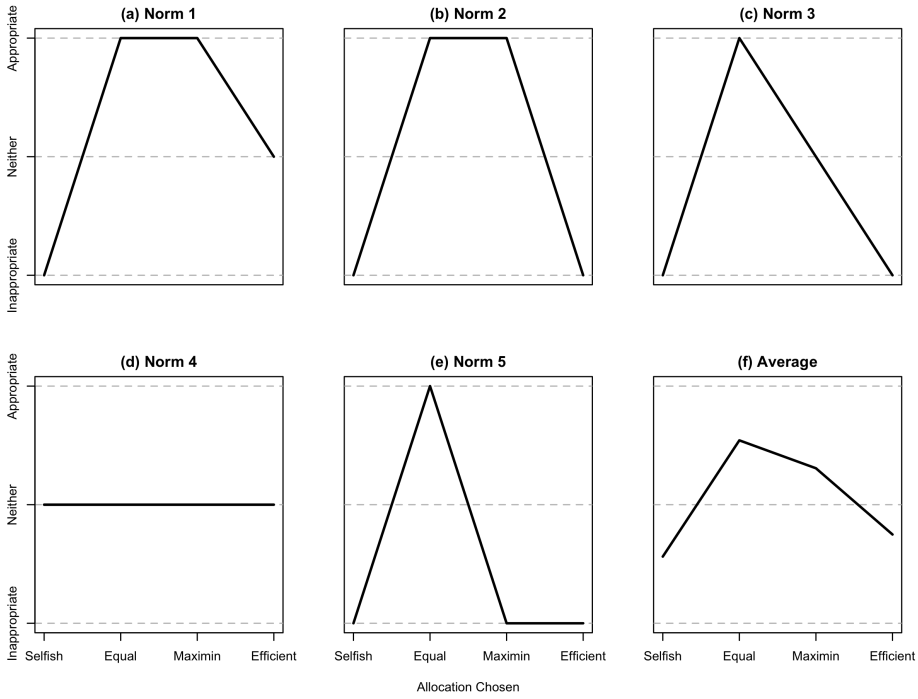

Figure C.11: Top 5 Most Commonly Drawn Norms, and the Belief-Weighted Average Norm, Standard Allocation Game.

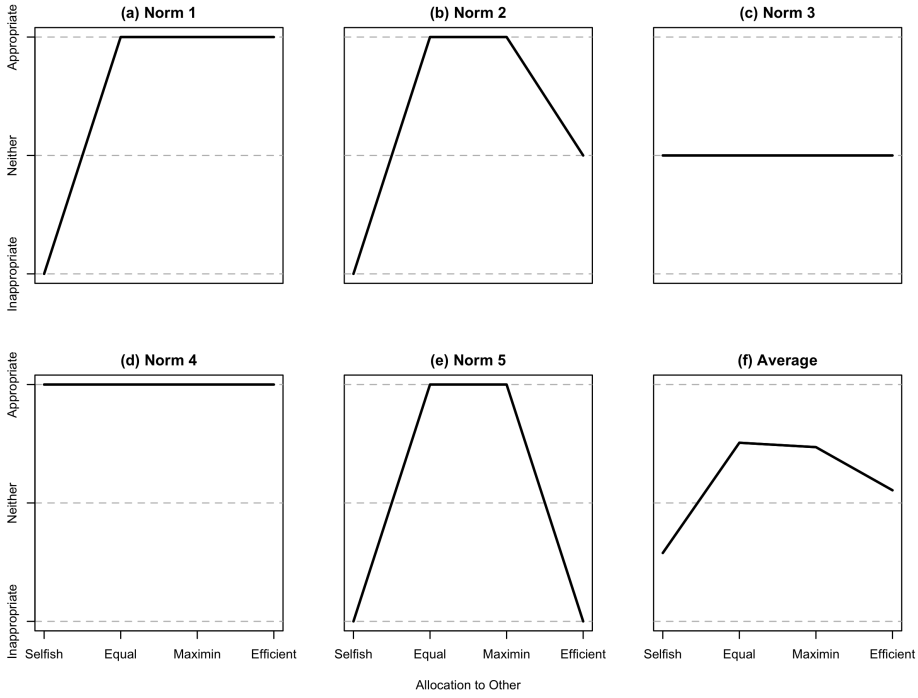

Figure C.12: Top 5 Most Commonly Drawn Norms, and the Belief-Weighted Average Norm, Charity Allocation Game.

Appendix D. Supplementary Tables

Appendix E. Experiment 2 amendements and exploratory analyses

Appendix E.1. Pre-registered miswording

To replicate the results in the first experiment, we consider punishment for all those actions that the participant rated *at least in one view* as "inappropriate" or "neither appropriate nor inappropriate". Due to a mistake in wording, the preregistration instead reported that we should only include those actions *always* rated as "inappropriate" or "neither appropriate nor inappropriate". In addition to deviating from the original analyses, this formulation reduces the sample available by more than 54%.

Despite these important differences, we report here the results as they provide a very high threshold for our tests. Indeed, to test for differences in punishment decisions, we narrowed our analyses to any action that was considered not rated as "appropriate" in at least one view reported by the participant. Our reasoning was that these views, even if not personally held by the respondent, could still influence punishment decisions. As a result, however, we also included in our analyses actions that the respondent would consider appropriate and, thus, artificially reducing the proportion of punishment decisions. Unfortunately, it is not possible to determine which particular view participants held, as we deliberately omitted this question for fear of spillover effects (i.e., eliciting personal beliefs could have influenced punishment decisions and vice versa, making responses difficult to interpret). However, if we assume that the respondent holds one of the reported views, restricting our analyses to actions that are not rated "appropriate" by *all* views arguably excludes all actions that the respondent considers appropriate. This selection is extremely conservative as it may still exclude actions that the participant might consider "inappropriate," but despite a considerable reduction in power, it could still reveal whether the trends are in the expected direction.

This is indeed the case and, for one analysis, the results still retain statistical significance. Indeed, while punishment rate is no more significantly different between one-view participants and the rest of the sample ($z = 1.511$,$p = .131$), punishment magnitude maintains its significance ($t(153.9) = 2.259$,$p = .025$). When comparing different treatments, there is no longer a significant difference (all $p > .260$), but differences trend in the same direction. Thus, despite the strong constraints, we find suggestive evidence that the effect could be found even when controlling for personal views. However, we leave this analysis for future studies.

Appendix E.2. Anti-social punishment

As an exploratory analysis, we repeat the pre-registered tests including also punishment decisions for those actions that participants rated as appropriate, namely anti-social punishment decisions. The results are largely consistent with the results reported in the main text. Punishment rate and magnitude are both significantly different between one-view participants and other participants (rate: $z = 3.319$, $p < .001$; magnitude: $t(382) = 4.153$, $p < .001$). Punishment rates also differ significantly between the Norm-Drawing treatment and the other two treatments (Norm-Drawing versus Krupka-Weber: $z = 2.661$, $p = .023$, Norm-Drawing versus control: $z = 2.141$, $p = .048$), but not between Krupka-Weber and control ($z = -0.439$, $p = .660$). Furthermore, we find no significant differences across treatments in terms of punishment magnitude (all $p > .234$). Equivalence tests suggest that the amount punished is similar across treatments.

Appendix E.3. View diversity

The categorization of participants into those who report only one view and those who report multiple views does not consider that participants who report multiple views may report views that are very similar to each other. We repeat the analyses of the second experiment using a measure of how

diverse the views reported in the Norm-Drawing Task are. To measure the diversity of views, we compute the following formula (see also Panizza et al., 2024):

$$d = 2 \times \sqrt{\sum_i s_i \delta_i} \tag{E.1}$$

Where s_i is the proportion of people that according to the participant hold the ith view and δ_i is the normalized Euclidean distance between view i, and the weighted average of all views reported:

$$\delta = \sqrt{\sum_j (a_j - \overline{a}_j)^2 / N_a} \tag{E.2}$$

Where a_j is the appropriateness rating for the jth action of the view (appropriate = 1; neither = 0.5; inappropriate = 0), $\overline{a}_j$ is the weighted average rating of the jth action among all the views reported, and N_a is the number of actions for the given issue. The diversity index d ranges from 0 to 1. It is 0 when only one view is reported and 1 when all views are maximally distant from the average view.

Using this measure instead of the one-view dummy does not change the original results: as the diversity of views increases, the likelihood and amount of punishment both decrease (likelihood: $z = -4.593$, $p < .001$; amount: $t(392.7) = -3.645$, $p < .001$).

Appendix E.4. Increase in punishment

The results reported in the main text do not allow us to tell whether the increase in money spent for punishment is due to the increased number of punishment decisions, an increased amount spent per participant, or both. To investigate these explanations, we repeat the analyses using a hurdle regression model that allows us to separate them. The hurdle model was chosen as it yielded the lowest Akaike Information Criterion compared to Poisson and Zero-Inflated Poisson regressions in all tests. The results reveal that the difference in punishment between participants reporting one view and the rest of the sample is significantly different in terms of likelihood ($z = 7.694$, $p < .001$) but punishers seem not to punish in different amounts ($z = 1.468$, $p = .142$). Similarly, differences between treatments seems to be explained by different rates of punishment decisions (control vs. Krupka-Weber: $z = -2.268$, $p = .035$; control vs. Norm-Drawing: $z = 1.793$, $p = .073$; Krupka-Weber vs. Norm-Drawing: $z = 4.150$, $p < .001$), rather than by different individual amounts (all $p > .525$). Taken together, these results suggest that differences in the amount spent on punishment are mostly driven by the number of participants who decide to punish rather than by an increase in the amount of punishment per see.