WORKING PAPERS

N° 1515

March 2025

"On Injunctive Norms: Theory and Experiment"

Pau Juan-Bartroli

Toulouse School of Economics

On Injunctive Norms: Theory and Experiment

Pau Juan-Bartroli*

pau.juanbartroli@tse-fr.eu

March 2025
First Draft: October 2023

Abstract

Recent studies show that individuals' decisions are shaped by their perceptions of socially appropriate behavior. However, these studies elicit such perceptions without developing a theory of how individuals determine social appropriateness. This paper proposes a framework in which social appropriateness judgments emerge endogenously from a utility function that combines payoff maximization with universalization reasoning. The framework allows one to compute the social appropriateness of any action without relying on beliefs, preferences, or choices. I test the theory's predictions using evidence from past studies and new data from a laboratory experiment.

JEL Classification: C91, D91

Keywords: Social Norms, Morality, Lab Experiments, Social Preferences.

*I am indebted to Ingela Alger for her support and guidance during this project. I also thank Amirreza Ahmadzadeh, Jean-François Bonnefon, Eric van Damme, Anna Dreber, Tore Ellingsen, Alice Hallman, Jona Krutaj, Gerard Maideu-Morera, Moritz Loewenfeld, Sophie Moinas, Esteban Muñoz-Sobrado, Jan Potters, Sigrid Suetens, Roberto Weber, Jörgen Weibull, and especially Enrico-Mattia Salonia, Sébastien Pouget and Boris van Leeuwen for their valuable discussions. I acknowledge the audiences of the PhD workshop at the Toulouse School of Economics, the 10th Warwick Economics PhD Conference, the Stockholm School of Economics Brown Bag Seminar, the Tilburg Experimental Workshop, the 2022 and 2023 ESA World meetings and the 2023 EEA Barcelona meeting for their useful feedback. The study was approved by the ethical committee of the Toulouse School of Economics and pre-registered with the Open Science Foundation (https://osf.io/g768h/). Finally, I acknowledge funding from the Toulouse School of Economics doctoral school and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789111 - ERC EvolvingEconomics).

1 Introduction

Decades of experimental economics have documented a range of motivations that influence behavior, such as altruism (Becker, 1976), warm glow (Andreoni, 1990), reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004), inequity aversion (Fehr and Schmidt, 1999), efficiency preferences (Charness and Rabin, 2002), and reputation concerns (Bénabou and Tirole, 2006; Ellingsen and Johannesson, 2008). However, some observations remain unexplained.[1]

Several studies propose models of norm-dependent preferences, arguing that they offer greater explanatory power than social preference models (e.g., Cappelen et al., 2007; López-Pérez, 2008; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2023).[2] Individuals with norm-dependent preferences seek to maximize their monetary payoffs while behaving in a socially appropriate manner. Recent studies elicit *injunctive norms*, the shared belief of how socially appropriate a given behavior is, in the lab. This framework has been shown to be consistent with observations that are otherwise difficult to reconcile with alternative theories. However, since norms can vary across contexts, these studies struggle to predict injunctive norms *ex-ante* and to explain individuals' evaluations *ex-post*. Having a theory on how individuals assess what is socially appropriate is therefore essential for evaluating the validity of norm-dependent preferences.

In this paper, I propose a framework in which social appropriateness judgments emerge endogenously from a utility function that combines payoff maximization and universalization reasoning (Brekke et al., 2003; Alger and Weibull, 2013). A key advantage of this approach is that, rather than relying on elicited norms, it provides an explicit function that ranks actions by their social appropriateness. The resulting injunctive norm, determined endogenously by the interaction, can be computed across various settings with minimal assumptions. Additionally, it provides a thought process on how individuals determine what is socially appropriate and what is not.

The theory is grounded in *universalization reasoning*, individuals' tendency to consider

[1] For example, Latane and Darley (1968) show that individuals are less likely to volunteer when others are present. Cherry et al. (2002) find that dictators transfer less to recipients when they have to work to generate their endowment. List (2007) shows that dictators' transfers decrease when they have the option to take from recipients' endowments.

[2] These models fall into two categories: those where norms are exogenous or elicited (Cappelen et al., 2007; Krupka and Weber, 2013), and those where norms are endogenous to the interaction (López-Pérez, 2008; Kimbrough and Vostroknutov, 2023). While models in the first category are useful, they may be criticized for having too many degrees of freedom, limiting their predictive power. I compare the proposed norm framework with the second category of models in Section 7.

what would happen if everyone acted as they do.[3] When evaluating how socially appropriate an action is, individuals consider what their material payoff would be if everyone also chose it. The most socially appropriate action is the one that yields the highest payoff if that action were to become a universal law. Such reasoning is reminiscent of Kant's categorical imperative (Kant, 1785), "act only on the maxim that you would at the same time will to be a universal law".

I present four remarks to clarify the framework. First, I focus on injunctive norms, what one should do, rather than *descriptive norms*, what most people do. Although descriptive norms explain individuals' behavior (Cialdini et al., 1990; Köbis et al., 2015), several studies have shown that a preference for conformity to injunctive norms is sufficient to explain a considerable amount of the variation (Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016). Second, consistent with previous literature, the injunctive norm prescribes the social appropriateness of behavior rather than outcomes (Elster, 1989; Krupka and Weber, 2013). Third, injunctive norms are homogeneous across individuals. Finally, individuals conform to injunctive norms because they derive utility from adhering to them.

The resulting model's main strengths are its simplicity and portability. It is simple because it allows us to compute the social appropriateness of any action with minimal degrees of freedom, without relying on beliefs, preferences, or choices. It is portable because because it allows us to compute the injunctive norm in many settings. The model rationalizes several observations, such as the effect of taking options in dictator games (List, 2007; Bardsley, 2008), the bystander effect (Latane and Darley, 1968; Fischer et al., 2011), and the effect of heterogeneous earnings and productivities in dictator games (Konow, 2000; Cherry et al., 2002; Oxoby and Spraggon, 2008).[4] In Section 7, I argue that other models of norm-dependent preferences have difficulties in explaining these observations.

I conduct a pre-registered lab experiment to test the predictions of the theory. The

[3]The concept of universalization reasoning is often referred to synonymously as *Kantian morality* (Alger and Weibull, 2013; van Leeuwen and Alger, 2024). I use the former terminology because it provides a more transparent description of the reasoning individuals follow, avoiding confusion with Kantian deontological reasoning, in which only one action is considered morally permissible (see footnote 38 for more details). Additionally, it aligns with the terminology used in psychology (e.g., Levine et al., 2020 and Kwon et al., 2023).

[4]Although the theory can explain the shift in norms in the dictator game with taking option (Krupka and Weber, 2013), it cannot explain the choice set effects reported in List (2007). This reversal (i.e., individuals choosing a positive transfer in the dictator game but zero when taking options are included) violates the Weak Axiom of Revealed Preference (WARP). Since complete and transitive preferences cannot accommodate this violation, the standard model cannot explain it. In Appendix D, I show that a modified version of the utility function can explain both norms and behavior.

experiment consists of seven game protocols.[5] In each protocol, participants are divided into two variants that differ in only one dimension. For example, in the dictator game with earnings, the endowment is generated by either the dictator or the recipient. I elicit the injunctive norm in each variant using the method introduced in Krupka and Weber (2013). This design allows me to compare (i) the injunctive norm elicited in a variant with its corresponding theoretical prediction and (ii) how changes in a variant's dimension affect both the elicited and predicted norm. Additionally, I use data from the experimental questionnaire to construct two measures of universalization reasoning (see Appendix C). These measures suggest that universalization reasoning plays a central role in individuals' elicitations.[6]

The theory applies to all games. In this paper, I test the predictions of the theory in three canonical classes of games. First, I examine symmetric two-player games with two actions, in which, in monetary payoffs, either (i) both actions can be part of a Nash equilibrium, and these equilibria are Pareto-ranked (e.g., stag hunt game), or (ii) one action is strictly dominant and leads to a Pareto inferior outcome (e.g., prisoner's dilemma). The theory predicts that individuals find it more socially appropriate to select the action that, if universalized, gives a higher material payoff. This implies that individuals evaluate the action part of the Pareto-dominant Nash equilibrium as more socially appropriate and that selecting a strictly dominant action may be socially inappropriate. The elicited norms from the lab experiment support these two predictions.

Second, I examine the standard dictator game (Forsythe et al., 1994), the dictator game with taking options (List, 2007), and dictator games with production (Konow, 2000). The theory predicts that the most socially appropriate transfer is proportional to individuals' relative efforts. As a result, the equal split is the most socially appropriate transfer in the standard dictator game but not in dictator games with production.[7] The elicited norms provide mixed support for the theory's predictions. On the one hand, the theory explains changes in the elicited norms when (i) a taking option is included, (ii) the recipient earns the endowment, or (iii) differences in production arise from exogenous factors. On the other hand, it fails to account for (i) the asymmetry in the standard dictator game and (ii) the equal split being the most socially appropriate transfer when the dictator exerts more

[5]The pre-registration and experiment instructions are available at https://osf.io/g768h/

[6]Specifically, (i) participants rated a universalization statement as the most relevant for their evaluations, and (ii) a measure of universalization reasoning (elicited using the Oxford Utilitarianism Scale from Kahane et al., 2018) is positively correlated with a measure of rule-following (elicited from Kimbrough and Vostroknutov, 2018).

[7]Therefore, individuals' legitimacy over the endowment is endogenous to the interaction. This contrasts with most prominent models in the literature, where fairness ideals are typically exogenous (Cappelen et al., 2007).

effort.

Finally, I consider the linear public goods games (Isaac and Walker, 1988) and the volunteer's dilemma (Diekmann, 1985). While the theory rationalizes several observations, it fails to capture key patterns in the elicited norms. For instance, in the linear public goods game, the theory predicts that it is socially inappropriate to contribute to the public account when doing so is socially inefficient. However, I document a robust positive relationship between contributions to the public account and social appropriateness, even when these contributions are socially inefficient.

While the empirical evidence supports many of the theory's predictions, it fails to account for certain results, suggesting that universalization reasoning alone is insufficient to explain all variations in individuals' evaluations. To explain these observations, I extend the theory to incorporate both universalization reasoning and social preferences in determining social appropriateness. I show that this extended theory accounts for most results unexplained by the simpler version and offers a framework for analyzing norm heterogeneity.

Related Literature: This paper contributes to the literature on social norms (Cialdini et al., 1990; Bicchieri, 2005). Krupka and Weber (2013) propose a method for eliciting injunctive norms which has been widely adopted.[8] I contribute to this literature in two ways. First, I propose a framework that allows one to compute the social appropriateness of any action without relying on beliefs, preferences, or choices. This offers a potential explanation for how individuals evaluate what is socially appropriate. Second, I provide new evidence on injunctive norms across several interactions, such as the stag hunt game, prisoners' dilemma, volunteer's dilemma, dictator game with joint production, and dictator game with earnings. The dataset can be used by other researchers to test alternative theories and predictions.

Several theories of social norms have been proposed (e.g., Cappelen et al., 2007; López-Pérez, 2008; Kessler and Leider, 2012). In these theories, norms are typically treated as exogenous, with the focus placed on the effects of norm conformity rather than on the determinants of the norm itself.[9] Ellingsen and Mohlin (2023) introduce a theory that classifies social duties according to their *conditionality* (i.e., if actions' consequences affect their

[8]For other studies using this method see Gächter et al. (2013), Veselỳ (2015), Kimbrough and Vostroknutov (2016), Krupka et al. (2017), Bašić and Verrina (2023), Ellingsen and Mohlin (2023) and Krupka et al. (2022). For more detailed discussions of the method, see Erkut (2020) and Nosenzo and Görges (2020). See Bicchieri and Xiao (2009) for an alternative method to elicit norms.

[9]For other theoretical models of social norms see Akerlof (1980), Lindbeck et al. (1999), Brekke et al. (2003), Levitt and List (2007) and Huck et al. (2012).

classification) and their *strictness* (i.e., if they provide a minimum acceptable standard of behavior). The key differences are as follows. First, their model is applied to situations without strategic interaction, while the proposed framework also extends to simultaneous games with an arbitrary number of players. Second, in Ellingsen and Mohlin (2023), the endowment's entitlement is partially exogenous, while in my framework, injunctive norms emerge endogenously from the interaction. The closest paper is Kimbrough and Vostroknutov (2023), which, to my knowledge, is the only other study that introduces a theory of the *content* of the injunctive norms. The main difference between the two papers is that Kimbrough and Vostroknutov (2023) define injunctive norms at the outcome level, while I define them at the action level. Additionally, I test the theory's predictions using new data from a lab experiment. In Section 7, I discuss the differences with López-Pérez (2008) and Kimbrough and Vostroknutov (2023) in more detail.

Finally, this paper contributes to the literature that uses universalization reasoning to explain individuals' behavior (Brekke et al., 2003; Alger and Weibull, 2013).[10] These preferences have been used in theoretical studies (Sarkisian, 2017; Eichner and Pethig, 2021; Alger and Laslier, 2022; Muñoz Sobrado, 2022; Juan-Bartroli and Karagözoğlu, 2024) and tested in laboratory settings (Miettinen et al., 2020; Van Leeuwen and Alger, 2024; Alger and Rivero-Wildemauwe, 2024). Levine et al. (2020) show that universalization reasoning explains individuals' (unincentivized) moral judgments in threshold games. This paper extends their findings by introducing a portable theoretical framework and conducting an incentivized experiment across various interactions. More broadly, I contribute to this literature by showing that models of universalization reasoning can be interpreted as models of norm-dependent preferences.[11]

Outline: The remainder of the paper is organized as follows: In Section 2, I present the theoretical framework. In Section 3, I describe the experimental design. In Section 4, I provide the theoretical predictions. In Section 5, I display the experimental results. In Section 6, I introduce the extended theoretical framework. In Section 7, I compare the theory with alternative models. In Section 8, I conclude.

[10] For the evolutionary foundations of universalization reasoning see Alger (2023). For its axiomatic foundations, see Salonia (2023). Universalization reasoning is also related to Sugden (2003)'s *team reasoning*, where individuals consider what they should do as a group rather than focusing on individual decisions.

[11] See Cox et al. (2019) and Roemer (2010) for studies that consider moral concerns in alternative ways.

2 The theoretical framework

In this section, I show how to reformulate models of norm compliance as models where individuals combine material payoff maximization and universalization reasoning.

2.1 The normalized injunctive norm

Consider a game with $n \geq 2$ individuals, and let X be the common set of strategies, assumed to be non-empty and compact. A strategy profile is denoted by $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, or by $\boldsymbol{x} = (x_i, x_{-i})$ if viewed from individual i's perspective. A strategy profile where everyone chooses the same strategy x_i is denoted by $\boldsymbol{x_i} = (x_i, ..., x_i)$. Let $\pi : X^n \to \mathbb{R}$ be individuals' material payoff function, which represents how individuals evaluate the resulting monetary payoffs. More concretely, $\pi(x_i, x_{-i}) \equiv v(\widetilde{\pi}(x_i, x_{-i}))$, where $\widetilde{\pi}(x_i, x_{-i})$ is individual $i's$ monetary payoff under strategy profile (x_i, x_{-i}), and $v : \mathbb{R} \to \mathbb{R}$ represents how individuals evaluate such monetary payments. I assume that individuals have *homo moralis* preferences (Alger and Weibull, 2013; Alger and Weibull, 2016). That is, individual i's utility function is

$$u_i(x_i, x_{-i}) = (1 - \kappa_i)\pi(x_i, x_{-i}) + \kappa_i \pi(\boldsymbol{x_i}). \tag{1}$$

Interpreting the last term in (1) as a representation of Kant's categorical imperative (Kant, 1785), I refer to $\kappa_i \in [0,1]$ as individual i's degree of morality. The larger κ_i, the larger the weight individual i attaches to $\pi(\boldsymbol{x_i})$, which represents individual i's material payoff in the (hypothetical) scenario where everyone selects x_i. Lemma 1 shows that equation (1) can be rewritten as a model of norm conformity, in which $\pi(\boldsymbol{x_i})$ represents the ranking of socially appropriate strategies.

Lemma 1. *Let* $\overline{x} \in \arg\max_{x \in X} \pi(x, ..., x)$ *and* $\underline{x} \in \arg\min_{x \in X} \pi(x, ..., x)$. *Then, (1) represents the same preferences as*

$$\widetilde{u}_i(x_i, x_{-i}) = \widetilde{v}(\pi(x_i, x_{-i})) + \gamma_i \widetilde{N}(\boldsymbol{x_i}), \tag{2}$$

with

- $\widetilde{v}(\pi(x_i, x_{-i})) \equiv 2\frac{\pi(x_i, x_{-i}) - \pi(\underline{x})}{\pi(\overline{x}) - \pi(\underline{x})} - 1$,
- $\gamma_i \equiv \frac{\kappa_i}{1 - \kappa_i}$,
- $N(x_i) \equiv \pi(\boldsymbol{x_i})$,

- $\widetilde{N}(\boldsymbol{x_i}) \equiv 2\frac{\pi(\boldsymbol{x_i})-\pi(\underline{x})}{\pi(\overline{x})-\pi(\underline{x})} - 1 \in [-1,1]$.

Proof. See Appendix B. □

Equation (2) is alike to the model in Krupka and Weber (2013). In this case, individual i cares about his material payoff under strategy profile (x_i, x_{-i}) and the degree to which x_i is collectively perceived as socially appropriate, denoted $\widetilde{N}(x_i)$.[12] The parameter $\gamma_i \geq 0$ is individual i's degree of *norm-following*: the extent to which the individual cares about adhering to injunctive norms (Kimbrough and Vostroknutov, 2016; Kimbrough and Vostroknutov, 2018). The function $\widetilde{N}(x_i) \in [-1,1]$ is the *normalized injunctive norm*: the shared belief on how socially appropriate selecting action x_i is. The key distinction between (2) and the framework in Krupka and Weber (2013) is that here $\widetilde{N}(x_i)$ can be directly derived from preferences over material outcomes. In contrast, Krupka and Weber (2013) treats social appropriateness rankings as an independent normative construct that cannot be inferred solely from material preferences and thus requires separate elicitation (see Section 3).

I discuss four remarks. First, I distinguish between the injunctive norm, $N(x_i)$, and the normalized injunctive norm, $\widetilde{N}(x_i)$. The latter maintains the ranking prescribed by the former while imposing $\widetilde{N}(\overline{x}) = 1$ and $\widetilde{N}(\underline{x}) = -1$.[13] Second, $\widetilde{N}(x_i)$ is homogeneous across individuals; everyone shares the same normalized injunctive norm, even when they differ in their willingness to conform to it. Third, contrary to Krupka and Weber (2013), equation (1) normalizes both the utility from the material payoff and the injunctive norm. If I were only to normalize the latter (i.e., having $\widetilde{v}(\pi(x_i, x_{-i})) = \pi(x_i, x_{-i})$ and $\widetilde{N}(x_i) \in [-1,1]$), then small increases in the stakes of the interaction would make individuals entirely selfish. Fourth, the injunctive norm does not depend on others' strategies, beliefs about others' strategies, or the idiosyncratic parameters of individuals' utility function.

In asymmetric games, the universalization counterfactual (i.e., what if everyone did that?) cannot be computed, as certain strategies may be unavailable to some individuals. To address this limitation, I consider the ex-ante symmetric version of the game, where

[12] Krupka and Weber (2013) only consider the case with $n = 2$. Here, I consider the extended version with $n \geq 2$.

[13] $\widetilde{N}(x_i)$ captures the idea that, when evaluating the social appropriateness of an action, we assign a *relative* rating compared to the most and least socially appropriate actions *in that situation*. Therefore, the same action may be perceived as more or less appropriate depending on the other available options (e.g., Section 5.2.2). This normalization approach is similar to the one used by Ferguson and Flynn (2016).

individuals have an equal probability of being assigned to any role and select their strategies behind a veil of ignorance.[14]

Assumption 1. *When the game is asymmetric, the injunctive norm is computed behind the veil of ignorance.*

Assumption 1 does not mean that the theory cannot be applied to asymmetric games without role uncertainty. The reason is that even if the individual is certain that he will only be playing in a certain role, he may still consider what would have happened if the roles had been reversed. For example, an individual assigned to be the dictator, who knows he will not be the recipient, may understand that the roles were randomly assigned and that, in another scenario, they could have been reversed. Therefore, even when he knows he will not be the recipient, he may still consider what would have happened if the roles were reversed and his recipient chose the same transfer as he did.[15]

Assumption 2 imposes structure on the function individuals use to evaluate the monetary payoffs resulting from the interaction.

Assumption 2. *v is increasing, strictly concave and differentiable.*

Assumption 2 implies that individuals exhibit decreasing marginal utility of money and are risk-averse behind the veil of ignorance. The theory's predictions remain unchanged when considering $v(x) = x$, except in the case of dictator games (see footnote 21 for more details).

3 The Experimental Design

To test the theory's predictions, I conduct a lab experiment that follows the methodology introduced in Krupka and Weber (2013). The experiment consists of seven distinct situations, each with two variants that differ in a single dimension. Participants read the descriptions of seven variants where Person A has to choose between different actions. The participants' task consists of rating the social appropriateness of the different actions

[14] This approach is common in studies examining universalization reasoning in asymmetric games (e.g., Alger and Weibull, 2013; Muñoz Sobrado, 2022; Wildemauwe, 2023; Salonia, 2023; van Leeuwen and Alger, 2024).

[15] An alternative interpretation of Assumption 1 is that the individual judges the action's appropriateness as an impartial spectator who cares equally about the two players. The spectator evaluates a transfer as the sum between the two players' utilities. In Appendix A, I discuss how the theory can be applied to analyze choice data of third parties.

on a 6-point scale (as in Bašić and Verrina, 2023). Participants' answers are converted to numerical scores (as in Krupka et al., 2017).[16] The main variable of interest is the average social appropriateness of an action, which is defined as the average of the numerical scores given by the participants.[17]

Participants earn a (€7) bonus payment if their evaluation of a randomly selected action is the same as the most selected by other participants in their session. This gives participants incentives to truthfully report their beliefs on their session's most common evaluation. After the participants read the description of a variant, they had to answer several comprehension questions about it. I refer the reader to the experimental instructions to see how the variants were presented to the participants. The situations evaluated in the experiment are:

1. **Coordination game with two Pareto-ranked Nash equilibria**. Participants evaluate how socially appropriate they find Person A selecting each of the two actions. The two variants differ in the payoff when coordinating in the Pareto-dominant equilibrium. (See Section 4.1.1)

2. **Stag hunt game**. Participants evaluate how socially appropriate they find Person A selecting each of the two actions. The two variants differ in the payoff when coordinating in the payoff dominant equilibrium. (See Section 4.1.2)

3. **Prisoner's dilemma**. Participants evaluate how socially appropriate they find Person A cooperating or defecting. The two variants differ in the payoff of cooperating when the other individual defects. (See Section 4.1.3)

4. **Dictator game with earnings**. Participants evaluate how socially appropriate they find Person A transferring €$x \in \{0,1,2,3,4,5,6,7,8,9,10\}$ in the dictator stage. The two variants differ in whether the dictator or the recipient works to generate the endowment. (See Section 4.2.3)

5. **Dictator game with joint production**. Participants evaluate how socially appropriate they find Person A transferring €$x \in \{0,1,2,3,4,5,6,7,8,9,10\}$ in the dictator

[16] Specifically, -1 to "Very Socially Inappropriate", -0.6 to "Socially Inappropriate", -0.2 to "Rather Socially Inappropriate", 0.2 to "Rather Socially Appropriate", 0.6 to "Somewhat Socially Appropriate", and 1 to "Very Socially Appropriate".

[17] As a secondary variable, I compute the fraction of participants that consider the action to be "Appropriate to some extent" by calculating the share of the participants who evaluate the action as "Very Socially Appropriate", "Socially Appropriate", or "Rather Very Socially Appropriate" (as in Ellingsen and Mohlin, 2023).

stage. The two variants differ in whether the differences in individuals' contributions are for exogenous or endogenous reasons. (See Section 4.2.3)

6. **Linear public goods game**. Participants evaluate how socially appropriate they find Person A depositing €$x \in \{0,2,4,6,8,10\}$ to the public account. The two variants differ in the return of the public account. (See Section 4.3.1)

7. **Volunteer's dilemma**. Participants evaluate how socially appropriate they find Person A volunteering with probability $x \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The two variants differ in the group size. (See Section 4.3.2)

The situations and variants were chosen for three main reasons. First, they include interactions of various natures, such as coordination, giving, and public good games. This allows testing the theory's predictions in a diverse set of situations. Second, the two variants were selected to test the key prediction of the theory in each situation. Third, they provide new evidence of the injunctive norms in several interactions.

The randomization of participants is conducted at the individual level, satisfying four conditions: (i) participants evaluate exactly one variant of each of the seven situations; (ii) for any situation, each participant is equally likely to evaluate any of the two variants; (iii) situations are presented in random order; and (iv) each variant is evaluated by half of the participants in the session.[18]

After the norm-elicitation stage, participants completed a questionnaire that included a norm-following task (Kimbrough and Vostroknutov, 2018) and the Oxford Utilitarianism Scale (Kahane et al., 2018). Appendix C describes these two tasks in detail, contains tables with the ratings given by participants in each variant and several robustness and secondary tests.

Procedures: The experiment was conducted in French at the lab of the Toulouse School of Economics with 203 participants. It was programmed with oTree (Chen et al., 2016), and participants were recruited with ORSEE (Greiner, 2015). A total of 18 sessions, which lasted 55 minutes on average, were conducted between the 15th of March and the 21st of March of 2023.[19] Participants earned an average of €10.08 (including a €5 participation fee), with a minimum of €6 and a maximum of €14. Participants were primarily French

[18]Note that (iv) does not guarantee that each variant is evaluated by the same number of participants at the end of the experiment since sessions were not restricted to an even number of participants.

[19]One of the sessions was cancelled at the beginning of the experiment because of internet issues. I do not consider this session for the analysis of the paper.

(82%), female (57%), studying economics-related majors (45%) and had 20.78 years old on average.

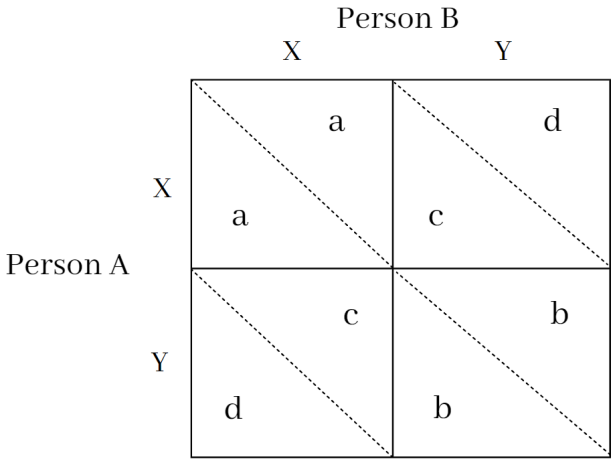
4 Theoretical predictions

In this section, I describe the situations considered, derive the theory's predictions, and discuss how these predictions are tested. I divide this section into three parts. In Section 4.1, I examine one-shot symmetric two-player games with two actions. In Section 4.2, I analyze dictator games. In Section 4.3, I consider public goods games.

4.1 One-shot symmetric two-player games

I begin by computing the injunctive norm in a symmetric two-player game with two actions and arbitrary payoffs. Figure 1 presents the material payoffs received by individuals for each action combination. I denote this game as $G = (a,b,c,d)$.

Figure 1: Two-player symmetric game with two actions.



Under the proposed norm, individuals assess the social appropriateness of an action based on their material payoff when the other individual also selects it. Proposition 1 characterizes the social appropriateness of choosing actions X and Y.

Proposition 1. *Let G be a symmetric two-player game with two actions, X and Y, and with $\pi(X,X) = a$, $\pi(X,Y) = c$, $\pi(Y,X) = d$, and $\pi(Y,Y) = b$. The social appropriateness of selecting actions X and Y is given by $N(X) = a$ and $N(Y) = b$.*

Proof. All the proofs are in Appendix B. □

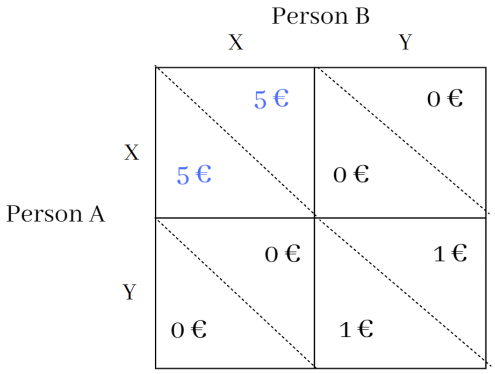
The theory has three predictions. First, when $a > b$ (resp. $a < b$), selecting X is more (resp. less) socially appropriate than Y. Second, changes in c and d do not affect the social appropriateness of selecting X and Y. Finally, the social appropriateness of selecting actions X and Y increases with a and b, respectively.[20] Proposition 1 implies that when (X,X) and (Y,Y) are Pareto-ranked Nash equilibria, selecting the action that may implement the Pareto-dominant Nash equilibrium is more socially appropriate. Additionally, selecting a strictly dominant action may be socially inappropriate. I consider three games: (i) coordination game with two Pareto-ranked Nash equilibria ($a > b > c = d$), (ii) stag hunt game ($a > b = d > c$), and (iii) prisoner's dilemma ($d > a > b > c$).

4.1.1 Coordination game with two Pareto-ranked Nash equilibria

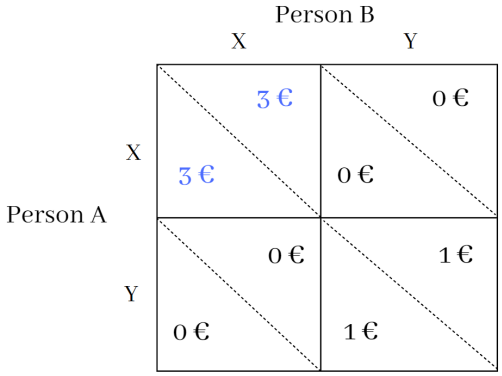
Figure 2 presents the two variants evaluated in the experiment. These describe situations where individuals prefer to coordinate rather than not coordinate, but if they do so, they prefer to do it for one specific action. Specifically, (X,X) and (Y,Y) are Pareto-ranked Nash equilibria, with (X,X) being Pareto-dominant. The only difference between the variants is the payoff when both individuals select X.

Figure 2: Coordination game.

(a) Coordination 1



(b) Coordination 2



Predictions (I) In both variants, selecting X is more socially appropriate than selecting Y. (II) Selecting X is more socially appropriate in *Coordination 1* than in *Coordination 2*. (III) Selecting Y is equally appropriate in both variants.

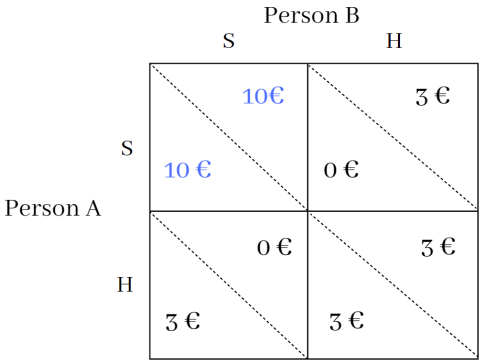
[20] As pre-registered, I use $N(x_i)$, and not $\widetilde{N}(x_i)$, to derive the predictions in two-action games. If I were to use $\widetilde{N}(x_i)$, the normalized appropriateness of the actions would only take values of -1 or 1, making comparisons within and across games more difficult.

4.1.2 Stag hunt game

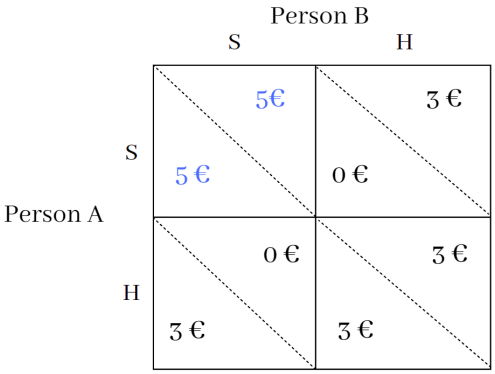
Figure 3 presents the two variants evaluated in the experiment. As before, (S,S) and (H,H) are Pareto-ranked Nash equilibria. The key difference from the previously considered game is that individuals can guarantee themselves the payoff of the Pareto-dominated Nash equilibrium by selecting H. Thus, while (S, S) is *payoff dominant*, (H, H) is *risk dominant* (Rydval and Ortmann, 2005). The only difference between the two variants is the payoff obtained when both individuals select S.

Figure 3: Stag hunt game.

(a) Stag Hunt 1



(b) Stag Hunt 2



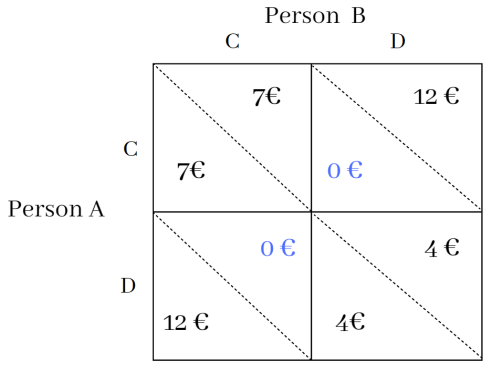
Predictions (I) In both variants, selecting S is more socially appropriate than selecting H. (II) Selecting S is more socially appropriate in *Stag Hunt 1* than in *Stag Hunt 2*. (III) Selecting H is equally appropriate in both variants.

4.1.3 Prisoner's dilemma

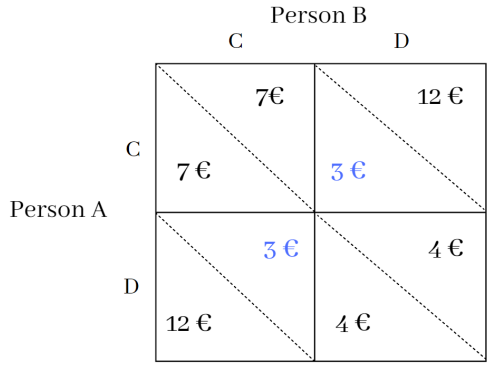
Figure 4 presents the two variants evaluated in the experiment. These describe situations in which individuals would benefit from mutual cooperation, yet each has an incentive to deviate. In this game, (D,D) is the unique Nash equilibrium, and D is a strictly dominant action. The only difference between the two variants is the payoff when one individual cooperates and the other defects.

Figure 4: Prisoner's dilemma.

(a) Prisoner's dilemma 1



(b) Prisoner's dilemma 2



Predictions (I) In both variants, selecting C is more socially appropriate than selecting D. (II) Selecting C is equally appropriate in both variants. (III) Selecting D is equally appropriate in both variants.

4.2 Dictator games

In this section, I examine the standard dictator game (Section 4.2.1), the dictator game with a taking option (Section 4.2.2), and the dictator game with production (Section 4.2.3).

4.2.1 Standard dictator game

Individuals are matched into pairs and are randomly assigned the roles of dictator or recipient. Dictators decide how to divide an endowment of $w > 0$ between themselves and their pairs (Forsythe et al., 1994). In the ex-ante symmetric dictator game, individuals have an equal chance of being assigned either role and select their transfer behind the veil of ignorance. The expected material payoff of individual 1 is

$$\pi(x_1, x_2) = \frac{1}{2}v(w - x_1) + \frac{1}{2}v(x_2), \tag{3}$$

where $x_1 \in [0, w]$ (resp. $x_2 \in [0, w]$) is the individual 1's (resp. 2's) transfer in the dictator role. The injunctive norm in the dictator game is

$$N(t) = \frac{1}{2}v(w - x) + \frac{1}{2}v(x). \tag{4}$$

Proposition 2. *The most socially appropriate transfer in the dictator game is* $x^* = \frac{w}{2}$. *Additionally, (i)* $\frac{\partial N(x)}{\partial x} > 0$ *for any* $x \in [0, \frac{w}{2})$ *and* $\frac{\partial N(x)}{\partial x} < 0$ *for any* $x \in (\frac{w}{2}, w]$, *and (ii)* $N(x) = N(w - x)$ *for any* $x \in [0, \frac{w}{2})$.

Proposition 2 shows that the most socially appropriate transfer is the equal split and that the injunctive norm is symmetric around $\frac{w}{2}$, increasing for transfers below $\frac{w}{2}$ and decreasing for transfers above it.[21]

Predictions (displayed in Figure 12): (I) $x = \frac{w}{2}$ is the most socially appropriate transfer. (II) For $x < \frac{w}{2}$, the larger the transfer, the more socially appropriate that transfer is. (III) For $x > \frac{w}{2}$, the larger the transfer, the less socially appropriate that transfer is. (IV) The injunctive norm is symmetric around $x = \frac{w}{2}$.

4.2.2 Dictator game with taking options.

Recent studies have examined modified dictator games where dictators can take from recipients' endowments. List (2007) show that dictators reduce their transfers when given the option to take \$1 from recipients' endowments. Levitt and List (2007) suggest that this observation may result from a shift in social norms due to the expanded choice set. Proposition 3 shows how the social appropriateness of *existing* alternatives changes when a *new* alternative is introduced.

Proposition 3. *Consider two games* $g \in \{A, B\}$ *such that (i)* $X_A = X$ *and* $X_B = X \cup \{x'\}$ *with* $x' \notin X$ *and (ii)* $N^A(x) = N^B(x)\ \forall x \in X$. *Let* $\overline{x}^A \in \arg\max_{x \in X} N^A(x)$ *and* $\underline{x}^A \in \arg\min_{x \in X} N^A(x)$ *denote the most and least socially appropriate strategies in game A. Let* $\widetilde{N}^g(x)$ *denote the normalized social appropriateness of strategy* x *in game* g.

- ***Case 1*** *(x' is neither the most nor the least socially appropriate action): If* $N^A(\overline{x}^A) > N^B(x') > N^A(\underline{x}^A)$, *then (i)* $\widetilde{N}^A(x) = \widetilde{N}^B(x)\ \forall x \in X$ *and (ii)* $\widetilde{N}^B(x') \in (-1, 1)$.

- ***Case 2*** *(x' is the most socially appropriate action): If* $N^B(x') > N^A(\overline{x}^A)$, *then (i)* $\widetilde{N}^A(x) > \widetilde{N}^B(x)\ \forall x \in X$ *with* $\widetilde{N}^A(x) > -1$ *and (ii)* $\widetilde{N}^B(x') = 1$.

- ***Case 3*** *(x' is the least socially appropriate action): If* $N^B(x') < N^A(\underline{x}^A)$, *then (i)* $\widetilde{N}^A(x) < \widetilde{N}^B(x)\ \forall x \in X$ *with* $\widetilde{N}^A(x) < 1$ *and (ii)* $\widetilde{N}^B(x') = -1$.

[21] When v is linear, $N(x)$ is constant in x. Thus, a strictly concave function v is required to generate these predictions.

Proposition 3 predicts that introducing an additional alternative affects the social appropriateness of existing alternatives only if the new option is more (or less) socially appropriate than *all* existing ones.[22]

Predictions (displayed in Figure 13b):[23] (I) In the dictator game with a taking option, taking \$1 is the most socially inappropriate action. (II) In both variants, transferring half of the endowment is the most socially appropriate action. (III) Except for $x = \frac{w}{2}$, the social appropriateness of any action common to both variants is higher in the dictator game with taking option than in the standard dictator game.

4.2.3 Dictator game with production

The dictator game with production consists of two stages. In the first stage, individuals are matched into pairs, and one (or both) works to generate an endowment. In the second stage, dictators divide the endowment generated at $t = 1$. Individual i chooses a strategy $\hat{x}_i = ({e_1}^i, {e_2}^i, x_i({e_1}^i, {e_2}^j))$ that specifies his effort as dictator (i.e., ${e_1}^i$) and recipient (i.e., ${e_2}^i$) at $t = 1$, and a transfer as dictator at $t = 2$ given any pair of efforts at $t = 1$ (i.e., $x_i({e_1}^i, {e_2}^j)$). Individual i's effort in role k has an associated strictly convex cost of $c({e_k}^i)$, which is assumed to be independent of individuals' identity and role. Finally, $w({e_1}^i, {e_2}^j)$ is the endowment generated at $t = 1$, which is assumed to be deterministic, increasing and concave in individuals' efforts.

Individual i's expected material payoff under strategy profile $(\hat{x}_i, \hat{x}_j)$ is:

$$\pi(\hat{x}_i, \hat{x}_j) = \frac{1}{2}\underbrace{[v(w({e_1}^i, {e_2}^j)) - x_i({e_1}^i, {e_2}^j) - c({e_1}^i)]}_{\text{Individual } i \text{ is dictator}} + \frac{1}{2}\underbrace{[v(x_j({e_1}^j, {e_2}^i) - c({e_2}^i)]}_{\text{Individual } i \text{ is recipient}}. \tag{5}$$

The social appropriateness of choosing a strategy $\hat{x} = (e_1, e_2, x(e_1, e_2))$ is given by

$$N(\hat{x}) = \frac{1}{2}v(w(e_1, e_2) - x(e_1, e_2) - c(e_1)) + \frac{1}{2}v(x(e_1, e_2) - c(e_2)). \tag{6}$$

Proposition 4. *The most socially appropriate transfer is*

$$x^*(e_1, e_2) = \min\left\{\max\left\{\frac{w(e_1, e_2)}{2} + \frac{c(e_2) - c(e_1)}{2}, 0\right\}, w(e_1, e_2)\right\}.$$

[22] This result is not driven by universalization reasoning per se, but rather by the fact that $\widetilde{N}(x)$ is normalized based on the most and least socially appropriate strategies in the given situation.

[23] In the dictator game with a \$1 taking option, we are in Case 3.

Additionally, $\frac{\partial N(x)}{\partial x} > 0$ *for any* $x \in [0, x^*)$ *and* $\frac{\partial N(x)}{\partial x} < 0$ *for any* $x \in (x^*, w(e_1, e_2)]$.

Proposition 4 shows that the most socially appropriate transfer allocates a larger share of the resources to the individual who exerts greater effort at $t = 1$. Intuitively, individuals maximize their expected utility by receiving the same material payoff in the two roles. Thus, the individual with the higher effort is compensated with a larger share of the aggregate resources. This is consistent with evidence documenting that most individuals hold meritocratic views in games with production (Konow, 2000; Cappelen et al., 2010; Luhan et al., 2019). I apply these predictions to the dictator game with earnings and with joint production.

Dictator game with earnings It is either the dictator or the recipient who works to generate the endowment (Cherry et al., 2002; Oxoby and Spraggon, 2008). In *Dictator Earns*, the effort of the recipients is restricted to zero. The injunctive norm is given by

$$N(\hat{x}) = \frac{1}{2}v(w(e_1, 0) - x(e_1, 0) - c(e_1)) + \frac{1}{2}v(x(e_1, 0)). \tag{7}$$

Corollary 1. *The most socially appropriate transfer in Dictator Earns is* $x^*(e_1) = \max\{\frac{w(e_1,0)}{2} - \frac{c(e_1)}{2}, 0\}$. *Additionally,* $\frac{\partial N(x)}{\partial x} > 0$ *for any* $x \in [0, x^*)$ *and* $\frac{\partial N(x)}{\partial x} < 0$ *for any* $x \in (x^*, w(e_1, 0)]$.

In *Dictator Earns*, the most socially appropriate transfer is below the equal split. This implies that, conditional on the same endowment at $t = 2$, the most socially appropriate transfer in *Dictator Earns* is lower than in the standard dictator game. A similar reasoning applies to deriving the injunctive norm in *Recipient Earns*.

Experimental details: Participants evaluated a situation in which either Person A (in *Dictator Earns*) or Person B (in *Recipient Earns*) worked for 30 minutes, counting 0s in tables with 0s and 1s. At the end of the 30 minutes, the worker solved 20 tables and generated €10. Participants evaluate how socially appropriate they find Person A giving €$x \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ to Person B at $t = 2$. This situation varies the role assigned to the worker while keeping the generated endowment and the worker's effort constant.

Predictions (displayed in Figure 8a): (I) Giving less than the equal split is more socially appropriate in *Dictator Earns* than in *Recipient Earns*. (II) Giving more than the equal split is more socially appropriate in *Recipient Earns* than in *Dictator Earns*. (III) The most socially appropriate transfer in *Recipient Earns* is above the equal split. (IV) The most socially appropriate transfer in *Dictator Earns* is below the equal split.

Dictator game with joint production Both individuals work to generate the endowment, and they may have different productivities. Following Konow (2000), individuals are paired and fold letters for mailing. Each individual's production is $q_i = p_i \cdot l_i$, where l_i is the number of letters folded by individual i and $p_i > 0$ is the piece-rate assigned to him. I assume that all individuals have the same ability to fold letters (i.e., ${l_k}^i({e_k}^i) = l({e_k}^i)$ $\forall i \in \{1,2\}$ and $k \in \{1,2\}$), implying that differences in output for individuals exerting the same effort arise solely from differences in their assigned piece rates. Under this assumption, the production of the pair is $w({e_1}^i, {e_2}^j) = p_i l({e_1}^i) + p_j l({e_2}^j)$. The social appropriateness of choosing strategy $\hat{x} = (e_1, e_2, x(e_1, e_2))$ is given by

$$N(\hat{x}) = \frac{1}{2} v(p_i l(e_1) + p_j l(e_2) - x(e_1, e_2) - c(e_1)) + \frac{1}{2} v(x(e_1, e_2) - c(e_2)). \tag{8}$$

Corollary 2. *The most socially appropriate transfer is* $x^*(e_1, e_2) = \min\{\max\{\frac{p_i l(e_1) + p_j l(e_2)}{2} + \frac{c(e_2) - c(e_1)}{2}, 0\}, w(e_1, e_2)\}$. *Additionally,* $\frac{\partial N(x)}{\partial x} > 0$ *for any* $x \in [0, x^*)$ *and* $\frac{\partial N(x)}{\partial x} < 0$ *for any* $x \in (x^*, w(e_1, e_2)]$.

Corollary 2 implies that when individuals exert the same effort, the most socially appropriate division is the egalitarian one. This is independent of the piece-rates assigned to them. On the other hand, when individuals fold different amounts of letters, the most socially appropriate division allocates a larger share of total production to the individual who folded more letters.

Experimental details: Participants evaluated a situation in which both Person A and Person B worked for 30 minutes, counting 0s in tables with 0s and 1s. In *Exogenous Inequality*, both individuals exerted the same effort, but Person A was assigned a higher piece-rate. In *Endogenous Inequality*, both were assigned the same piece-rate, but Person A exerted more effort. In both variants, Person A contributes €8 to the joint endowment while Person B contributes €2. This situation varies the source of inequality in contributions while keeping the total endowment and individual contributions constant.

Predictions (displayed in Figure 9a): (I) Transferring a low amount is more socially appropriate in *Endogenous Inequality* than in *Exogenous Inequality*. (II) Transferring an intermediate and high amount is more socially appropriate in *Exogenous Inequality* than in *Endogenous Inequality*. (III) In *Exogenous Inequality*, the most socially appropriate transfer is the equal split. (IV) In *Endogenous Inequality*, the most socially appropriate transfer is below the equal split.

4.3 Public goods games

In this section, I consider the linear public goods game (Section 4.3.1) and the volunteer's dilemma (Section 4.3.2).

4.3.1 Linear public goods game.

Individuals divide an endowment of $w > 0$ between a public and a private account. The private account returns 1, while the public account returns $\hat{A} \in (0,1)$ to each of the $n \geq 2$ group members. Individuals decrease their material payoff when contributing to the public account (as $\hat{A} < 1$). However, when $\hat{A} \in \left(\frac{1}{n},1\right)$, it is socially efficient to do so. Individual 1's material payoff under contribution profile $(x_1,...,x_n)$ is given by

$$\pi(x_1,...,x_n) = v(w - x_1 + \hat{A}(x_1 + ... + x_n)), \tag{9}$$

which gives the following injunctive norm:

$$N(x) = v(w - x + \hat{A}nx). \tag{10}$$

Proposition 5. *The most socially appropriate contribution to the public account is:*

$$x^* = \begin{cases} w & \text{if } \hat{A} \in (\frac{1}{n},1) \\ [0,w] & \text{if } \hat{A} = \frac{1}{n} \\ 0 & \text{if } \hat{A} \in (0,\frac{1}{n}) \end{cases}$$

Additionally, $N(x)$ is strictly increasing in x when $\hat{A} \in (\frac{1}{n},1)$, constant in x when $\hat{A} = \frac{1}{n}$, and strictly decreasing in x when $\hat{A} \in (0,\frac{1}{n})$.

Proposition 5 predicts that individuals consider contributions to the public account socially appropriate only when they are socially efficient. Specifically, it implies a positive relationship between contributions and social appropriateness when $\hat{A}n > 1$ and a negative relationship when $\hat{A}n < 1$.

Experimental details: Participants evaluated a situation in which Person A is matched with three other individuals. Each individual receives €10 and allocates it between the private and public accounts. Each €1 deposited in the private account returns €1 to that individual. On the other hand, each €1 deposited in the public account gives a return

of either €0.3 (*Efficient PGG*) or €0.2 (*Inefficient PGG*) to each group member. Thus, contributing to the public account is socially efficient in *Efficient PGG* and socially inefficient in *Inefficient PGG*. Participants evaluate how socially appropriate they find Person A contributing €$x \in \{0,2,4,6,8,10\}$ to the public account and €$(10-x)$ to the private account.

Predictions (displayed in Figure 10a): (I) Contributing a low amount to the public account is more socially appropriate in *Inefficient PGG* than in *Efficient PGG*. (II) Contributing a high amount to the public account is more socially appropriate in *Efficient PGG* than in *Inefficient PGG*. (III) In *Efficient PGG*, there is a positive relationship between contributions to the public account and social appropriateness. (IV) In *Inefficient PGG*, there is a negative relationship between contributions to the public account and social appropriateness.

4.3.2 The volunteer's dilemma.

Individuals are assigned to a group of size $n \geq 2$ and decide whether to volunteer. If at least one individual volunteers, each individual receives a benefit of $b > 0$, while volunteers suffer a cost of $c \in (0,b)$, regardless of the number of volunteers. When no individuals volunteer, all individuals receive zero. Thus, individuals prefer volunteering over no one volunteering but would rather have someone else volunteer. Let $x_i \in [0,1]$ denote the probability that individual i volunteers. Individual i's expected material payoff under action profile $(x_1,...,x_n)$ is

$$\pi(x_1,...,x_n) = v(b(1-\prod_{j=1}^{n}(1-x_j)) - cx_i). \tag{11}$$

The volunteer's dilemma has been used to study the bystander effect (Diekmann, 1985; Campos-Mercade, 2021), the observation that one's likelihood of helping others decreases when other individuals who can volunteer are present (Latané and Nida, 1981; Fischer et al., 2011; Kettrey and Marx, 2021). Here, the injunctive norm captures a trade-off between increasing the probability that someone volunteers and decreasing the probability of multiple individuals volunteering simultaneously, which is socially inefficient. The injunctive norm is

$$N(x) = v(b(1-(1-x)^n) - cx). \tag{12}$$

Proposition 6. *The most socially appropriate volunteering probability is* $x^* = 1 - (\frac{c}{bn})^{\frac{1}{n-1}} \in (0,1)$. *Additionally, (i)* $\frac{\partial x^*}{\partial n} \leq 0$ *and (ii)* $\frac{\partial N(x)}{\partial x} > 0$ *for any* $x \in [0,x^*)$ *and* $\frac{\partial N(x)}{\partial x} < 0$ *for any* $x \in (x^*,1]$.

Proposition 6 predicts that the most socially appropriate probability of volunteering decreases with group size. This result aligns with the *diffusion of responsibility principle*, which suggests that individuals distribute their responsibility to help among the number of people present (Latane and Darley, 1968).

Experimental details: Participants evaluated a situation in which Person A is in a group with either $N = 3$ (*VD 3*) or $N = 16$ (*VD 16*). Individuals simultaneously decided whether to volunteer. If no individual volunteers, all individuals earn €0. If at least one individual volunteers, all individuals earn €10 at a cost of €5 for the volunteers. Participants evaluate how socially appropriate they find Person A volunteering with probability $x \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

Predictions (displayed in Figure 11a): (I) Volunteering with low probability is more socially appropriate in *VD 16* than in *VD 3*. (II) Volunteering with high probability is more socially appropriate in *VD 3* than in *VD 16*. (III) In both variants, volunteering with certainty is more socially appropriate than not volunteering. (IV) The most socially appropriate probability of volunteering is lower in *VD 16* than in *VD 3*.

5 Results

In this section, I test the predictions of the theory using data from the lab experiment (Section 5.1) and existing evidence (Section 5.2). I discuss these results in Section 5.3. In Appendix C, I present two additional results from the experimental questionnaire. First, a universalization statement was the most relevant for participants to justify their evaluations (Appendix C.2). Second, participants' degrees of norm-following and universalization reasoning are positively correlated (Appendix C.3).

5.1 Experimental Results

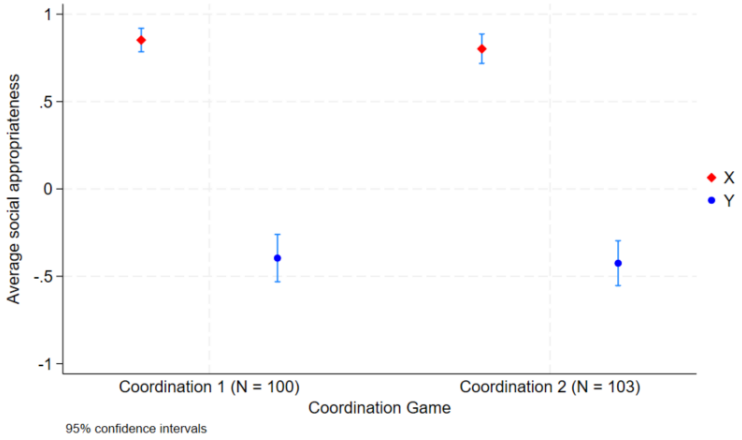
5.1.1 Coordination game with two Pareto-ranked Nash equilibria

Figure 5 supports the predictions derived in Section 4.1.1. First, in both variants, selecting X is more socially appropriate than Y ($p < 0.0001$ in both cases).[24] Therefore, although

[24] Within-participant tests are conducted using paired comparison t-tests, while between-participant tests are conducted using two-sample t-tests. I perform one-sided (t-)tests when the theory has a prediction on the direction of the effect and two-sided (t-)tests otherwise.

both actions may implement a Nash equilibrium, selecting the action that may implement the Pareto-dominant Nash equilibrium is considered more socially appropriate. Second, selecting X is more socially appropriate in *Coordination 1* than in *Coordination 2*, although this difference is not statistically significant ($p = 0.17$).[25] Finally, selecting Y is equally appropriate in both variants ($p = 0.75$).

Figure 5: Average social appropriateness ratings in the coordination game.



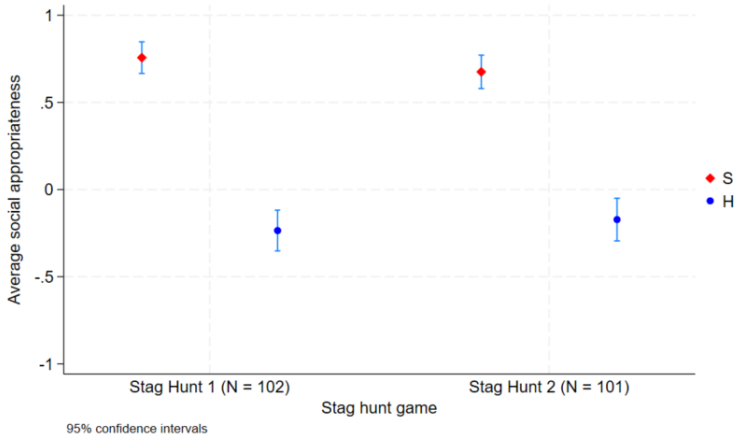
Note: The left column shows the average social appropriateness of X (red diamond) and Y (blue round) in Coordination 1. The right column shows the average social appropriateness of X (red diamond) and Y (blue round) in Coordination 2

5.1.2 Stag Hunt game

Figure 6 supports the predictions derived in Section 4.1.2. First, in both variants, selecting S is more socially appropriate than selecting H ($p < 0.0001$ for both tests). Second, selecting S is more socially appropriate in *Stag Hunt 1* than in *Stag Hunt 2*, although this difference is not statistically significant ($p = 0.11$). Finally, selecting H is equally appropriate in both variants ($p = 0.46$).

[25] In two-action games, differences in social appropriateness for the same action tend to be small. Therefore, caution is needed when interpreting them.

Figure 6: Average social appropriateness ratings in the stag hunt game.

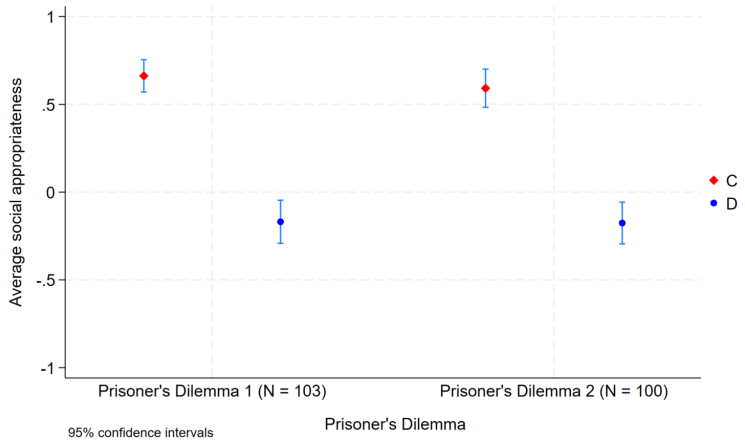


Note: The left column shows the average social appropriateness of S (red diamond) and H (blue round) in Stag Hunt 1. The right column shows the average social appropriateness of S (red diamond) and H (blue round) in Stag Hunt 2.

5.1.3 Prisoners' Dilemma

Figure 7 supports the predictions derived in Section 4.1.3. First, in both variants, selecting C is more socially appropriate than selecting D ($p < 0.0001$ for both tests). Therefore, selecting D is perceived as socially inappropriate despite being strictly dominant. Second, selecting C is equally appropriate in both variants ($p = 0.32$). Finally, selecting D is equally appropriate in both variants ($p = 0.93$).

Figure 7: Average social appropriateness ratings in the Prisoner's dilemma.



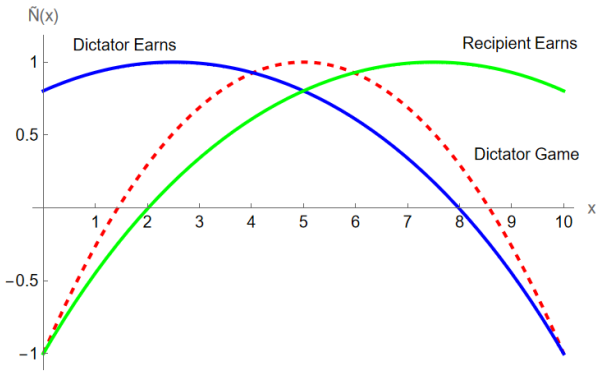
Note: The left column shows the average social appropriateness of C (red diamond) and D (blue round) in Prisoner's Dilemma 1. The right column shows the average social appropriateness of C (red diamond) and D (blue round) in Prisoner's Dilemma 2.

5.1.4 Dictator game with earned endowment

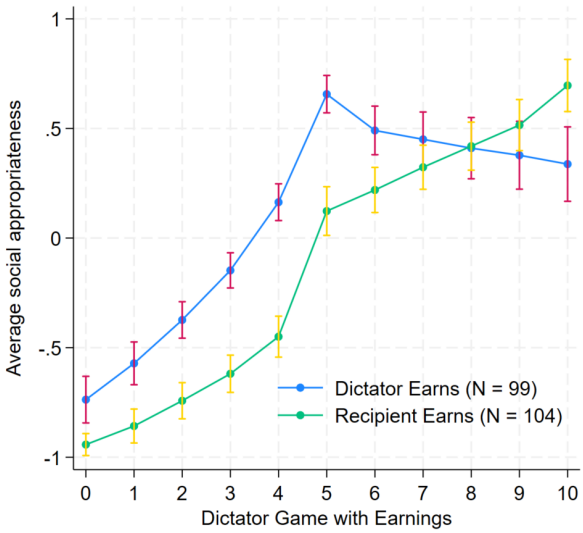
Figures 8a and 8b display the theory's predictions and the elicited norms, respectively.[26] The elicited norms provide mixed support for the theory's predictions. Compared to the standard dictator game, when the recipient is the one exerting effort, there is a clear change in the injunctive norm: larger transfers are perceived as increasingly socially appropriate. On the other hand, when the dictator is the one exerting effort, the norm remains similar to that of the standard dictator game. Thus, despite identical effort levels in both variants, the change in evaluations is asymmetric.

Figure 8: Dictator game with earnings

(a) Normalized injunctive norms predicted



(b) Average social appropriateness



Note: x-axis: Amount given to the recipient. y-axis: Average social appropriateness.

Returning to the theory's predictions, the actions' social appropriateness varies substantially across the two variants. First, giving $x = 0$, $x = 1$ or $x = 2$ is significantly more socially appropriate in *Dictator Earns* than in *Recipient Earns* ($p = 0.0003$, $p < 0.0001$, and $p < 0.0001$, respectively). Second, giving $x = 9$ or $x = 10$ is significantly more socially appropriate in *Recipient Earns* than in *Dictator Earns* ($p = 0.0788$ and $p = 0.0003$, respectively), whereas this is not the case for $x = 8$ ($p = 0.45$). Third, the most socially appropriate transfer is above the equal split in *Recipient Earns* ($p < 0.001$). Finally, there is no evidence that the most socially appropriate transfer in *Dictator Earns* is below the equal split.

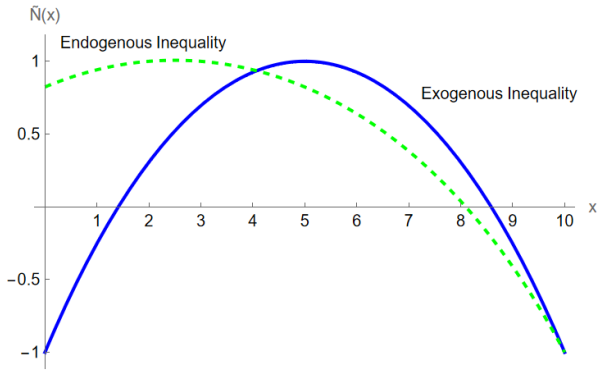
[26] Ellingsen and Mohlin (2023) elicit the *Dictator Earns* condition with similar qualitative patterns. Kassas and Palma (2019) and Bašić and Verrina (2023) examine a modified dictator game where individuals' roles are decided with a contest.

5.1.5 Dictator game with joint production

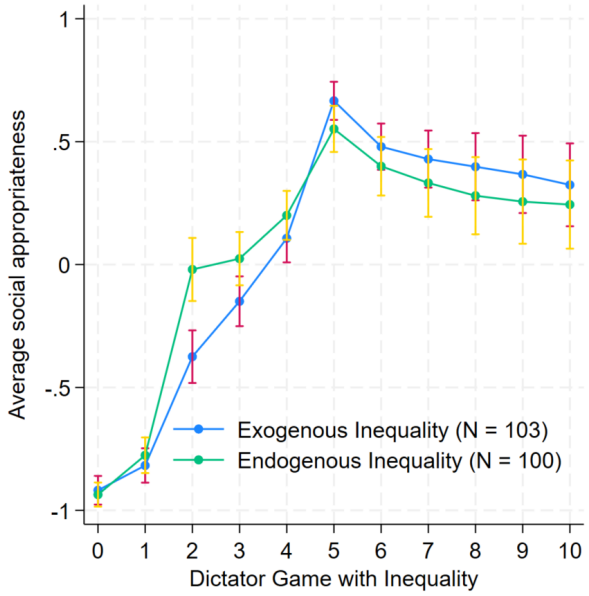
Figures 9a and 9b display the theory's predictions and the elicited norms, respectively. The elicited norms provide mixed support for the theory's predictions. Consistent with the theory, when both individuals exert the same effort and differences in production arise from exogenous reasons, the injunctive norm is similar to the one in the standard dictator game. However, contrary to the theory's predictions, the injunctive norm does not change substantially when inequality in production arises from endogenous reasons (i.e., the dictator exerts more effort than the recipient). Even in that case, the equal division remains the most socially appropriate action. Thus, there is no evidence suggesting that the injunctive norm varies based on the source of inequality.

Figure 9: Dictator game with joint production

(a) Normalized injunctive norms predicted



(b) Average social appropriateness



Note: x-axis: Amount given to the recipient. y-axis: Average social appropriateness.

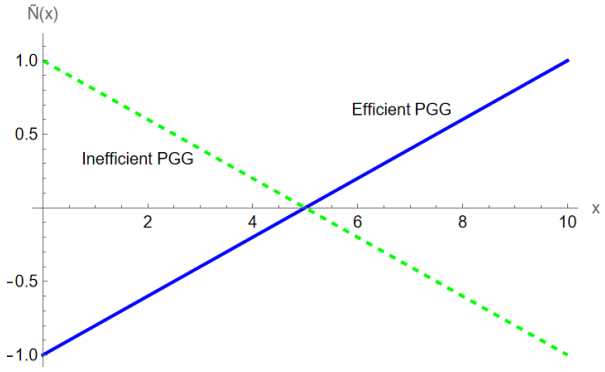
Although the elicited norms do not exhibit stark differences, there are still some differences in the actions' social appropriateness. First, giving $x = 2$ is significantly more socially appropriate in *Endogenous Inequality* than in *Exogenous Inequality* ($p < 0.001$), whereas this is not the case for for $x = 0$ ($p = 0.67$) and $x = 1$ ($p = 0.20$). Second, giving $x \geq 5$ is more socially appropriate in *Exogenous Inequality* than in *Endogenous Inequality*. However, this difference is statistically significant only for $x = 5$ ($p = 0.03$).

5.1.6 Linear public goods game

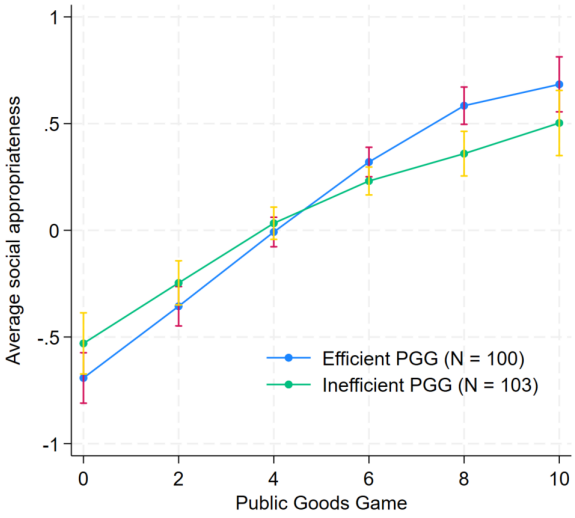
Figures 10a and 10b display the theory's predictions and the elicited norms in the linear public goods game, respectively.[27] The key theoretical prediction is inconsistent with the empirical evidence. Specifically, the theory predicts a negative relationship between contributions to the public account and social appropriateness in *Inefficient PGG*. However, I document a robust positive relationship in both *Efficient PGG* and *Inefficient PGG* (see Appendix C for the regression tables).[28] This suggests that universalization reasoning alone cannot fully explain injunctive norms. In Section 6, I show how this result can be rationalized using the extended version of the theory.

Figure 10: Linear public goods game

(a) Normalized injunctive norms predicted



(b) Average social appropriateness



Note: x-axis: Amount contributed to the public account. y-axis: Average social appropriateness.

Beyond this negative result, one can observe differences between the appropriateness of the different contributions. Contributing $x = 0$ and $x = 2$ is more socially appropriate in *Inefficient PGG* than in *Efficient PGG*, while contributing $x = 8$ and $x = 10$ is more socially appropriate in *Efficient PGG* than in *Inefficient PGG*. All these differences are statistically significant ($p = 0.043$ and $p = 0.059$, $p = 0.0006$ and $p = 0.037$, respectively) and

[27] Kimbrough and Vostroknutov (2016) elicit the *Efficient PGG* condition with similar qualitative patterns. Abbink et al. (2017) elicit a modified *Inefficient PGG* condition where the interaction is repeated for twenty periods, and individuals can punish after the contribution stage. They find similar qualitative patterns as the ones presented here.

[28] These results cannot be attributed to participants' misunderstanding of the task, as they remain robust when considering only those who correctly answered both comprehension questions (see Appendix C). Importantly, one of the comprehension questions is specifically related to the public good multiplier.

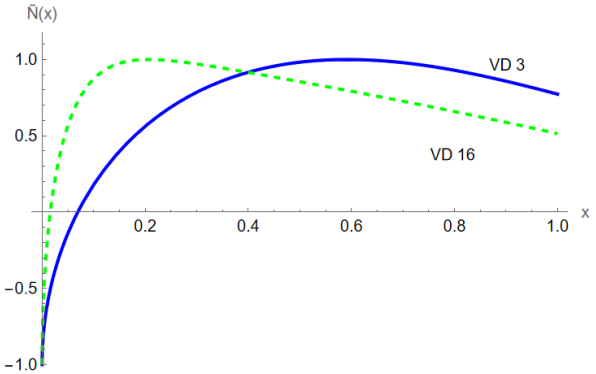
in line with the theory's predictions.

5.1.7 Volunteer's Dilemma

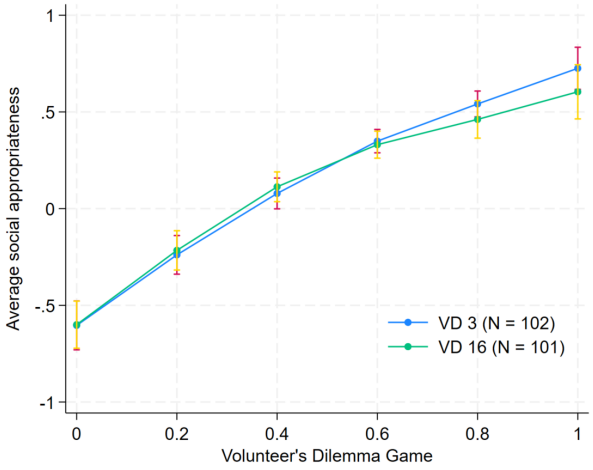
Figures 11a and 11b display the theory's predictions and the elicited norms in the volunteer's dilemma, respectively. The elicited norms provide mixed support for the theory's predictions. Consistent with the theory, the social appropriateness of volunteering varies with group size. More concretely, selecting $x = 0$ and $x = 0.2$ is more socially appropriate in *VD 16*, while selecting $x = 0.8$ and $x = 1$ is more socially appropriate in *VD 3*. These differences are small and only statistically significant for $x = 0.8$ and $x = 1$ ($p = 0.48$ and $p = 0.37$, $p = 0.08$, and $p = 0.08$, respectively). These differences provide suggestive evidence of the diffusion of responsibility effect predicted by the theory. At the same time, they contradict the predictions of a utilitarian norm, which assumes individuals prioritize helping the maximum number of others.

Figure 11: Volunteer's dilemma

(a) Normalized injunctive norms predicted



(b) Average social appropriateness



Note: x-axis: Probability of volunteering. y-axis: Average social appropriateness.

On the other hand, contrary to the theory's predictions, I do not find evidence that the most socially appropriate probability of volunteering is lower in *VD 16* than in *VD 3*, as it equals one in both cases. An exploratory analysis shows that a higher fraction of participants evaluated as *Very Socially Appropriate* to volunteer with a probability strictly below one in *VG 16* than in *VG 3* ($p = 0.068$ and $p = 0.116$ with one- and two-sided Fisher exact tests).[29] This provides suggestive evidence in support of this prediction.

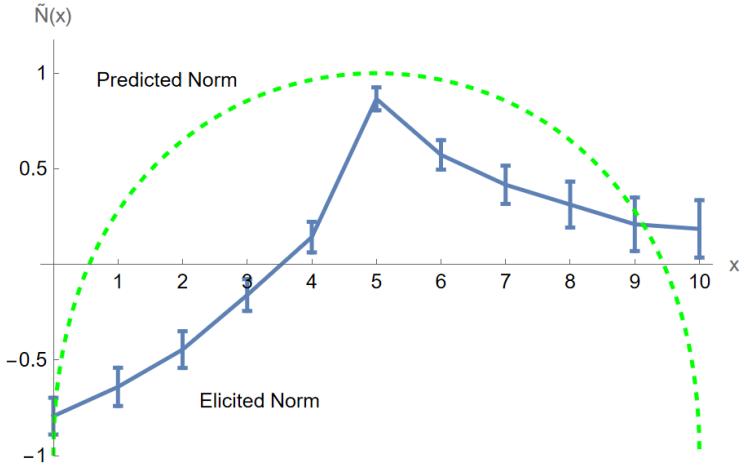
[29]Specifically, 45% of the participants (46 out of 101) evaluated in *VG 16* some action $x < 1$ as *Very Socially Appropriate*, whereas 34% (35 out of 102) did so in *VG 3*.

5.2 Past Evidence

5.2.1 Standard dictator game

Figure 12 displays both the injunctive norm predicted by the theory and the one elicited in Krupka and Weber (2013).[30] The norm elicited in Krupka and Weber (2013) provides mixed support for the theory's predictions. On the one hand, the theory correctly predicts that the most socially appropriate transfer is $x = \frac{w}{2}$, that $N(x)$ is increasing when $x < \frac{w}{2}$, and that it is decreasing when $x > \frac{w}{2}$. On the other hand, the elicited injunctive norm exhibits a clear asymmetry: transfers above the equal split are perceived as more socially appropriate than their complementary transfers below it. In Section 6, I show how this asymmetry can be explained using the extended version of the theory.

Figure 12: Normalized injunctive norm predicted (in dashed green) and elicited in Krupka and Weber (2013) (in blue) in the $10 dictator game.



Note: x-axis: Amount given to the recipient. y-axis: Average social appropriateness.

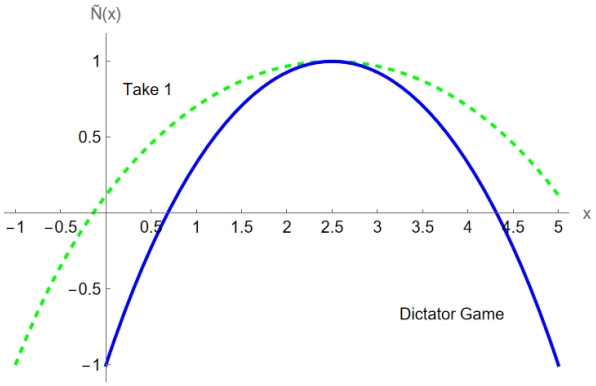
5.2.2 Dictator game with taking option

Figure 13b supports the predictions derived in Section 4.2.2. Specifically, taking $1 is the most socially inappropriate transfer in Take-1. Transferring $\frac{w}{2}$ is the most socially appropriate transfer in both variants. Finally, transfers common in both variants are generally more socially appropriate in Take-1.

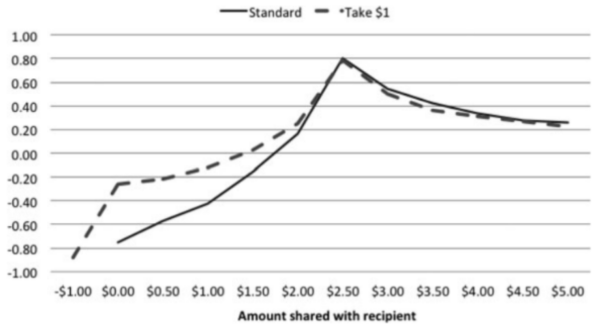
[30]The injunctive norm in the dictator game has also been elicited in other studies (e.g., Kimbrough and Vostroknutov, 2016; Bašić and Verrina, 2023) with similar qualitative patterns.

Figure 13: Dictator game and Take-1 condition

(a) Normalized injunctive norms predicted



(b) Average social appropriateness (Krupka and Weber, 2013)



Note: x-axis: Amount given to the recipient. y-axis: Average social appropriateness.

Note that although $\widetilde{N}(x_i)$ can explain the shift in norms documented in Krupka and Weber (2013), the utility function (2) cannot explain the choice set effects reported in List (2007). This reversal documented in List (2007) (i.e., individuals choosing a positive transfer in the dictator game but zero in Take 1) violates the Weak Axiom of Revealed Preference (WARP). It is well-known that this violation cannot be accommodated by complete and transitive preferences, such as *homo moralis*.[31] In Appendix D, I show that a minor modification of (2) can explain this reversal while preserving the same injunctive norm.

5.3 Discussion of the results

I summarize the main results described above and discuss two potential explanations on why the theory does not explain several key predictions of the presented evidence.

- **Symmetric one-shot two-player games:** The evidence is generally in line with the theory's predictions. More concretely, it correctly predicts which action is more socially appropriate in all three situations. Additionally, despite the small differences between variants, it correctly predicts (at least directionally) the relative appropriateness of the actions in the different variants.

- **Dictator games:** The evidence provides mixed support to the theory's predictions. In some of the variants, the theory provides a good match for the elicited norms. For example, it can explain the change in norms when (i) an additional action is included, (ii) the recipient earns the endowment, and (iii) differences in production

[31] See, for example, Propositions 1.D.1 and 1.D.2 in Mas-Colell et al. (1995).

are due to exogenous reasons. However, it fails to explain (i) the asymmetry in the standard dictator game and (ii) the equal split being the most socially appropriate transfer when the dictator exerts more effort.

- **Public good games:** Despite the theory is in line with several observations (e.g., it is more socially appropriate to volunteer when fewer individuals are present), it *fails* to account for several important ones. More concretely, it does not predict (i) the positive relationship between contributions and social appropriateness in *Inefficient PGG*, and (ii) that volunteering with certainty is the most socially appropriate action in the volunteer's dilemma.

Two limitations: I consider two possible reasons for the differences between predicted and elicited norms. I exemplify them using the standard dictator game.

- **Intentions and Context**: The social appropriateness of an action may differ depending on whether the decision-maker is directly involved in the interaction. For example, when the dictator gives €8 (out of €10) to the recipient, it may be considered socially appropriate if the dictator is Person A (as he sacrifices his own money to give it to Person B). Conversely, it may be seen as socially inappropriate when the dictator is a third party (as it generates inequality between Person A and Person B without a specific reason).

- **Conflicting norms**: While research of injunctive norms tends to assume a single shared norm, recent studies document that participants differ on their perceptions of what is socially appropriate. In the dictator game, two main conflicting norms have been identified (Panizza et al., 2023; Kimbrough et al., 2024): (i) *equality norm* (as predicted by the theory; see Figure 12), and (ii) *generosity norm*, in which the higher the transfer the more socially appropriate that transfer is. Comparing the predictions of the theory with the average injunctive norm may mask substantial heterogeneity in participants' responses.

6 The extended theoretical framework

In Section 6.1, I present the extended injunctive norm. In the following sections, I examine its predictions in the standard dictator game (Section 6.2) and the linear public goods game (Section 6.3). For the similarity between the linear public goods game and the volunteer's dilemma, I report the latter in Appendix A.

6.1 The extended injunctive norm

The injunctive norm proposed in Section 2 does not explain the following three observations:

- **Standard dictator game**: Any transfer $x \in (\frac{w}{2}, w]$ is perceived as more socially appropriate than the complementary transfer $w - x$ (see Figure 12).
- **Linear public goods game**: There is a positive relationship between contributions to the public account and their perceived social appropriateness, even when contributions are socially inefficient (see Figure 10b).
- **Volunteer's dilemma**: Volunteering with certainty is viewed as the most socially appropriate action (see Figure 11b).

To account for these observations, I extend the preferences in (1) by assuming that individuals' utility function combines universalization reasoning and social concerns (i.e., attaching a weight to others' payoff). For simplicity, I only consider strategy profiles where all other individuals select the same strategy $\widetilde{x}$. In that case, individual i's utility function is

$$\begin{aligned} u_i(x_i, \widetilde{x}) &= (1-\kappa_i)\pi_i(x_i, \widetilde{x}) \\ &- \alpha_i(n-1)\max[\pi_j(x_i, \widetilde{x}) - \pi_i(x_i, \widetilde{x}), 0] \\ &- \beta_i(n-1)\max[\pi_i(x_i, \widetilde{x}) - \pi_j(x_i, \widetilde{x}), 0] \\ &+ \kappa_i \pi_i(\boldsymbol{x_i}). \end{aligned} \tag{13}$$

Here, $\pi_j(x_i, \widetilde{x})$ represents others' material payoff under strategy profile $(x_i, \widetilde{x})$. Equation (13) is motivated by two reasons. First, it nests several prominent preferences as special cases: payoff-maximization (i.e., $\alpha_i = \beta_i = \kappa_i = 0$), altruism (i.e., $\kappa_i = 0$ and $\alpha_i = -\beta_i$), inequity aversion (i.e., $\kappa_i = 0$ and $\alpha_i \geq \beta_i > 0$), spitefulness (i.e., $\kappa_i = 0$ and $\alpha_i = -\beta_i$ for a $\beta_i \in (-1, 0)$) and homo moralis (i.e., $\alpha_i = \beta_i = 0$ and $\kappa_i \in (0, 1]$). Second, van Leeuwen and Alger (2024) structurally estimate (13) using experimental data and show that a model combining universalization reasoning and social preferences improves substantially the fit compared to models based on either component alone.[32]

[32] Model comparisons are based on the Integrated Completed Likelihood criterion (Biernacki et al., 2002), which penalizes models for additional parameters.

Lemma 2 shows that under the assumption that individuals attach a positive weight to others' payoff, equation (13) can be rewritten as a model of norm conformity, in which the injunctive norm is a combination of universalization reasoning and altruism.

Lemma 2. *Let* $\alpha_i = -\beta_i < 0$, $\widetilde{\beta}_i \equiv \beta_i(n-1)$, $\hat{\pi}(x_i, \widetilde{x}) \equiv \widetilde{\beta}_i \pi_j(x_i, \widetilde{x}) + \kappa_i \pi_i(x_i, ..., x_i)$, $\overline{x_i} \in \arg\max_{x \in X} \hat{\pi}(x, \widetilde{x})$, *and* $\underline{x_i} \in \arg\min_{x \in X} \hat{\pi}(x, \widetilde{x})$. *Then, (13) represents the same preferences as*

$$\widetilde{u}_i(x_i, \widetilde{x}) = \widetilde{v}(\pi_i(x_i, \widetilde{x})) + \widetilde{\gamma}_i \widetilde{N}_i(x_i, \widetilde{x}), \tag{14}$$

with

- $\widetilde{v}(\pi_i(x_i, \widetilde{x})) \equiv \frac{2}{\hat{\pi}(\overline{x_i}, \widetilde{x}) - \hat{\pi}(\underline{x_i}, \widetilde{x})} \times \pi_i(x_i, \widetilde{x})$,
- $\widetilde{\gamma}_i \equiv \frac{1}{1 - \widetilde{\beta}_i - \kappa_i}$,
- $N_i(x_i, \widetilde{x}) \equiv \widetilde{\beta}_i \pi_j(x_i, \widetilde{x}) + \kappa_i \pi_i(\boldsymbol{x_i})$,
- $\widetilde{N}_i(x_i, \widetilde{x}) \equiv 2 \frac{N_i(x_i, \widetilde{x}) - \hat{\pi}(\underline{x_i}, \widetilde{x})}{\hat{\pi}(\overline{x_i}, \widetilde{x}) - \hat{\pi}(\underline{x_i}, \widetilde{x})} - 1 \in [-1, 1]$.

Proof. See Appendix B. □

For ease of exposition, I normalize $\widetilde{\beta}_i + \kappa_i = 1$ which allows to write $N_i(x_i, \widetilde{x})$ as

$$N_i(x_i, \widetilde{x}) = (1 - \tau_i) N(x_i) + \tau_i \pi_j(x_i, \widetilde{x}). \tag{15}$$

Here, $\tau_i \in [0, 1]$ is the weight that individual i attaches to others' material payoff in evaluating the social appropriateness of selecting x_i. When $\tau_i = 0$, $N_i(x_i, \widetilde{x})$ coincides with $N(x_i)$, while it coincides with others' payoffs when $\tau_i = 1$.[33] Therefore, the social appropriateness of a strategy is now the weighted sum of the universalization norm and others' payoff. I refer to this latter concern as a *kindness* type of motive.[34]

In Section 2, I assume that injunctive norms are homogeneous across individuals. However, several studies have documented heterogeneity in individuals' beliefs on the injunctive norm (Bursztyn et al., 2020; Andre et al., 2022; Panizza et al., 2023; Kimbrough

[33] When $\tau_i > 0$, the social appropriateness of strategies may depend on one's beliefs about others' behavior. The idea is that the impact on others' can vary depending on what others are doing. This effect may be relevant in some games (e.g., games with Pareto-ranked Nash equilibria) but not in others (e.g., linear public goods games).

[34] Kindness is defined in the Cambridge Academic Content Dictionary as "*the quality of being generous, helpful, and caring about other people, or an act showing this quality*". For experimental papers studying kindness see Andreoni (1995), Di Mauro and Finocchiaro Castro (2011) and Koch and Nafziger (2016).

et al., 2024). Under (15), norm heterogeneity may emerge due to different beliefs about (i) others' choices (i.e., x_j), or (ii) the weight others attach to universalization reasoning and social concerns (i.e., τ_i). In the following sections, I compute $N_i(x, \widetilde{x})$ in several interactions and study how it varies with τ.

6.2 Standard dictator game

In the standard dictator game, the extended norm is given by

$$N_i(x,\widetilde{x}) = (1-\tau)\underbrace{\frac{1}{2}[v(w-x)+v(x)]}_{\text{Universalization norm}} + \tau\underbrace{\frac{1}{2}[v(x)+v(w-\widetilde{x})]}_{\text{Kindness norm}}, \tag{16}$$

where $\widetilde{x} \in [0,w]$ is transfer of the other individual. The first-order condition of (16) illustrates how the extended norm can accommodate for the role of intentions and context. More concretely,

$$\frac{\partial N_i(x,\widetilde{x})}{\partial x} = -\frac{1}{2}(1-\tau)v'(w-x) + \frac{1}{2}v'(x). \tag{17}$$

The most socially appropriate transfer x^* is computed from a weighted sum of the two individuals' marginal utilities. These weights may differ across individuals and situations. For example, when the recipient is a charity or when the decision-maker is in the dictator role, one might expect $\tau > 0$ which over-weights the recipient's utility. In this case, transfers above the equal split are more socially appropriate than those below it. On the other hand, when third parties have to divide money between the two players, one might expect $\tau = 0$ equally weighting the two individuals' payoffs. In this case, the norm is symmetric around the equal split.[35]

Proposition 7 characterizes the main properties of the extended injunctive norm in the standard dictator game.

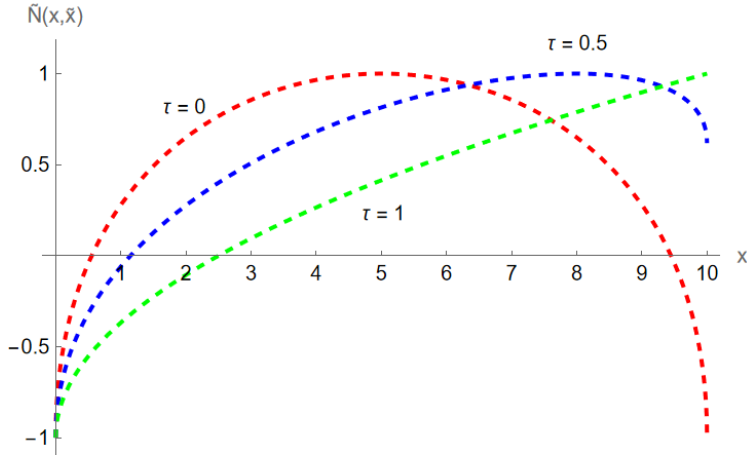
Proposition 7. *Let $\hat{x}$ be such that $\frac{\partial N_i(x,\widetilde{x})}{\partial x}\Big|_{x=\hat{x}} = 0$. Then, there exists $\overline{\tau} \equiv 1 - \frac{v'(w)}{v'(0)} \in (0,1)$ such that the most socially appropriate transfer in the dictator game is:*

$$x^* = \begin{cases} \hat{x} \in [\frac{w}{2}, w) & \text{if } \tau < \overline{\tau} \\ w & \text{if } \tau \geq \overline{\tau} \end{cases} \tag{18}$$

[35] Bašić and Verrina (2023) elicit the injunctive norm in a dictator game with a charity recipient and find that it is increasing in the dictator's transfer. This is consistent with a case of high τ. On the other hand, third parties mostly divide the endowment equally between two players who exert the same effort. This is consistent with the case where $\tau = 0$. See Figure 14 for the predicted injunctive norm in these two scenarios.

Additionally, (i) $\frac{\partial \hat{x}}{\partial \tau} \geq 0$*, (ii) when* $\tau > 0$ *and* $x \in [0, \frac{w}{2})$*,* $N_i(w-x, \widetilde{x}) > N_i(x, \widetilde{x})$*, (iii) when* $x^* \in [\frac{w}{2}, w)$*,* $\frac{\partial N_i(x,\widetilde{x})}{\partial x} > 0$ *for* $x \in [0, x^*)$ *and* $\frac{\partial N_i(x,\widetilde{x})}{\partial x} < 0$ *for* $x \in (x^*, w]$*, and (iv) when* $x^* = w$*,* $\frac{\partial N_i(x,\widetilde{x})}{\partial x} > 0$ *for* $x \in [0, w)$*.*

Figure 14: Normalized extended injunctive norm in the €10 dictator game when $\tau = 0$ (dashed red), $\tau = 0.5$ (dashed blue) or $\tau = 1$ (dashed green).



Note: x-axis: Amount given to the recipient. y-axis: Average social appropriateness.

Figure 14 displays the extended injunctive norm for different values of τ. When $x \in [0, \frac{w}{2})$, both universalization and kindness motives are aligned. Therefore, for any $\tau \in [0,1]$, the injunctive norm is increasing for transfers below $\frac{w}{2}$. On the other hand, when $x \in (\frac{w}{2}, w]$, the two motives are in conflict, implying that the relationship between transfers and social appropriateness depends on τ. This rationalizes why transfers above the equal split are perceived as more socially appropriate than the complementary ones.

Regarding the most socially appropriate transfer, Corollary 3 shows that if a population is divided into two types (one with $\tau = 0$ and one with $\tau > 0$), then the equal split is (on average) the most socially appropriate transfer if there is a sufficiently large fraction of the population with $\tau = 0$.

Corollary 3. *Let* $s_0 \in [0,1]$ *denote the share of the population with* $\tau = 0$*, and let* $s_\tau = 1 - s_0$ *denote the share of the population with* $\tau \in (0,1]$*. Then, there exists a unique* $\underline{s} \in (0,1]$ *such that for any* $s_0 \geq \underline{s}$ *and* $\tau \in (0,1]$*,* $x = \frac{w}{2}$ *is (on average) the most socially appropriate transfer.*

This seems in line with the evidence from *Exogenous Inequality*, which has the same predictions as in the standard dictator game (see Figure 9). By (exploratory) classifying participants by their norm elicited in that variant, I find that of the 103 participants, 35

can be classified as $\tau = 0$, while 37 can be classified with any $\tau > 0$. The remaining participants either display (i) a mix of $\tau = 0$ and $\tau > 0$ (14 participants) or (ii) evaluations that can not explained by the theory (17 participants).[36] Therefore, the extended version of the theory can jointly explain the asymmetry observed at the equal split and the equal split being the most socially appropriate transfer.

6.3 Linear public goods game

In the linear public goods games, the extended norm is given by

$$N_i(x,\widetilde{x}) = (1-\tau)\underbrace{v(w-x+\hat{A}nx)}_{\text{Universalization norm}} + \tau\underbrace{v(w-\widetilde{x}+(n-1)\hat{A}\widetilde{x}+\hat{A}x)}_{\text{Kindness norm}}, \tag{19}$$

where $\widetilde{x} \in [0,w]$ is the contribution to the public good selected by others. Importantly, the kindness norm increases in x even when contributing to the public account is socially inefficient. This occurs as, even in that case, contributing to the public account increases others' payoffs. Proposition 8 characterizes the main properties of the extended injunctive norm in the linear public goods game.

Proposition 8. *Let $\hat{x}$ be such that $\frac{\partial N_i(x,\widetilde{x})}{\partial x}\Big|_{x=\hat{x}} = 0$. Then, there exist $\overline{\tau} \in (0,1)$ and $\underline{\tau} \in (0,\overline{\tau})$ such that the most socially appropriate contribution in the linear public goods game is:*

- ***Case 1**: $\hat{A}n \geq 1$ (Socially efficient case)*

$$x^* = w \qquad \forall \tau \in [0,1] \tag{20}$$

- ***Case 2**: $\hat{A}n < 1$ (Socially inefficient case)*

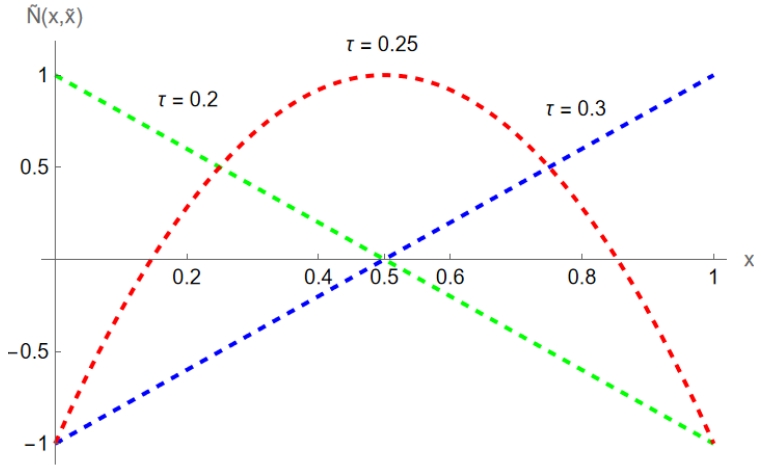
$$x^* = \begin{cases} 0 & \text{if } \tau \in [0,\underline{\tau}] \\ \hat{x} \in (0,w) & \text{if } \tau \in (\underline{\tau},\overline{\tau}) \\ w & \text{if } \tau \in [\overline{\tau},1] \end{cases} \tag{21}$$

[36] Participants are classified as $\tau = 0$ if their elicited norm is represented by the red dashed line in Figure 14. They are classified as $\tau > 0$ if they evaluate a concave relationship with the most appropriate transfer above $\frac{w}{2}$ (as in the blue and green lines in Figure 14). They are classified as having a mix of the two motives if (i) all transfers above or equal to $\frac{w}{2}$ are evaluated as equally (and most) socially appropriate, and (ii) there is an increasing relationship for transfers below $\frac{w}{2}$. See the Online Appendix for the elicited norms and the corresponding classification.

Additionally, (i) $\frac{\partial \hat{x}}{\partial \tau} \geq 0$*, (ii) when* $x^* = 0$*,* $\frac{\partial N_i(x,\widetilde{x})}{\partial x} < 0$ *for* $x \in (0,w]$*, (iii) when* $x^* = w$*,* $\frac{\partial N_i(x,\widetilde{x})}{\partial x} > 0$ *for* $x \in [0,w)$*, and (iv) when* $x^* \in (0,w)$*,* $\frac{\partial N_i(x,\widetilde{x})}{\partial x} > 0$ *for* $x \in [0,x^*)$ *and* $\frac{\partial N_i(x,\widetilde{x})}{\partial x} < 0$ *for* $x \in (x^*,w]$*.*

Proposition 8 states that there can be a positive relationship between contributions to the public account and social appropriateness even when $\hat{A}n < 1$. This occurs when individuals attach a sufficiently high weight to the kindness motive (i.e., $\tau > \overline{\tau}$).

Figure 15: Normalized extended injunctive norm when $\hat{A} = 0.2$, $n = 4$, $w = 1$, $\widetilde{x} = 0.5$, $v(\cdot) = \sqrt{(\cdot)}$, and $\tau = 0.2$ (dashed green), $\tau = 0.25$ (dashed red) or $\tau = 0.3$ (dashed blue).



Note: x-axis: Amount contributed to the public account. y-axis: Average social appropriateness.

Figure 15 displays the extended injunctive norm when $\hat{A} = 0.2$ and $n = 4$ (i.e., $\hat{A} \times n < 1$) for different values of τ. When τ is low, there is a negative relationship between contributions to the public account and social appropriateness. When τ is intermediate, there is a non-linear relationship between contributions to the public account and social appropriateness. Finally, when τ is high, there is a positive relationship between contributions to the public account and social appropriateness.

7 Comparison with alternative models

In this section, I compare the norm proposed in Section 2 with the ones introduced in Kimbrough and Vostroknutov (2023) (hereafter KV) and in López-Pérez (2008) (hereafter LP). I do not intend to conduct a horse race between the theories but rather show that they have different predictions in several settings. For each comparison, I first briefly

describe the alternative model and then discuss the main differences. For a more detailed description, I refer the reader to the corresponding papers.

7.1 KV's theory

KV's theory provides a measure of the social appropriateness of each outcome in an interaction based on individuals' dissatisfaction with those outcomes. An individual is dissatisfied with an outcome $x \in C$ if they could have achieved a higher utility with another feasible outcome. Therefore, dissatisfaction with an outcome is computed relative to *all* feasible outcomes. More concretely,

$$d_i(u_i(x), u_i(y)) = \max\{u_i(y) - u_i(x), 0\} \tag{22}$$

is individual i's dissatisfaction with consequence x because of possibility y, with $u_i(k)$ being individual i's utility with outcome k. Individual i's aggregate dissatisfaction with consequence x, given the set of consequences C, is given by

$$D_i(x|C) = \sum_{y \in C \setminus \{x\}} d_i(u_i(x), u_i(y)). \tag{23}$$

The overall dissatisfaction with an outcome x is defined as the weighted sum of all individual's dissatisfactions:

$$D(x|C) = \sum_{i \in N} w_i D_i(x|C), \tag{24}$$

where w_i is individual i's social weight. The authors propose the following injunctive norm:

Definition 1 (Kimbrough and Vostroknutov, 2023): For an environment $\langle N, C, u, D \rangle$, call $\eta : C \to [-1, 1]$, defined as

$$\eta(x|C) := [-D(x|C)],$$

where $[-D(x|C)]$ is the linear normalization of $-D$ to interval $[-1, 1]$, a **norm function** associated with $\langle N, C, u, D \rangle$. If D is a constant function, set $\eta_C(x) = 1$ for all $x \in C$.

Therefore, the social appropriateness of an outcome, $\eta(x|C)$, is inversely proportional to its aggregate dissatisfaction, $D(x|C)$, and normalized to the interval $[-1, 1]$. Fixing C, the larger $\eta(x|C)$, the more socially appropriate outcome x is, with $\eta(x|C) = 1$ (resp. $\eta(x|C) = -1$) being the most (resp. least) socially appropriate outcome.

Main differences:

- **Standard dictator game**: Both theories predict that the equal split is the most socially appropriate transfer. However, they do so for distinct reasons. In KV, the most socially appropriate transfer is the *midpoint* consequence (i.e., the consequence that has an equal number of better and worse consequences for both individuals). On the other hand, the proposed norm predicts that the most socially appropriate transfer is the one that implements the smallest difference in monetary payoffs. If we consider a modified dictator game with $w = 10$ and $x = [0,5]$ (resp. $x = [5,10]$), KV's theory predicts that the most socially appropriate transfer is $x^* = 2.5$ (resp. $x^* = 7.5$), while the proposed norm predicts $x^* = 5$ in both cases. To my best knowledge, this prediction remains untested.

- **Dictator game with taking option**: Both theories predict that low transfers are more socially appropriate when dictators can take from recipients' endowments. However, when the dictator's choice set changes, so does the most socially appropriate transfer predicted in KV. In the modified dictator game in List (2007) (i.e., $w = 5$ and \$1 of taking option), KV predicts $x^* = 2$ (see Figure 6 in KV). However, the elicited norms in Krupka and Weber (2013) show that the most socially appropriate transfer is $x^* = 2.5$ (see Figure 13). Similarly, Ellingsen and Mohlin (2023) document that the equal split is the most socially appropriate action in all the variations of taking games they consider (see Figure 4 in Ellingsen and Mohlin, 2023). This is in line with the proposed norm, but not with the one in KV.

- **Dictator game with earned endowment**: Both theories predict that the injunctive norm is sensitive to the effort exerted by individuals.[37] More concretely, it is socially appropriate to allocate a higher share of the endowment to the individual who exerted higher effort. KV's theory predicts this result by assuming that the (exogenous) social weights w_i on (24) depend on individuals' relative effort. That is, individuals that exerted a larger effort are assigned to a higher (ownership) weight. On the other hand, the proposed norm generates fairness ideals that are endogenous to the interaction (see Section 4.2.3).

- **Other situations**: The comparison in situations where outcome and strategy sets do not coincide is challenging. In this case, one could try to distinguish between the theories by examining whether the social appropriateness of different strategies

[37] KV's predictions in the dictator game with earned endowment are displayed in an older version of their manuscript (Kimbrough and Vostroknutov, 2021).

changes with the decision-makers' beliefs about others' behavior. While the proposed norm predicts that the injunctive norm should be unaffected by the decision-maker's beliefs, this should not be the case in KV. Krupka et al. (2017) show that injunctive norms are unresponsive to information about others' behavior, which is consistent with the proposed norm.

7.2 LP's theory

LP's theory prescribes which actions should or should not be taken. In this model, all actions within a category are considered equally appropriate, based on the assumption that individuals experience disutility when deviating from an internalized norm of morally permissible actions. This norm is defined by the author as follows.

Definition 1 (López-Pérez, 2008): A norm ψ is a nonempty correspondence $\psi : h \to A(h)$ that applies on any information set h of any material game. Action $a \in A(h)$ is said to be consistent with norm ψ if $a \in \psi(h)$. Otherwise, a is a deviation from ψ.

Therefore, an action a is socially appropriate if it belongs to $A(h)$, and socially inappropriate if it does not. LP's theory focuses on what actions one *ought* to choose, without ranking them in terms of social appropriateness. In other words, an action can either be appropriate (if $a \in A(h)$) or inappropriate (if $a \notin A(h)$), with all actions in each category being equally appropriate. Definition 2 shows how to compute these socially appropriate actions.

Definition 2 (López-Pérez, 2008): Allocation $x = \{x_1, ..., x_n\} \in X(t_0)$ is an (ϵ, δ)—fairmax distribution of a material game if it maximizes function

$$F_{\epsilon\delta} = \epsilon \sum_{i \in N} x_i - \delta(\max_{i \in N} x_i - \min_{i \in N} x_i), \tag{25}$$

over $X(t_0)$ for at least one node t_0. A path connecting node t_0 and one of its (ϵ, δ)-fairmax distributions is an (ϵ, δ)-fairmax path of the material game.

Here, $\epsilon > 0$ and $\delta > 0$ represent the (exogenous) weights attached to efficiency and inequality concerns, respectively. $F_{\epsilon\delta}$ depends positively on the social efficiency of x and negatively on the degree of inequality of x. One limitation of $F_{\epsilon\delta}$ is that different combinations of (ϵ, δ) may lead to different predictions, introducing several degrees of freedom. The author suggests normalizing $\epsilon = 1$ and $\delta < 1$ (i.e., efficiency is more important than inequality), which partially addresses this issue.

Main differences:

- **Standard dictator game**: In both theories, the equal split is the most socially appropriate transfer. However, in LP's theory, any transfer $x \neq \frac{w}{2}$ does not belong to the norm (i.e., $x \notin \psi(h)$) and is, therefore, equally socially inappropriate. This is consistent with a *deontological* norm, in which any deviation from $\frac{w}{2}$ is not morally permissible.[38]

- **Allocation games: Giving money to rich individuals**: When $\epsilon = 1$ and $\delta < 1$, the decision-maker is willing to increase inequality for a constant efficiency gain $k = \frac{2\delta}{1-\delta} > 0$. In other words, the decision-maker is always willing to take 1 dollar from the poor individual to give $1 + \frac{2\delta}{1-\delta}$ dollars to the rich individual, regardless of the initial inequality. This implies that, in a dictator game with multiplier $m > 0$, the decision-maker considers it socially appropriate to transfer zero if $m < \frac{1-\delta}{1+\delta}$. On the other hand, the proposed norm generally prescribes interior transfers (see Appendix A for a formal comparison).

- **Allocation games with multiple players**: As evident from (25), a prediction of LP's theory is that, conditional on having the same efficiency, the social appropriateness of allocations depends only on the inequality between the highest and lowest payoffs. Suppose we denote an allocation by $A = (x_1, x_2, x_3, x_4)$, where x_i is the payoff to player $i \in \{1,2,3,4\}$. Then, $A = (100, 0, x_3, x_4)$ is equally appropriate for any $x_3 \in [0,100]$ and $x_4 \in [0,100]$ as long as $x_3 + x_4$ remains constant. Therefore, $A_1 = (100,0,100,0)$ is as socially appropriate as $A_2 = (100,0,50,50)$. With the proposed norm, the social appropriateness of A is given by $\frac{1}{4}[v(x_1) + v(x_2) + v(x_3) + v(x_4)]$, which depends on the distribution of all payoffs.

8 Conclusion

Recent studies have demonstrated the explanatory power of injunctive norms in shaping individual behavior. However, prior research has been limited to empirical settings where injunctive norms are elicited using the method proposed in Krupka and Weber

[38] Kimbrough et al. (2024) study the individual-level variation in reported injunctive norms in the standard dictator game and classify norms depending "*of what respondents view as "appropriate" (e.g. equality vs. generosity) and how they view deviations (e.g. deontological vs. consequentialist).*" (p.1). They classify a norm as deontological when all transfers different from the equal split are evaluated similarly inappropriate. In contrast, they classify a norm as consequentialist when it assesses deviations from the equal transfer based on their magnitude from it.

(2013). In this paper, I propose a framework in which social appropriateness judgments emerge from a utility function that combines payoff maximization and universalization reasoning (Brekke et al., 2003; Alger and Weibull, 2013). A key advantage of this approach is that, rather than relying on elicited norms, it provides an explicit function that ranks actions by their social appropriateness. Additionally, it allows one to compute the social appropriateness of any action without relying on beliefs, preferences, or choices.

I test the theory's predictions using both past data and new evidence from a lab experiment. While the theory explains many patterns, it fails to account for certain key observations. For example, it does not explain why transfers above the equal split in the standard dictator game are perceived as more socially appropriate than their complementary transfers or why contributing to the public account remains socially appropriate even when it is socially inefficient. To address these gaps, I extend the injunctive norm by incorporating social concerns. I show that this extended framework rationalizes previously unexplained observations and provides a novel approach to studying norm heterogeneity.

A natural question is how to test the theory's predictions outside the lab, where norms are elicited from real-life behaviors (e.g., Lane et al., 2023). One approach is to follow, for example, Muñoz Sobrado (2022) and Alger and Laslier (2022), which use universalization reasoning in taxation and voting contexts. With their proposed settings, one could study how the social appropriateness of paying taxes and voting varies with the parameters of the interaction and compare these predictions with empirical elicitations. I leave this as an open avenue for future research.

References

Abbink, K., Gangadharan, L., Handfield, T., and Thrasher, J. (2017). Peer punishment promotes enforcement of bad social norms. *Nature Communications*, 8(1):609.

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *The Quarterly Journal of Economics*, 94(4):749–775.

Alger, I. (2023). Evolutionarily stable preferences. *Philosophical Transactions of the Royal Society B*, 378(1876):20210505.

Alger, I. and Laslier, J.-F. (2022). Homo moralis goes to the voting booth: Coordination and information aggregation. *Journal of Theoretical Politics*, 34(2):280–312.

Alger, I. and Rivero-Wildemauwe, J. I. (2024). Doing the right thing (or not) in a lemons-like situation: on the role of social preferences and kantian moral concerns. *arXiv preprint arXiv:2405.13186*.

Alger, I. and Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, 81(6):2269–2302.

Alger, I. and Weibull, J. W. (2016). Evolution and Kantian morality. *Games and Economic Behavior*, 98:56–67.

Andre, P., Boneva, T., Chopra, F., and Falk, A. (2022). Misperceived Social Norms and Willingness to Act Against Climate Change. *Econtribute Discuss. Pap*.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401):464–477.

Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion? *The American Economic Review*, pages 891–904.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11(2):122–133.

Bašić, Z. and Verrina, E. (2023). Personal norms—and not only social norms—shape economic behavior. *MPI Collective Goods Discussion Paper*, (2020/25).

Becker, G. S. (1976). Altruism, egoism, and genetic fitness: Economics and sociobiology. *Journal of Economic Literature*, 14(3):817–826.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

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

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.

Biernacki, C., Celeux, G., and Govaert, G. (2002). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

Brekke, K. A., Kverndokk, S., and Nyborg, K. (2003). An economic model of moral motivation. *Journal of Public Economics*, 87(9-10):1967–1983.

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in Saudi Arabia. *American Economic Review*, 110(10):2997–3029.

Campos-Mercade, P. (2021). The volunteer's dilemma explains the bystander effect. *Journal of Economic Behavior & Organization*, 186:646–661.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., and Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, 97(3):818–827.

Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2010). Responsibility for what? Fairness and individual responsibility. *European Economic Review*, 54(3):429–441.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Cherry, T. L., Frykblom, P., and Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, 92(4):1218–1221.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015.

Cox, J., List, J., Price, M., Sadiraj, V., and Samek, A. (2019). Moral costs and rational choice: Theory and experimental evidence. *ExCEN Working Papers*.

Di Mauro, C. and Finocchiaro Castro, M. (2011). Kindness, confusion, or… ambiguity? *Experimental Economics*, 14:611–633.

Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution*, 29(4):605–610.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and economic behavior*, 47(2):268–298.

Eichner, T. and Pethig, R. (2021). Climate policy and moral consumers. *The Scandinavian Journal of Economics*, 123(4):1190–1226.

Ellingsen, T. and Johannesson, M. (2008). Pride and prejudice: The human side of incentive theory. *American Economic Review*, 98(3):990–1008.

Ellingsen, T. and Mohlin, E. (2023). A Model of Social Duties. *Working Paper*.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4):99–117.

Erkut, H. (2020). Incentivized Measurement of Social Norms Using Coordination Games. *Analyse & Kritik*, 42(1):97–106.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Ferguson, E. and Flynn, N. (2016). Moral relativism as a disconnect between behavioural and experienced warm glow. *Journal of Economic Psychology*, 56:163–175.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., and Kainbacher, M. (2011). The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, 137(4):517.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3):347–369.

Gächter, S., Nosenzo, D., and Sefton, M. (2013). Peer effects in pro-social behavior: Social norms or social preferences? *Journal of the European Economic Association*, 11(3):548–573.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125.

Huck, S., Kübler, D., and Weibull, J. (2012). Social norms and economic incentives in firms. *Journal of Economic Behavior & Organization*, 83(2):173–185.

Isaac, R. M. and Walker, J. M. (1988). Group size effects in public goods provision: The voluntary contributions mechanism. *The Quarterly Journal of Economics*, 103(1):179–199.

Juan-Bartroli, P. and Karagözoğlu, E. (2024). Moral preferences in bargaining. *Economic Theory*, pages 1–24.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., and Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, 125(2):131.

Kant, I. (1785). Groundwork of the metaphysics of morals.

Kassas, B. and Palma, M. A. (2019). Self-serving biases in social norm compliance. *Journal of Economic Behavior & Organization*, 159:388–408.

Kessler, J. B. and Leider, S. (2012). Norms and contracting. *Management Science*, 58(1):62–77.

Kettrey, H. H. and Marx, R. A. (2021). Effects of bystander sexual assault prevention programs on promoting intervention skills and combatting the bystander effect: A systematic review and meta-analysis. *Journal of Experimental Criminology*, (17):343–367.

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.

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

Kimbrough, E. O. and Vostroknutov, A. (2021). A theory of injunctive norms. *Working Paper*.

Kimbrough, E. O. and Vostroknutov, A. (2023). A Theory of Injunctive Norms. *Working Paper*.

Köbis, N. C., Van Prooijen, J.-W., Righetti, F., and Van Lange, P. A. (2015). "who doesn't?"— The impact of descriptive norms on corruption. *PloS one*, 10(6):e0131830.

Koch, A. K. and Nafziger, J. (2016). Gift exchange, control, and cyberloafing: A real-effort experiment. *Journal of Economic Behavior & Organization*, 131:409–426.

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, 90(4):1072–1091.

Krupka, E. L., Leider, S., and Jiang, M. (2017). A meeting of the minds: informal agreements and social norms. *Management Science*, 63(6):1708–1729.

Krupka, E. L., Weber, R., Crosno, R. T., and Hoover, H. (2022). "When in Rome": Identifying social norms using coordination 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.

Kwon, J., Zhi-Xuan, T., Tenenbaum, J., and Levine, S. (2023). When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments.

Lane, T., Nosenzo, D., and Sonderegger, S. (2023). Law and norms: Empirical evidence. *American Economic Review*, 113(5):1255–1293.

Latane, B. and Darley, J. M. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology*, 10(3):215.

Latané, B. and Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, 89(2):308.

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., and Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, 117(42):26158–26169.

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2):153–174.

Lindbeck, A., Nyberg, S., and Weibull, J. W. (1999). Social norms and economic incentives in the welfare state. *The Quarterly Journal of Economics*, 114(1):1–35.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3):482–493.

López-Pérez, R. (2008). Aversion to norm-breaking: A model. *Games and Economic Behavior*, 64(1):237–267.

Luhan, W. J., Poulsen, O., and Roos, M. W. (2019). Money or morality: fairness ideals in unstructured bargaining. *Social Choice and Welfare*, 53:655–675.

Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995). *Microeconomic theory*. Oxford University Press.

Miettinen, T., Kosfeld, M., Fehr, E., and Weibull, J. (2020). Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. *Journal of Economic Behavior & Organization*, 173:1–25.

Muñoz Sobrado, E. (2022). Taxing Moral Agents. *CESifo Working Paper*.

Nosenzo, D. and Görges, L. (2020). Measuring social norms in economics: why it is important and how it is done. *Analyse & Kritik*, 42(2):285–312.

Oxoby, R. J. and Spraggon, J. (2008). Mine and yours: Property rights in dictator games. *Journal of Economic Behavior & Organization*, 65(3-4):703–713.

Panizza, F., Dimant, E., Kimbrough, E. O., and Vostroknutov, A. (2023). Measuring norm pluralism and tolerance. *Available at SSRN*.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, pages 1281–1302.

Roemer, J. E. (2010). Kantian Equilibrium. *Scandinavian Journal of Economics*, 112(1):1–24.

Rydval, O. and Ortmann, A. (2005). Loss avoidance as selection principle: evidence from simple stag-hunt games. *Economics Letters*, 88(1):101–107.

Salonia, E.-M. (2023). A Foundation for Universalization in Games. *Working Paper*.

Sarkisian, R. (2017). Team Incentives under Moral and Altruistic Preferences: Which Team to Choose? *Games*, 8(3):37.

Sugden, R. (2003). The logic of team reasoning. *Philosophical Explorations*, 6(3):165–181.

van Leeuwen, B. and Alger, I. (2024). Estimating Social Preferences and Kantian Morality in Strategic Interactions. *Journal of Political Economy Microeconomics*.

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

Wildemauwe, J. I. R. (2023). Trade among moral agents with information asymmetries. Technical report, THEMA (THéorie Economique, Modélisation et Applications).

Appendices

A Other interactions.

In this section, I consider interactions not included in the main text. More concretely, I study the trust game (Section A.1), the ultimatum game (Section A.2), the dictator game with different prices of giving (Section A.3), and the linear public goods game with heterogeneous endowments (Section A.4). Additionally, I consider the comparison with third-party choice data in Konow (2000) (Section A.5) and the volunteer's dilemma with the extended norm (Section A.6).

A.1 Trust Game

Individuals are matched into pairs and randomly assigned the trustor or trustee roles. Both trustors and trustees receive an endowment of $w > 0$. Trustors send an amount $x \in [0, w]$ to trustees, which is multiplied by a rate of return $m > 0$. Trustees return an amount $r \in [0, xm]$ to trustors. Thus, trustors' monetary payoff is $w - x + r$, while trustees' monetary payoff is $w + mx - r$ (Berg et al., 1995).

In the ex-ante symmetric version of the game, a strategy is a tuple $x = (x_1, f(x_2))$ where $x_1 \in [0, w]$ is the amount sent in the trustor role, and $f(x_2) : [0, mx_2] \to [0, mx_2]$ is the amount returned as trustee. I start by computing the injunctive norm in the trust game and show how it varies with the rate of return m and trustees' initial endowment. In the standard trust game, the injunctive norm is given by

$$N(x) = \frac{1}{2}v(w - x_1 + f(x_2)) + \frac{1}{2}v(w + mx_1 - f(x_2)). \tag{26}$$

Note that x_1 and $f(x_2)$ can be interpreted as measures of efficiency and equality. The most socially appropriate strategy $x^* = ({x_1}^*, f^*(x_2))$ consists of ${x_1}^* = w$ and $f^*(x_2) = \frac{(m+1)x_2}{2}$.[39]

When trustors' endowments (i.e., w_R) are larger than trustees' endowments (i.e., w_P),

[39] Intuitively, trustees' decision is analogous to dictators' decision with $w = (m+1)x_2$. On the other hand, given that $f^*(x_2) = \frac{(m+1)x_2}{2}$, the most socially appropriate amount sent is ${x_1}^* = w$.

the injunctive norm is given by

$$N(x) = \frac{1}{2}v(w_R - x_1 + f(x_2)) + \frac{1}{2}v(w_P + mx_1 - f(x_2)). \tag{27}$$

As before, the most socially appropriate return function $f^*(x_2)$ equals roles' ex-post material payoffs. However, this is now adjusted by the fact that trustors receive a larger endowment. The most socially appropriate strategy is $x_1^* = w_R$ and $f^*(x_2) = max[g^*(x_2), 0]$, where $g^*(x_2) \equiv \frac{w_P - w_R}{2} + \frac{(m+1)x_2}{2}$.

Finally, note that $f^*(x_2) = \frac{(m+1)x_2}{2}$ is increasing in m, as the higher m, the higher the inequality generated by x_2. On the other hand, $x_1^* = w$ if $m > 1$ and $x_1^* = 0$ if $m < 1$.

A.2 Ultimatum Game

Individuals are matched into pairs and randomly assigned the proposer or responder roles. Proposers receive an endowment of size $w > 0$ and decide an offer $x \in [0, w]$ to responders. Responders can accept or reject this offer. If responders accept it, this is implemented, while if responders reject it, both individuals receive zero (Güth et al., 1982).

In the ex-ante symmetric version of the game, a strategy is a tuple $x = (x_1, x_2)$ where $x_1 \in [0, w]$ is the amount offered in the proposer role and $x_2 \in [0, w]$ is the rejection threshold in the responder role (i.e., the responder accepts any offer equal or above x_2 and rejects otherwise).

Individual 1's material payoff when he selects $x = (x_1, x_2)$ and individual 2 selects $y = (y_1, y_2)$ is given by

$$\pi(x, y) = v(0) + \frac{1}{2} \cdot \mathbf{1}_{\{x_1 \geq y_2\}} \cdot [v(w - x_1) - v(0)] + \frac{1}{2} \cdot \mathbf{1}_{\{y_1 \geq x_2\}} \cdot [v(y_1) - v(0)]. \tag{28}$$

Thus, the social appropriateness of strategy $x = (x_1, x_2)$ is:

$$N(x) = v(0) + \frac{1}{2} \cdot \mathbf{1}_{\{x_1 \geq x_2\}} \cdot [v(x_1) + v(w - x_1) - 2v(0)]. \tag{29}$$

As $v(x_1) + v(w - x_1) - 2v(0) \geq 0$, $N(x)$ is maximized when ${x_1}^* = \frac{w}{2}$ and ${x_2}^* \in [0, \frac{w}{2}]$. Thus, offers become more socially inappropriate as the associated distribution of payoffs becomes unequal. On the other hand, it is considered socially inappropriate to reject any offer above the one you would have chosen as the proposer. In contrast, any rejection threshold below this offer is considered equally appropriate.

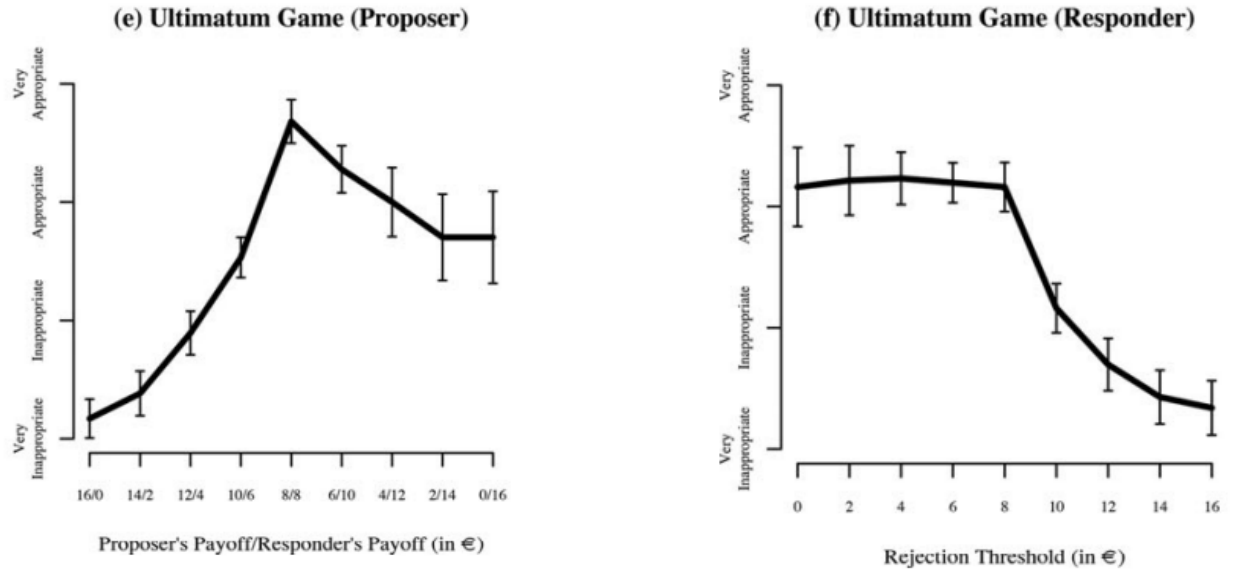


Figure 16: Injunctive norms elicited in the ultimatum game (Kimbrough and Vostroknutov, 2016).

Figure 16 shows the injunctive norm elicited in the ultimatum game with $w = 16€$ (Kimbrough and Vostroknutov, 2016). The elicited norms support the main predictions of the theory.

A.3 Dictator game with different prices of giving

A.3.1 Predictions universalization norm

In this modified dictator game, dictators' transfers are multiplied by a constant parameter $m > 0$ (Andreoni and Miller, 2002). Thus, the injunctive norm is given by

$$N(x) = \frac{1}{2}v(w - x) + \frac{1}{2}v(mx). \tag{30}$$

In this case, there is a trade-off between equality and efficiency as the allocations' sum of material payoffs are not constant. The relative weight of these two concerns depends on the relative risk aversion of v (i.e., $RRA(x) = \frac{-xv''(x)}{v'(x)}$).

Proposition 9. *Let* $RRA(x) = \frac{-xv''(x)}{v'(x)}$ *denote the relative risk aversion of function* v *at* x*. The*

most socially appropriate transfer x^* satisfies:

$$\frac{\partial x^*}{\partial m}\begin{cases} > 0 & \text{if } RRA(mx^*) < 1 \\ = 0 & \text{if } RRA(mx^*) = 1 \\ < 0 & \text{if } RRA(mx^*) > 1 \end{cases} \tag{31}$$

Additionally, $\frac{\partial N(x)}{\partial x} > 0$ *for* $x \in [0, x^*)$ *and* $\frac{\partial N(x)}{\partial x} < 0$ *for* $x \in (x^*, w]$.

To exemplify Proposition 9, let $v(c) = \frac{c^{(1-\rho)}-1}{1-\rho}$ with $\rho \neq 1$, where ρ represents the (constant) relative risk aversion of v. In this case, the injunctive norm is given by

$$N(x) = \frac{1}{2}\frac{(w-x)^{(1-\rho)}-1}{1-\rho} + \frac{1}{2}\frac{(mx)^{(1-\rho)}-1}{1-\rho}, \tag{32}$$

which leads to

$$x^* = \frac{m^{\frac{1}{\rho}}}{m + m^{\frac{1}{\rho}}} w \tag{33}$$

$$\frac{\partial x^*}{\partial m} = -\frac{\rho - 1}{\rho}\frac{w m^{\frac{1}{\rho}}}{(m + x^{\frac{1}{\rho}})^2}. \tag{34}$$

Thus, the sign of $\frac{\partial x^*}{\partial m}$ depends on the relative risk aversion of v. When $\rho > 1$, the most socially appropriate transfer decreases in m. When $\rho < 1$, the most socially appropriate transfer increases in m. Finally, when $\rho = 1$, the most socially appropriate transfer does not depend on m and is equal to $\frac{w}{2}$.

A.3.2 Predictions LP's theory

As shown in the main text, LP's fairmax distribution, which determines the most socially appropriate transfer, is given by:

$$F_{\epsilon\delta} = \epsilon \sum_{i \in N} x_i - \delta(\max_{i \in N} x_i - \min_{i \in N} x_i), \tag{35}$$

A transfer x implements a payoff distribution of $w - x$ for the dictator and xm for the recipient. By fixing $\epsilon = 1$ and $\delta < 0$, and denoting $F_{\epsilon\delta}(x)$ the value of (35) at transfer x, I compute $F_{1,\delta}(0)$ and $F_{1,\delta}(w)$:

$$F_{1,\delta}(0) = w(1 - \delta), \tag{36}$$

and

$$F_{1,\delta}(w) = wm(1-\delta). \tag{37}$$

Therefore, transferring zero is more socially appropriate than transferring w when $m < 1$, while the opposite is true for $m > 1$. On the other hand, let $\hat{x} \equiv \frac{w}{m+1}$ denote the transfer that implements equality of payoffs between the two players. In this case,

$$F_{1,\delta}(\hat{x}) = w\frac{2m}{m+1}. \tag{38}$$

One can show that there exists $\underline{m} \equiv \frac{1-\delta}{1+\delta} \in (0,1)$ and $\overline{m} \equiv \frac{1+\delta}{1-\delta} > 1$, such that (i) when $m = \underline{m}$, $F_{1,\delta}(0) = F_{1,\delta}(\hat{x})$ and (ii) $F_{1,\delta}(w) = F_{1,\delta}(\hat{x})$ when $m = \overline{m}$. Therefore, the most socially appropriate transfer is given by:

$$x^* = \begin{cases} 0 & \text{if } m \in [0, \underline{m}) \\ \hat{x} & \text{if } m \in (\underline{m}, \overline{m}) \\ w & \text{if } m > \overline{m} \end{cases} \tag{39}$$

A.4 Linear Public Goods Game with Heterogeneous Endowments

In the main text, I discussed the linear public goods game with homogeneous endowments. Here, I consider the case with heterogeneous endowments (Cherry et al., 2005; Hofmeyr et al., 2007). This is of particular interest as adding heterogeneity in endowments may lead to different norms, such as equal contributions, proportional contributions, and equal earnings (Reuben and Riedl, 2013; Kingsley, 2016).

For simplicity, I consider the case with two individuals and two endowment levels (i.e., $w_H > w_L > 0$). A strategy is a pair $x_i = (x_i^H, x_i^L)$ where $x_i^H \in [0, w_H]$ (resp. $x_i^L \in [0, w_L]$) is the contribution of the individual i when having a high (resp. low) endowment. The individual 1's (expected) material payoff is given by

$$\pi(x_1, x_2) = \frac{1}{2}\underbrace{v(w_H - x_1^H + \hat{A}(x_1^H + x_2^L))}_{\text{Individual 1 has high endowment}} + \frac{1}{2}\underbrace{v(w_L - x_1^L + \hat{A}(x_1^L + x_2^H))}_{\text{Individual 1 has low endowment}}. \tag{40}$$

Therefore, the injunctive norm is the following:

$$N(x) = \frac{1}{2}v(w_H - x_H + \hat{A}(x_H + x_L)) + \frac{1}{2}v(w_L - x_L + \hat{A}(x_L + x_H)). \tag{41}$$

When $\hat{A} \in (\frac{1}{2},1)$, the most socially appropriate strategy is $x^* = (w_H, w_L)$. Thus, the theory predicts that the equal earnings and proportional contributions norms (with individuals contributing all their endowment) will be perceived by individuals as the most socially appropriate strategy.

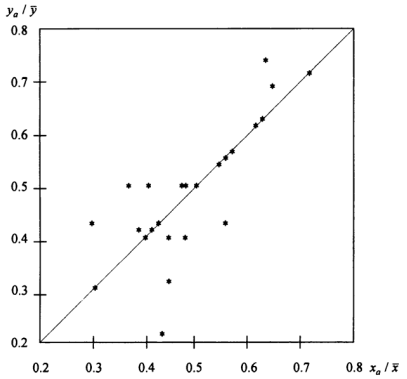
A.5 Third-party data in Konow (2000)

The experiment in Konow (2000) has two conditions. In the first condition, individuals were assigned to the same piece-rate and given a large number of letters to fold. In the second condition, individuals were assigned to different piece-rates and given a low number of letters, which ensured that all individuals folded all the letters.

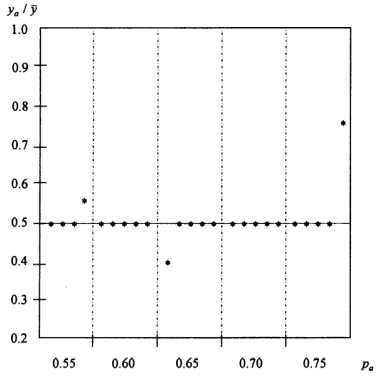
Therefore, while differences in production in the first condition are due to differences in the letters produced, they are due to differences in the piece-rates assigned in the second condition. Third parties, which are paid a fixed compensation, divide the production generated by pairs after observing the number of letters each of them produced and their assigned piece-rates. I consider the choices of third parties as a proxy of their perceived most socially appropriate division.[40] Figure 17 shows the divisions of third parties when individuals have the same (left) or different (right) piece-rates.

Figure 17: Third-party choice data in Konow (2000).

(a) Condition with the same piece-rate.



(b) Condition with different piece-rates.



Notes: x-axis left: Share of letters produced by individual A. x-axis right: Piece-rate assigned to individual A (with (i) $p_A + p_B = 1$ and (ii) $p_A > p_B$). y-axis: Share of the total production assigned to individual A.

[40] It is standard in the literature to consider the decisions of non-involved parties as proxies of their fairness and normative views (e.g., Konow, 2000; Cappelen et al., 2007; Mollerstrom et al., 2015). In the context of the theory, third parties' decisions can be interpreted as choosing the division that maximizes $N(x)$.

The data from Konow (2000) supports the theory's predictions. When individuals were assigned the same piece-rate, most third parties allocated a larger share of the production to the individual who produced the largest share of letters. On the other hand, when individuals were assigned to different piece-rates, most third parties divided the total production in half, regardless of the piece-rates assigned to the individuals.

A.6 Voluntary game with extended norm

The extended norm is given by

$$\begin{aligned} N_i(x,\widetilde{x}) &= (1-\tau)\underbrace{v(b(1-(1-x)^n)-cx)}_{\text{Universalization norm}} \\ &+ \tau\underbrace{v(b(1-(1-\widetilde{x})^{n-1})+bx(1-\widetilde{x})^{n-1}-c\widetilde{x})}_{\text{Kindness norm}}, \end{aligned} \tag{42}$$

where $\widetilde{x} \in [0,1]$ is the probability others' volunteer. As in the linear public goods game, the kindness norm is always increasing in x as volunteering always increases others' payoff, although it may be socially inefficient.

Proposition 10. *Let $\hat{x}$ be such that $\frac{\partial N_i(x,\widetilde{x})}{\partial x}\Big|_{x=\hat{x}} = 0$. Then, there exists $\overline{\tau} \in (0,1]$ such that the most socially appropriate volunteering probability in the volunteer's dilemma is:*

$$x^* = \begin{cases} \hat{x} \in [1-(\frac{c}{bn})^{\frac{1}{n-1}},1) & \text{if } \tau < \overline{\tau} \\ 1 & \text{if } \tau \geq \overline{\tau} \end{cases} \tag{43}$$

Additionally, (i) $\frac{\partial \hat{x}}{\partial \tau} \geq 0$, (ii) when $x^* \in [1-(\frac{c}{bn})^{\frac{1}{n-1}},1)$, $\frac{\partial N_i(x,\widetilde{x})}{\partial x} > 0$ for $x \in [0,x^*)$ and $\frac{\partial N_i(x,\widetilde{x})}{\partial x} < 0$ for $x \in (x^*,1]$, and (iii) when $x^* = 1$, $\frac{\partial N_i(x,\widetilde{x})}{\partial x} > 0$ for $x \in [0,1)$.

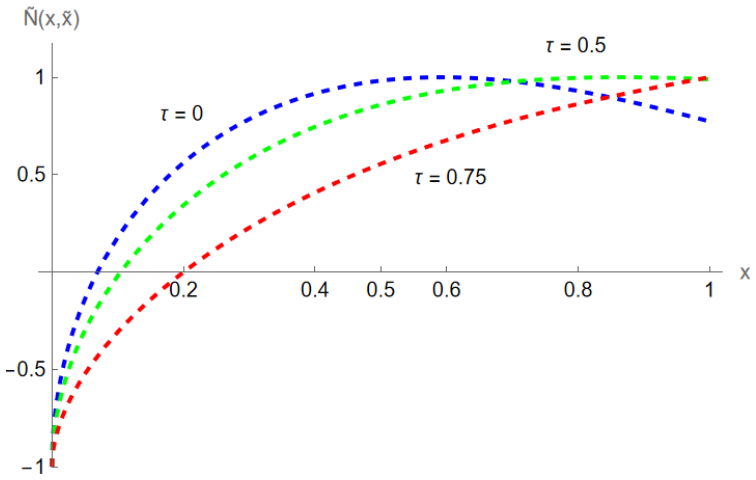


Figure 18: Normalized injunctive norms when $n = 3$, $b = 10$, $c = 5$, $\tilde{x} = 0.5$, $v(\cdot) = \sqrt{(\cdot)}$, and $\tau = 0$ (dashed blue), $\tau = 0.5$ (dashed green) or $\tau = 0.8$ (dashed red).

Figure 18 shows the normalized injunctive norms in the volunteer's dilemma for different values of τ. When τ increases, the most socially appropriate volunteering probability increases, and volunteering with high probability becomes more socially appropriate.

B Mathematical proofs.

Proof. Lemma 1: For simplicity, I show Lemma 1 when $n = 2$, but similar derivations follow when considering $n > 2$. When individuals have homo moralis preferences (Alger and Weibull, 2013) individual i's utility function is

$$u(x_i, x_j) = (1 - \kappa_i)\pi(x_i, x_j) + \kappa_i \pi(x_i, x_i). \tag{44}$$

I then define:[41]

- $\overline{x} \in \arg\max_{x \in X} \pi(x, x)$
- $\underline{x} \in \arg\min_{x \in X} \pi(x, x)$
- $a \equiv \frac{\pi(\overline{x},\overline{x}) + \pi(\underline{x},\underline{x})}{2}$
- $b \equiv \frac{\pi(\overline{x},\overline{x}) - \pi(\underline{x},\underline{x})}{2}$

[41] I consider situations where $\pi(x_i, x_i)$ is not constant in x_i, implying that $\pi(\overline{x},\overline{x}) > \pi(\underline{x},\underline{x})$. In some cases, $\overline{x}$ (or $\underline{x}$) may not be unique. This is not problematic, as any $x \in \overline{x}$ will have the same value of $\pi(\overline{x},\overline{x})$.

I proceed in two steps. First, I subtract a from (44), which gives

$$\hat{u}_i(x_i, x_j) = (1-\kappa_i)[\pi(x_i, x_j) - a] + \kappa_i[\pi(x_i, x_i) - a]. \tag{45}$$

Second, I divide (45) by $(1-\kappa_i)b$, which gives

$$\begin{aligned} \widetilde{u}_i(x_i, x_j) &= \frac{1}{b}[\pi(x_i, x_j) - a] + \frac{\kappa_i}{(1-\kappa_i)b}[\pi(x_i, x_i) - a] \\ &= 2\frac{\pi(x_i, x_j) - \pi(\underline{x}, \underline{x})}{\pi(\overline{x}, \overline{x}) - \pi(\underline{x}, \underline{x})} - 1 + \frac{\kappa_i}{(1-\kappa_i)}(2\frac{\pi(x_i, x_i) - \pi(\underline{x}, \underline{x})}{\pi(\overline{x}, \overline{x}) - \pi(\underline{x}, \underline{x})} - 1). \end{aligned} \tag{46}$$

Finally, I define

- $\widetilde{v}(\pi(x_i, x_j)) \equiv 2\frac{\pi(x_i, x_j) - \pi(\underline{x}, \underline{x})}{\pi(\overline{x}, \overline{x}) - \pi(\underline{x}, \underline{x})} - 1$,
- $\gamma_i \equiv \frac{\kappa_i}{1-\kappa_i}$,
- $N(x_i) \equiv \pi(x_i, x_i)$,
- $\widetilde{N}(x_i) \equiv 2\frac{N(x_i) - \pi(\underline{x}, \underline{x})}{\pi(\overline{x}, \overline{x}) - \pi(\underline{x}, \underline{x})} - 1$,

which leads to:

$$\widetilde{u}_i(x_i, x_j) = \widetilde{v}(\pi(x_i, x_j)) + \gamma_i \widetilde{N}(x_i). \tag{47}$$

□

Proof. Proposition 1: Follows from $\pi(X, X) = a$ and $\pi(Y, Y) = b$. □

Proof. Proposition 2: $N(x)$ is a strictly concave function as is the sum of two strictly concave functions.

$$\frac{\partial N(x)}{\partial x} = -\frac{1}{2}v'(w - x) + \frac{1}{2}v'(x), \tag{48}$$

which implies that $x^* = \frac{w}{2}$. Additionally, for the strict concavity of $N(x)$, $\frac{\partial N(x)}{\partial x} > 0$ for $x \in [0, \frac{w}{2})$ and $\frac{\partial N(x)}{\partial x} < 0$ for $x \in (\frac{w}{2}, w]$. □

Proof. Proposition 3: I distinguish between three cases: (i) $N^A(\overline{x}^A) > N^B(x') > N^A(\underline{x}^A)$, (ii) $N^B(x') > N^A(\overline{x}^A)$ and (iii) $N^B(x') < N^A(\underline{x}^A)$.

In (i), x' is neither the most nor the least socially appropriate strategy. Then, $\overline{x}^A = \overline{x}^B$ and $\underline{x}^A = \underline{x}^B$, which implies that $\widetilde{N}^A(x) = \widetilde{N}^B(x)$ for all $x \in X$. To see this, it is sufficient to note that $\widetilde{N}^g(x)$ only depends on $N(\overline{x}^g)$, $N(\underline{x}^g)$ and $N(x)$. Finally, $\widetilde{N}^B(x') \in (-1, 1)$

when $N^A(\overline{x}^A) > N^B(x') > N^A(\underline{x}^A)$ and $\widetilde{N}^B(x') = 1$ (resp. $\widetilde{N}^B(x') = -1$) when $N^B(x') = N^A(\overline{x}^A)$ (resp. $N^B(x') = N^A(\underline{x}^A)$).

In (ii), x' is the least appropriate strategy of B. Thus, $N(\overline{x}^A) = N(\overline{x}^B)$, $N(\underline{x}^A) > N(x')$, $\widetilde{N}^A(\overline{x}^A) = \widetilde{N}^B(\overline{x}^B) = 1$ and $\widetilde{N}^A(\underline{x}^A) = \widetilde{N}^B(x') = -1$. On the other hand, $\widetilde{N}^A(x) < \widetilde{N}^B(x)$ for all $x \in X$ with $\widetilde{N}^g(x) < 1$. To see this, let $N(x') = N(\underline{x}^A) - k$ with $k > 0$. Then,

$$\begin{aligned} \widetilde{N}^B(x) &= 2\frac{N(x) - N(x')}{N(\overline{x}^B) - N(x')} - 1 \\ &= 2\frac{N(x) - (N(\underline{x}^A) - k)}{N(\overline{x}^A) - (N(\underline{x}^A) - k))} - 1 \\ &= 2\frac{N(x) - N(\underline{x}^A) + k}{N(\overline{x}^A) - N(\underline{x}^A) + k} - 1 \\ &> \widetilde{N}^A(x) \end{aligned} \tag{49}$$

for any $k > 0$.

In (iii), x' is the most appropriate strategy of B. Thus, $N(\overline{x}^A) < N(x')$, $N(\underline{x}^A) = N(\underline{x}^B)$, $\widetilde{N}^A(\overline{x}^A) = \widetilde{N}^B(x') = 1$ and $\widetilde{N}^A(\underline{x}^A) = \widetilde{N}^B(\underline{x}^B) = -1$. On the other hand, $\widetilde{N}^A(x) > \widetilde{N}^B(x)$ for all $x \in X$ with $\widetilde{N}^g(x) > -1$. To see this, let $N(x') = N(\overline{x}^A) + k$ with $k > 0$. Then,

$$\begin{aligned} \widetilde{N}^B(x) &= 2\frac{N(x) - N(\underline{x}^B)}{N(x') - N(\underline{x}^B)} - 1 \\ &= 2\frac{N(x) - N(\underline{x}^A)}{N(\overline{x}^A) + k - N(\underline{x}^A)} - 1 \\ &< \widetilde{N}^A(x) \end{aligned} \tag{50}$$

for any $k > 0$. □

Proof. Proposition 4: The injunctive norm is given by

$$N(\hat{x}) = \frac{1}{2}v(w(e_1,e_2) - x(e_1,e_2) - c(e_1)) + \frac{1}{2}v(x(e_1,e_2) - c(e_2)), \tag{51}$$

where $\hat{x} = (e_1,e_2,x(e_1,e_2))$. Differentiating $N(\hat{x})$ with respect to e_1, e_2 and $x(e_1,e_2)$ gives $x^*(e_1,e_2) = \frac{w(e_1,e_2)}{2} - \frac{c(e_1)-c(e_2)}{2}$. □

Proof. Proposition 5: As $v' > 0$ and $w - x + \hat{A}nx$ is linear in x, x^* can be computed by determining for which values of $\hat{A}n$ is $w - x + \hat{A}nx$ increasing in x. Thus, $x^* = w$ when $\hat{A}n > 1$, $x^* = [0,w]$ when $\hat{A}n = 1$, and $x^* = 0$ when $\hat{A}n < 1$.

□

Proof. Proposition 6: As $N(x)$ is a strictly concave function, I compute x^* by taking the first order condition of $N(x)$ with respect to x.

$$\frac{\partial N(x)}{\partial x} = v'(b(1-(1-x)^n) - cx)[bn(1-x)^{n-1} - c], \tag{52}$$

which implies that

$$x^* = 1 - (\frac{c}{bn})^{\frac{1}{n-1}}. \tag{53}$$

To show that x^* is decreasing in n, I compute $\frac{\partial x^*}{\partial n}$ and show that it is always negative.

$$\frac{\partial x^*}{\partial n} = \frac{(\frac{b}{cn})^{\frac{1}{n-1}}[n-1+n\log(\frac{c}{bn})]}{n(n-1)^2}. \tag{54}$$

Thus, the sign of $\frac{\partial x^*}{\partial n}$ only depends on the sign of $n-1+n\log(\frac{c}{bn})$. This term can be rewritten as follows:

$$n[\log(c) - \log(b)] - n[\log(n) - 1] - 1. \tag{55}$$

Note that the first term is always negative (as $b > c$), and the second term is negative for any $n \geq 2$. Thus, $n[\log(c) - \log(b)] - n[\log(n) - 1] - 1 < 0$, which implies $\frac{\partial x^*}{\partial n} < 0$.

□

Proof. Lemma 2: For simplicity, I show Lemma 2 when $n = 2$, but similar derivations follow when considering $n > 2$. Individual i's utility function is

$$\begin{aligned} u_i(x_i, \widetilde{x}) &= (1-\kappa_i)\pi_i(x_i, \widetilde{x}) \\ &- \alpha_i(n-1)\max[\pi_j(x_i, \widetilde{x}) - \pi_i(x_i, \widetilde{x}), 0] \\ &- \beta_i(n-1)\max[\pi_i(x_i, \widetilde{x}) - \pi_j(x_i, \widetilde{x}), 0] \\ &+ \kappa_i\pi_i(\boldsymbol{x_i}). \end{aligned} \tag{56}$$

By setting $\alpha_i = -\beta_i$ and $\widetilde{\beta}_i \equiv \beta_i(n-1)$, I get

$$\begin{aligned} u_i(x_i, \widetilde{x}) &= (1-\kappa_i-\widetilde{\beta}_i)\pi_i(x_i, \widetilde{x}) \\ &+ \kappa_i\pi_i(\boldsymbol{x_i}) + \widetilde{\beta}_i\pi_j(x_i, \widetilde{x}). \end{aligned} \tag{57}$$

Following the approach of Lemma 1, I define:

- $\hat{\pi}(x_i, \widetilde{x}) \equiv \widetilde{\beta}_i\pi_j(x_i, \widetilde{x}) + \kappa_i\pi_i(x_i, x_i)$,

- $\overline{x_i} \in \arg\max_{x \in X} \hat{\pi}(x, \widetilde{x})$,
- $\underline{x_i} \in \arg\min_{x \in X} \hat{\pi}(x, \widetilde{x})$,
- $\hat{a} \equiv \frac{\hat{\pi}(\overline{x_i}, \widetilde{x}) + \hat{\pi}(\underline{x_i}, \widetilde{x})}{2}$,
- $\hat{b} \equiv \frac{\hat{\pi}(\overline{x_i}, \widetilde{x}) - \hat{\pi}(\underline{x_i}, \widetilde{x})}{2}$.

By subtracting $\hat{a}$ from (57) and dividing it by $(1 - \kappa_i - \widetilde{\beta}_i)\hat{b}$, I obtain:

$$\widetilde{u}_i(x_i, \widetilde{x}) = \widetilde{v}(\pi_i(x_i, \widetilde{x})) + \widetilde{\gamma}_i \widetilde{N}_i(x_i, \widetilde{x}), \tag{58}$$

where

- $\widetilde{v}(\pi_i(x_i, \widetilde{x})) \equiv 2\frac{\pi_i(x_i, \widetilde{x})}{\hat{\pi}(\overline{x_i}, \widetilde{x}) - \hat{\pi}(\underline{x_i}, \widetilde{x})}$,
- $\widetilde{\gamma}_i \equiv \frac{1}{1 - \widetilde{\beta}_i - \kappa_i}$,
- $N_i(x_i, \widetilde{x}) \equiv \widetilde{\beta}_i \pi_j(x_i, \widetilde{x}) + \kappa_i \pi_i(x_i, x_i)$,
- $\widetilde{N}_i(x_i, \widetilde{x}) \equiv 2\frac{N_i(x_i, \widetilde{x}) - \hat{\pi}(\underline{x_i}, \widetilde{x})}{\hat{\pi}(\overline{x_i}, \widetilde{x}) - \hat{\pi}(\underline{x_i}, \widetilde{x})} - 1 \in [-1, 1]$.

Finally, I normalize $\widetilde{\beta}_i + \kappa_i = 1$ which allows to write $N_i(x_i, \widetilde{x})$ as

$$N_i(x_i, \widetilde{x}) = (1 - \tau_i) N(x_i) + \tau_i \pi_j(x_i, \widetilde{x}), \tag{59}$$

with $\tau_i \in [0, 1]$. □

Proof. Proposition 7: The extended injunctive norm in the dictator game is given by

$$N_i(x, \widetilde{x}) = \frac{1}{2} v(x) + \frac{1}{2}(1 - \tau) v(w - x) + \tau \frac{1}{2} v(w - \widetilde{x}), \tag{60}$$

First, I consider the two polar cases. When $\tau = 0$, $x^* = \frac{w}{2}$ (see Proposition 2). When $\tau = 1$, $x^* = w$. When $\tau \in (0, 1)$, $N_i(x, \widetilde{x})$ is strictly concave in x, implying that x^* can be interior (i.e., $x^* \in (0, w)$) or in the boundary (i.e., $x^* = w$). When $x^* = \hat{x} \in (0, w)$, $\hat{x}$ satisfies

$$\frac{1}{2} v'(\hat{x}) - \frac{1}{2}(1 - \tau) v(w - \hat{x}) = 0. \tag{61}$$

By differentiating the previous expression with respect to τ, I find that $\hat{x}$ is increasing in τ,

$$\frac{\partial \hat{x}}{\partial \tau} = \frac{-\frac{1}{2}v'(w-\hat{x})}{\frac{1}{2}v''(\hat{x}) + \frac{1}{2}(1-\tau)v''(w-\hat{x})} \geq 0, \tag{62}$$

as both numerator and denominator are negative. This implies that $\hat{x} \in [\frac{w}{2}, w)$ (as $x^* = \frac{w}{2}$ when $\tau = 0$). On the other hand, a sufficient condition for having $x^* = w$ is that $\frac{\partial N_i(x,\widetilde{x})}{\partial x}\Big|_{x=w} \geq 0$, as if the derivative at $x = w$ is non-negative, it means that the derivative at any $x < w$ is also non-negative (for the strict concavity of $N_i(x,\widetilde{x})$). This implies that (61) can not be satisfied. This condition can be expressed as:

$$v'(w) - (1-\tau)v'(0) \geq 0, \tag{63}$$

which is equivalent to

$$\tau \geq 1 - \frac{v'(w)}{v'(0)} \equiv \overline{\tau}. \tag{64}$$

For the (strict) concavity of v, we have $v'(0) > v'(w) > 0$, which implies $\overline{\tau} \in (0,1)$. □

Proof. Proposition 8: The extended injunctive norm in the linear public goods game is given by

$$N_i(x,\widetilde{x}) = (1-\tau)v(w - x + \hat{A}nx) + \tau v(w - \widetilde{x} + (n-1)\hat{A}\widetilde{x} + \hat{A}x). \tag{65}$$

First, I consider the case with $\hat{A}n \geq 1$. When $\hat{A}n > 1$, the two terms of $N_i(x,\widetilde{x})$ are increasing in x. When $\hat{A}n = 1$, the first term does not depend in x, while the second term is strictly increasing in x. In both cases, $x^* = w \ \forall \tau \in [0,1]$.

Now, I consider the case with $\hat{A}n < 1$. In this case, the first term is strictly decreasing in x, while the second term is strictly increasing in x. Note that $N_i(x,\widetilde{x})$ is strictly concave in x as it is the sum of two strictly concave functions. When $x^* = \hat{x} \in (0,w)$ is interior, $\hat{x}$ satisfies:

$$(1-\tau)v'(w - \hat{x} + \hat{A}n\hat{x})(-1 + \hat{A}n) + \tau v'(w - \widetilde{x} + (n-1)\hat{A}\widetilde{x} + \hat{A}\hat{x})\hat{A} = 0. \tag{66}$$

By differentiating the previous expression with respect to τ and isolating $\frac{\partial \hat{x}}{\partial \tau}$, I find that

$$\frac{\partial \hat{x}}{\partial \tau} = \frac{v'(w - \hat{x} + \hat{A}n\hat{x})(\hat{A}n - 1) - v'(w - \widetilde{x} + (n-1)\hat{A}\widetilde{x} + \hat{A}\hat{x})\hat{A}}{v''(w - \hat{x} + \hat{A}n\hat{x})(\hat{A}n-1)^2(1-\tau) + v''(w - \widetilde{x} + (n-1)\hat{A}\widetilde{x} + \hat{A}\hat{x})\hat{A}^2\tau} \geq 0, \tag{67}$$

as both numerator and denominator are negative. On the other hand, the sufficient conditions for having $x^* = 0$ and $x^* = w$ are (i) $\frac{\partial N_i(x,\widetilde{x})}{\partial x}\Big|_{x=0} \leq 0$ and (ii) $\frac{\partial N_i(x,\widetilde{x})}{\partial x}\Big|_{x=w} \geq 0$. In the first case $N_i(x,\widetilde{x})$ is decreasing in x and $x^* = 0$, while in the second case $N_i(x,\widetilde{x})$ is increasing in x and $x^* = w$.

The first condition can be expressed as:

$$(1-\tau)v'(w)(-1+\hat{A}n) + \tau v'(w-\widetilde{x}+(n-1)\hat{A}\widetilde{x})\hat{A} \leq 0, \tag{68}$$

which is equivalent to:

$$\tau \leq \frac{v'(w)(1-\hat{A}n)}{v'(w)(1-\hat{A}n) + \hat{A}v'(w-\widetilde{x}+(n-1)\hat{A}\widetilde{x})} \equiv \underline{\tau}. \tag{69}$$

Note that $v'(w)(1-\hat{A}n) > 0$ and $\hat{A}v'(w-\widetilde{x}+(n-1)\hat{A}\widetilde{x}) > 0$ imply that $\underline{\tau} \in (0,w)$. On the other hand, the second condition can be expressed as:

$$(1-\tau)v'(\hat{A}nw)(-1+\hat{A}n) + \tau v'(w-\widetilde{x}+(n-1)\hat{A}\widetilde{x}+\hat{A}w)\hat{A} \geq 0, \tag{70}$$

which is equivalent to:

$$\tau \geq \frac{v'(\hat{A}nw)(1-\hat{A}n)}{v'(\hat{A}nw)(1-\hat{A}n) + \hat{A}v'(w-\widetilde{x}+(n-1)\hat{A}\widetilde{x}+\hat{A}w)} \equiv \overline{\tau}. \tag{71}$$

In this case $v'(\hat{A}nw)(1-\hat{A}n)$ and $\hat{A}v'(w-\widetilde{x}+(n-1)\hat{A}\widetilde{x}+\hat{A}w) > 0$ imply that $\overline{\tau} \in (0,w)$. □

Finally, to show that $\overline{\tau} > \underline{\tau}$, I define $\overline{\tau} = \frac{A}{A+B}$ with $A \equiv v'(\hat{A}nw)(1-\hat{A}n)$ and $B \equiv \hat{A}v'(w-\widetilde{x}+(n-1)\hat{A}\widetilde{x}+\hat{A}w)$ and $\underline{\tau} = \frac{C}{C+D}$ with $C \equiv v'(w)(1-\hat{A}n)$ and $D \equiv \hat{A}v'(w-\widetilde{x}+(n-1)\hat{A}\widetilde{x})$. Showing $\overline{\tau} > \underline{\tau}$ is equivalent to show that $A \times D > C \times B$. For the strict concavity of v and for $\hat{A}n < 1$, (i) $A > C$ and (ii) $D > B$, which gives the result.

Proof. Proposition 9: The injunctive norm is given by

$$N(x) = \frac{1}{2}v(w-x) + \frac{1}{2}v(mx), \tag{72}$$

where $m > 0$. When x^* is interior, then it satisfies

$$\frac{1}{2}v'(w-x^*) - \frac{1}{2}v'(mx^*)m = 0. \tag{73}$$

To determine $\frac{\partial x^*}{\partial m}$, I differentiate the previous expression with respect to m:

$$-v''(w-x^*)\frac{\partial x^*}{\partial m} - v''(x^*m)(x^* + \frac{\partial x^*}{\partial m}m)m - v'(mx^*) = 0, \tag{74}$$

which gives

$$\frac{\partial x^*}{\partial m} = \frac{-[mx^*v''(mx^*) + v'(mx^*)]}{[m^2v''(mx^*) + v''(w-x^*)]}. \tag{75}$$

As $v'' < 0$, the sign of $\frac{\partial x^*}{\partial m}$ depends on the sign of $mx^*v''(mx^*) + v'(mx^*)$. Using simple algebra, I obtain the result:

$$mx^*v''(mx^*) + v'(mx^*) = \tag{76}$$

$$= \frac{v'(mx^*)}{v'(mx^*)}[v''(mx^*)mx^* + v'(mx^*)] = \tag{77}$$

$$= v'(mx^*)[1 + \frac{mx^*v''(mx^*)}{v'(mx^*)}] = \tag{78}$$

$$= v'(mx^*)[1 - RRA(mx^*)], \tag{79}$$

where $RRA(mx^*) \equiv -\frac{mx^*v''(mx^*)}{v'(mx^*)}$. Therefore, x^* is increasing in m when $RRA(mx^*) < 1$, decreasing in m when $RRA(mx^*) > 1$ and constant in m when $RRA(mx^*) = 1$. □

Proof. Proposition 10: The extended injunctive norm in the volunteer's dilemma is given by

$$\begin{aligned} N_i(x,\widetilde{x}) &= (1-\tau)v(b(1-(1-x)^n) - cx) \\ &+ \tau v(b(1-(1-\widetilde{x})^{n-1}) + bx(1-\widetilde{x})^{n-1} - c\widetilde{x}). \end{aligned} \tag{80}$$

First, I consider three polar cases. When $\tau = 0$ or when $\widetilde{x} = 1$, $x^* = 1 - (\frac{c}{bn})^{\frac{1}{n-1}}$(see Proposition 6). When $\tau = 1$, $x^* = 1$.

When $\tau \in (0,1)$ and $\widetilde{x} \in [0,1)$, $N_i(x,\widetilde{x})$ is strictly concave in x. Thus, x^* can be interior (i.e., $x \in (0,1)$) or in the boundary (i.e., $x = 1$). When $x^* = \hat{x} \in (0,1)$, $\hat{x}$ satisfies $\frac{\partial N_i(x,\widetilde{x})}{\partial x}\Big|_{x=\hat{x}} = 0$, or equivalently

$$\begin{aligned} &(1-\tau)v'(b(1-(1-\hat{x})^n) - c\hat{x})(nb(1-\hat{x})^{n-1} - c) \\ + \quad &\tau v'(b(1-(1-\hat{x})(1-\widetilde{x})^{n-1}) - c\widetilde{x})b(1-\widetilde{x})^{n-1} = 0. \end{aligned} \tag{81}$$

By differentiating the previous expression with respect to τ and isolating $\frac{\partial \hat{x}}{\partial \tau}$, I get

$$\frac{\partial \hat{x}}{\partial \tau} = \frac{v'(b(1-(1-\hat{x})^n)-c\hat{x})-b(1-\widetilde{x})^{n-1}v'(b(1-(1-\hat{x})(1-\widetilde{x})^{n-1})-c\widetilde{x})}{(1-\tau)v''(b(1-(1-\hat{x})^n)-cx)(nb(1-\hat{x})^{n-1}-c)^2-(1-\tau)bnv'(b(1-(1-\hat{x})^n)-c\hat{x})+\tau v'(b(1-(1-\hat{x})(1-\widetilde{x})^{n-1})-c\widetilde{x})(b\hat{x}(1-\widetilde{x})^{n-1})^2}. \tag{82}$$

The denominator of the previous expression is negative, and therefore the sign of $\frac{\partial \hat{x}}{\partial \tau}$ depends on the sign of

$$v'(b(1-(1-\hat{x})^n)-c\hat{x})-b(1-\widetilde{x})^{n-1}v'(b(1-(1-\hat{x})(1-\widetilde{x})^{n-1})-c\widetilde{x}). \tag{83}$$

As $\hat{x}$ is an interior solution, (81) must be satisfied, implying that

$$\begin{aligned} &-\frac{\tau}{1-\tau}b(1-\widetilde{x})^{n-1}v'(b(1-(1-\hat{x})(1-\widetilde{x})^{n-1})-c\widetilde{x}) \\ &-b(1-\widetilde{x})^{n-1}v'(b(1-(1-\hat{x})(1-\widetilde{x})^{n-1})-c\widetilde{x}) \leq 0. \end{aligned} \tag{84}$$

Thus, $\frac{\partial \hat{x}}{\partial \tau} \geq 0$, which implies that $\hat{x} \in [1-(\frac{c}{bn})^{\frac{1}{n-1}},1)$ (as $x^* = 1-(\frac{c}{bn})^{\frac{1}{n-1}}$ when $\tau = 0$). On the other hand, a sufficient condition for having $x^* = 1$ is that $\frac{\partial N_i(x,\widetilde{x})}{\partial x}\Big|_{x=1} \geq 0$, as if the derivative at $x = 1$ is non-negative, the derivative at any $x < 1$ is also non-negative (for the strict concavity of $N_i(x,\widetilde{x})$). Therefore, (81) cannot be satisfied. This condition can be expressed as:

$$-(1-\tau)v'(b(1-c)(c)+\tau v'(b-c\widetilde{x})b(1-\widetilde{x})^{n-1} \geq 0, \tag{85}$$

which is equivalent to

$$\tau \geq \frac{v'(b-c)c}{v'(b-c)(c)+v'(b-c\widetilde{x})b(1-\widetilde{x})^{n-1}} \equiv \overline{\tau}. \tag{86}$$

Note that $v'(b-c)c > 0$ and $v'(b-c\widetilde{x})b(1-\widetilde{x})^{n-1} > 0$ implying $\overline{\tau} \in (0,1)$. □

Proof. Corollary 3: Let $\widetilde{N}_\tau(x) \in [-1,1]$ be the normalized appropriateness of strategy x for an individual with type $\tau \in [0,1]$. This is computed with the extended norm $N_i(x,\widetilde{x})$ in (16). Corollary 3 is equivalent to show that for any $x \neq \frac{w}{2}$ and $s_0 \geq \underline{s}$

$$\underbrace{s_0\widetilde{N}_0(\frac{w}{2})+(1-s_0)\widetilde{N}_\tau(\frac{w}{2})}_{\text{Average normalized appropriateness of } \frac{w}{2}} \geq \underbrace{s_0\widetilde{N}_0(x)+(1-s_0)\widetilde{N}_\tau(x)}_{\text{Average normalized appropriateness of } x}. \tag{87}$$

This is evident for any $x \neq \frac{w}{2}$ with $\widetilde{N}_\tau(x) \leq \widetilde{N}_\tau(\frac{w}{2})$.[42] Thus, I restrict attention to the case

[42] If type τ finds $\frac{w}{2}$ more socially appropriate than x, then the average appropriateness of $\frac{w}{2}$ is larger than the one of x for any $s_0 \in [0,1]$.

with $x \neq \frac{w}{2}$ with $\widetilde{N}_\tau(x) > \widetilde{N}_\tau(\frac{w}{2})$. Given that $\widetilde{N}_0(\frac{w}{2}) = 1$, (87) is given by

$$s_0 \geq \frac{\widetilde{N}_\tau(x) - \widetilde{N}_\tau(\frac{w}{2})}{1 - \widetilde{N}_0(x) + \widetilde{N}_\tau(x) - \widetilde{N}_\tau(\frac{w}{2})} \equiv s(x,\tau). \tag{88}$$

Then, I define $\underline{s}(\tau) \equiv \arg\max_{x \in [0,w]} s(x,\tau)$ and show that $\underline{s}(\tau) \in (0,1)$ for any $\tau \in (0,1]$. Note that both numerator and denominator of (88) are positive (as (i) $\widetilde{N}_\tau(x) > \widetilde{N}_\tau(\frac{w}{2})$, and (ii) $\widetilde{N}_0(x) < 1$). On the other hand, $1 - \widetilde{N}_0(x) > 0$ implies

$$1 - \widetilde{N}_0(x) + \widetilde{N}_\tau(x) - \widetilde{N}_\tau(\frac{w}{2}) > \widetilde{N}_\tau(x) - \widetilde{N}_\tau(\frac{w}{2}). \tag{89}$$

These two observations show that $\underline{s}(\tau) \in (0,1)$. Finally, defining $\underline{s} \equiv \arg\max_{\tau \in (0,1]} s(\tau)$ shows the result. □

C Lab Experiment

C.1 Social appropriateness ratings

The following tables display the number of participants that selected a given rating in the experiment.

Action	VSI	SI	SSI	SSA	SA	VSA
Give 0€	94	3	2	2	2	0
Give 1€	74	20	1	7	1	0
Give 2€	25	36	16	13	11	2
Give 3€	10	26	29	20	15	3
Give 4€	4	9	33	29	17	11
Give 5€	0	1	5	22	23	52
Give 6€	2	3	10	26	30	32
Give 7€	5	8	12	10	34	34
Give 8€	7	16	7	8	19	46
Give 9€	16	13	3	7	8	56
Give 10€	24	7	3	6	5	58

Dictator game with exogenous inequality (N = 103)

Action	VSI	SI	SSI	SSA	SA	VSA
Give 0€	90	7	2	0	0	1
Give 1€	62	27	7	2	1	1
Give 2€	7	34	10	23	8	18
Give 3€	5	21	22	26	17	9
Give 4€	5	3	26	32	21	13
Give 5€	2	2	5	29	21	41
Give 6€	6	6	12	17	26	33
Give 7€	9	12	12	7	24	36
Give 8€	16	13	8	4	16	43
Give 9€	24	8	7	2	9	50
Give 10€	29	6	3	2	7	53

Dictator game with endogenous inequality (N = 100)

Figure 19: Dictator game with joint production

Action	VSI	SI	SSI	SSA	SA	VSA
Give 0€	73	8	8	4	1	5
Give 1€	41	33	8	11	6	0
Give 2€	10	48	22	13	6	0
Give 3€	5	18	45	22	8	1
Give 4€	4	3	22	43	23	4
Give 5€	1	2	5	13	31	47
Give 6€	4	4	11	15	27	38
Give 7€	6	7	11	10	25	40
Give 8€	8	13	6	9	18	45
Give 9€	15	7	10	5	11	51
Give 10€	23	3	8	4	5	56

Dictator works (N = 99)

Action	VSI	SI	SSI	SSA	SA	VSA
Give 0€	97	3	2	1	0	1
Give 1€	86	11	1	2	2	2
Give 2€	64	26	6	4	3	1
Give 3€	42	41	10	7	3	1
Give 4€	28	33	28	7	7	1
Give 5€	8	9	26	29	16	16
Give 6€	8	6	12	35	33	10
Give 7€	8	2	10	25	48	11
Give 8€	7	3	11	15	41	27
Give 9€	7	4	8	11	29	45
Give 10€	8	2	5	4	8	77

Recipient works (N = 104)

Figure 20: Dictator game with earnings

Action	VSI	SI	SSI	SSA	SA	VSA
0€ Public Good	71	11	4	5	2	7
2€ Public Good	7	60	10	12	10	1
4€ Public Good	1	4	57	24	12	2
6€ Public Good	2	1	6	53	32	6
8€ Public Good	2	6	1	3	61	27
10€ Public Good	8	6	1	3	6	76

Efficient public goods game (N = 100)

Action	VSI	SI	SSI	SSA	SA	VSA
0€ Public Good	63	12	8	2	3	15
2€ Public Good	5	56	10	13	16	3
4€ Public Good	2	4	51	27	16	3
6€ Public Good	2	1	13	62	21	4
8€ Public Good	2	17	4	7	61	12
10€ Public Good	16	6	3	5	5	68

Inefficient public goods game (N = 103)

Figure 21: Linear public goods game

Action	VSI	SI	SSI	SSA	SA	VSA
Volunteer 0%	61	19	6	4	2	10
Volunteer 20%	4	49	23	10	11	5
Volunteer 40%	1	2	53	24	15	7
Volunteer 60%	0	1	6	58	28	9
Volunteer 80%	0	5	2	9	73	13
Volunteer 100%	6	2	2	8	10	74

Volunteer's dilemma with 3 group members. (N = 102)

Action	VSI	SI	SSI	SSA	SA	VSA
Volunteer 0%	58	19	9	4	3	8
Volunteer 20%	1	53	19	11	11	6
Volunteer 40%	1	1	45	33	14	7
Volunteer 60%	0	3	9	51	28	10
Volunteer 80%	0	12	7	6	55	21
Volunteer 100%	13	3	2	3	11	69

Volunteer's dilemma with 16 group members. (N = 101)

Figure 22: Volunteer's dilemma

Action	VSI	SI	SSI	SSA	SA	VSA
Choose S	4	0	3	6	21	68
Choose H	21	23	26	15	10	7

Stag hunt 1 (N = 102)

Action	VSI	SI	SSI	SSA	SA	VSA
Choose S	2	3	6	10	22	58
Choose H	22	17	21	21	13	7

Stag hunt 2 (N = 101)

Figure 23: Stag hunt game

Action	VSI	SI	SSI	SSA	SA	VSA
Choose A	2	0	0	4	19	75
Choose B	42	18	15	6	10	9

Coordination 1 (N = 100)

Action	VSI	SI	SSI	SSA	SA	VSA
Choose A	4	0	0	6	19	74
Choose B	43	22	13	7	11	7

Coordination 2 (N = 103)

Figure 24: Coordination game

Action	VSI	SI	SSI	SSA	SA	VSA
Choose C	2	3	4	14	25	55
Choose D	20	21	25	14	14	9

Prisoner's dilemma 1 (N = 103)

Action	VSI	SI	SSI	SSA	SA	VSA
Choose C	4	6	2	12	28	48
Choose D	15	26	25	15	10	9

Prisoner's dilemma 2 (N = 100)

Figure 25: Prisoner's dilemma

C.2 Evidence from the experimental questionnaire

In this section, I discuss two pieces of evidence that suggest that universalization reasoning is a primary determinant of individuals' evaluations. Two considerations are important to emphasize. First, several measures in this section are non-incentivized, so caution is needed when interpreting them. Second, despite both results supporting the proposed theory, the observed effects are quantitatively small. In Section C.2.1, I show that a universalization statement was the most relevant for participants to justify their evaluations. In Section C.2.2, I demonstrate that individual's degrees of universalization reasoning and norm-following are positively correlated.

C.2.1 Most relevant justification

Participants were asked to indicate on a seven-point scale how relevant five statements were when they evaluated an action as socially appropriate. I normalize participants' answers between -1 and 1 and compute the average score of each statement by averaging participants' answers. The five statements, evaluated on the same screen in random order, were based on mechanisms proposed in the literature.

- I considered that Person A was fair in selecting this action (*Fair*)
- I considered that Person A did not harm others in selecting this action (*Harm*)
- I considered that there were other actions that were more socially inappropriate (*Relative*)
- I considered that if everyone were to choose this action, the resulting outcome would be good for everyone (*Universalization*)
- I considered this action as something I would have chosen myself (*Myself*)

In Section 2, I assume that individuals' evaluations are determined solely through universalization reasoning. However, individuals may use other types of reasoning or combine several motivations. Therefore, while one should not expect to observe all the participants selecting only the universalization statement as relevant, it should nonetheless be selected as relevant by a considerable proportion of participants. Figure 26 shows

that the participants deemed the universalization statement most relevant to justify their decisions in the experiment.[43]

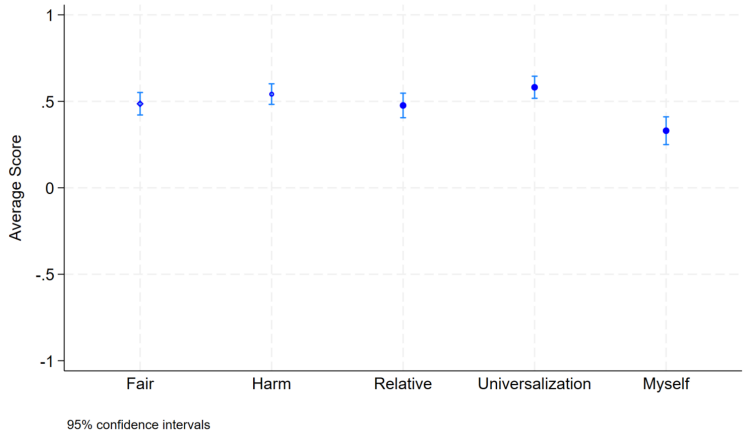


Figure 26: Average score of the five statements.

Although these differences are not quantitatively sizable, they provide evidence that the proposed thought process is relevant when evaluating social appropriateness, and, at the very least, comparable in magnitude to more standard explanations. This is in line with Levine et al. (2020) that find that a universalization statement was the most selected to explain why an action in a threshold problem was morally wrong. An explanatory analysis shows that the universalization statement was also the most preferred by the largest fraction of participants. More concretely, 55% of the participants evaluated the universalization statement as their most relevant statement. This is higher than the 41% for the fair statement, 45% for the harm statement, 44% for the relative statement, and 38% for the myself statement.[44]

C.2.2 Correlation between universalization reasoning and norm-following

In this section, I show that measures of γ_i and κ_i are positively correlated. I measure the participant's degree of norm-following with the task introduced in Kimbrough and

[43] Besides the harm statement, any difference between the average score of the universalization statement and other statements is statistically significant. I conduct paired comparison two-sided t-tests between the universalization statement and the fair ($p = 0.0142$), harm ($p = 0.3387$), relative ($p = 0.0189$) and myself ($p < 0.0001$) statements.

[44] Any difference between the universalization statement and another statement is statistically significant. I conduct paired comparison two-sided t-tests between the universalization statement and the fair ($p = 0.0025$), harm ($p = 0.0440$), relative ($p = 0.0258$), and myself ($p = 0.0004$) statements.

Vostroknutov (2018).[45] In this task, participants allocate 20 balls between a yellow and a blue bucket. Each ball they deposit in the blue (yellow) bucket gives them 0.05€ (0.10€). Participants are told, "The rule of the experiment is to put the balls only into the blue bucket." Therefore, they face a trade-off between maximizing their material payoff (by depositing the balls in the yellow bucket) and following the norm (by depositing the balls in the blue bucket). I use the number of balls participants deposit in the blue bucket to measure their norm-following degree.

On the other hand, I use the Oxford Utilitarianism Scale (Kahane et al., 2018) to measure participants' degree of universalization reasoning. For more details on the scale and on the distribution of participants' answers see Section C.3. Participants indicate, on a seven-point scale, their agreement with nine items. The scale is divided into two dimensions: (i) *impartial beneficence*, which measures individuals' support to maximize aggregate well-being, and (ii) *instrumental harm*, which evaluates their willingness to harm others to promote a greater good. I define participants' degree of universalization reasoning using their *inverse* score on the instrumental harm dimension. Intuitively, participants with a low score in the instrumental harm dimension are not willing to sacrifice their deontological moral principles over the greater good (e.g., killing someone over saving a larger number of individuals).

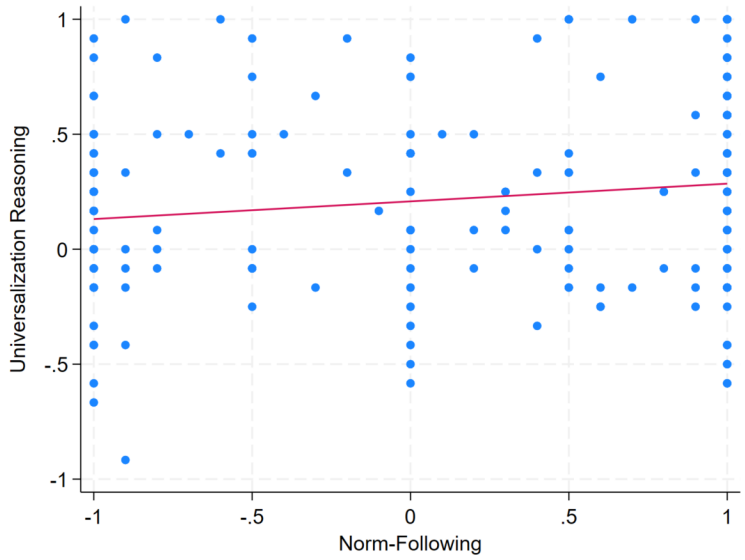


Figure 27: Relationship between norm-following and universalization reasoning.

[45] I thank Alexander Schneeberger for sharing his code for implementing the rule-following task in oTree. For other studies using this task, see Kimbrough and Vostroknutov (2018), Schneeberger and Krupka (2021) and Panizza et al. (2021).

Figure 27 shows a positive correlation between the two measures. This correlation is weak but statistically significant both with a Spearman rank correlation test ($\rho = 0.1529$, $p = 0.0295$) and a Person correlation test ($r = 0.1533$, $p = 0.029$). On the other hand, I do not find evidence that participants' degree of utilitarianism, measured with the impartial beneficence dimension, is correlated with their degree of norm-following, as indicated by Spearman and Person correlation tests ($p = 0.56$ and $p = 0.39$, respectively).

C.3 Oxford Utilitarianism Scale

The following tables display participants' answers in the Oxford Utilitarianism Scale and the distribution of their normalized scores. Figure 28 shows participants' answers to the nine items of the scale.

Item	Strongly Disagree	Disagree	Somewhat Disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly Agree
Item 1	22	27	48	11	38	36	21
Item 2	14	20	34	20	49	33	33
Item 3	19	32	46	16	33	35	22
Item 4	19	29	39	12	42	34	28
Item 5	13	17	43	22	60	35	13
Item 6	50	48	49	17	23	12	4
Item 7	39	28	27	32	49	15	13
Item 8	34	36	42	17	44	21	9
Item 9	35	37	41	22	44	13	11

Figure 28: Distribution of answer in the Oxford Utilitarianism Scale.

Items:

1. If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice.
2. From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we do not need two kidneys to survive, but really only one to be healthy.
3. From a moral perspective, people should care about the well-being of all human beings on the planet equally; they should not favor the well-being of people who are especially close to them either physically or emotionally.
4. It is just as wrong to fail to help someone as it is to actively harm them yourself.

5. It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal.
6. It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people.
7. If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used.
8. It is permissible to torture an innocent person if this would be necessary to provide information to prevent a bomb going off that would kill hundreds of people.
9. Sometimes it is morally necessary for innocent people to die as collateral damage—if more people are saved overall.

Figure 29 shows the distribution of participants' normalized scores on the universalization scale, constructed as the inverse of the Instrumental Harm score. The distribution of Instrumental Harm scores in this study aligns with the two studies reported in Kahane et al. (2018). Study 1 reports a mean score of $M_1 = 3.37$ ($SD_1 = 1.24$), while Study 2 reports $M_2 = 3.31$ ($SD_2 = 1.22$). These are similar to $M = 3.33$ ($SD = 1.23$) in the present study.

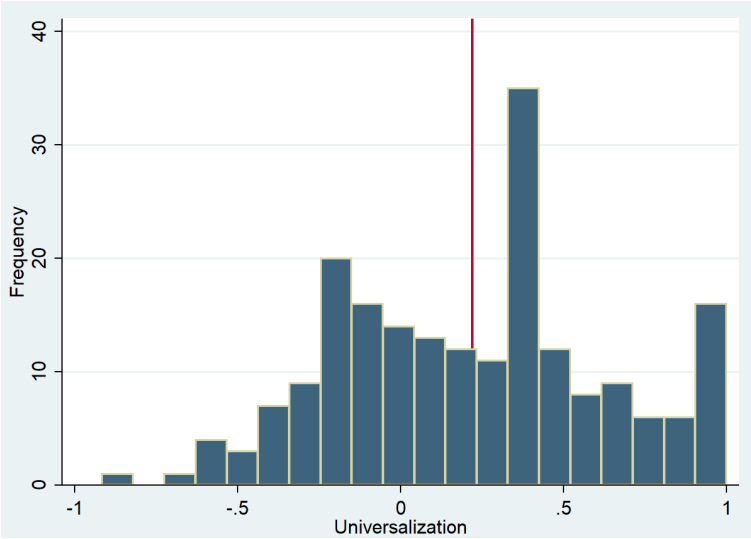


Figure 29: Universalization reasoning scale (normalized between -1 and 1).

Figure 30 shows participants' normalized Oxford Utilitarianism Scale (computed as the mean score over the nine items). The distribution of OUS scores is in line with the

two studies reported in Kahane et al. (2018) (see Table 8 in Kahane et al. (2018)). Study 1 reports a mean score of $M_1 = 3.58$ ($SD = 0.86$), while Study 2 reports $M_2 = 3.50$ ($SD = 0.92$). These are similar to $M = 3.81$ ($SD = 0.86$) from the present study.

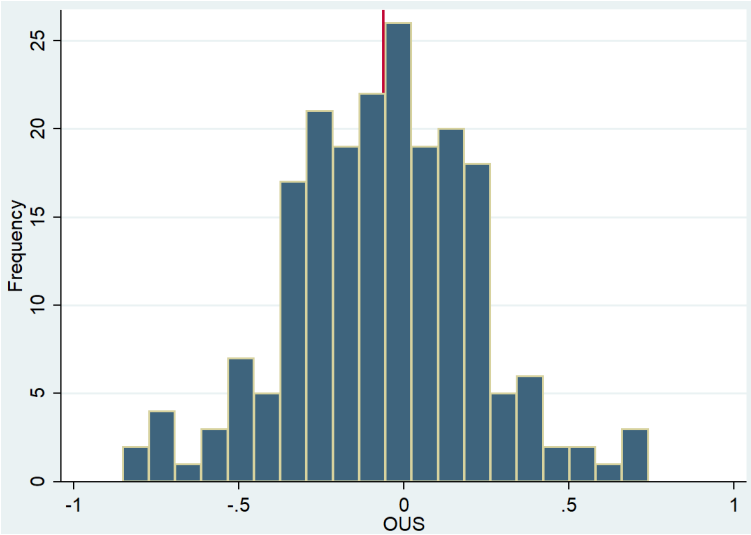


Figure 30: Oxford Utilitarianism Scale (normalized between -1 and 1).

C.4 Secondary and Robustness Tests

I this section, I conduct secondary tests detailed in the pre-registration and several robustness checks.

C.4.1 Linear public goods game

Regression Tables

Dependent Variable: Social appropriateness

	(1) Efficient	(2) Efficient	(3) Efficient	(4) Inefficient	(5) Inefficient	(6) Inefficient
Action	0.143***	0.143***	0.158***	0.103***	0.105***	0.093***
	(0.011)	(0.011)	(0.011)	(0.014)	(0.014)	(0.016)
Male	-	-0.079**	-	-	0.059	-
	-	(0.038)	-	-	(0.044)	-
Age	-	0.009	-	-	0.022	-
	-	(0.006)	-	-	(0.014)	-
Rule_Following	-	-0.003	-	-	0.001	-
	-	(0.002)	-	-	(0.003)	-
Constant	-0.628***	-0.756***	-0.699***	-0.455***	-0.944***	-0.421***
	(0.061)	(0.139)	(0.062)	(0.072)	(0.280)	(0.084)
Sample Passed Both CQ	No	No	Yes	No	No	Yes
Observations	600	594	420	618	612	474

Standard errors clustered at the individual level (in parentheses)

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

C.4.2 Volunteer's Dilemma

Test 1: More than half of the subjects evaluate $y = 1$ as "Appropriate to some extent".

For each variant, I conduct the following one-sided paired comparison t-test:

$$H_0 : \tilde{f}_{y=1}(Appropriate)(Variant) = 0.5$$
$$H_A : \tilde{f}_{y=1}(Appropriate)(Variant) > 0.5$$

I find that in *VD 3* (resp. *VD 16*), 90% (resp. 82%) of the subjects evaluate $y = 1$ as "Appropriate to some extent". This is statistically higher than 50% ($p < 0.0001$ in both cases).
Test 2: Less than half of the subjects evaluate $y = 0$ as "Appropriate to some extent".

For each variant, I conduct the following one-sided paired comparison t-test:

$$H_0 : \tilde{f}_{y=0}(Appropriate)(Variant) = 0.5$$
$$H_A : \tilde{f}_{y=0}(Appropriate)(Variant) < 0.5$$

I find that in *VD 3* (resp. *VD 16*), 15% (resp. 14%) of the subjects evaluate $y = 1$ as "Appropriate to some extent". This is statistically lower than 50% ($p < 0.0001$ in both cases).

C.4.3 Coordination game

Test 1: More than half of the subjects evaluate $y = X$ as "Appropriate to some extent".

For each variant, I conduct the following one-sided paired comparison t-test:

$$H_0 : \tilde{f}_{y=X}(Appropriate)(Variant) = 0.5$$
$$H_A : \tilde{f}_{y=X}(Appropriate)(Variant) > 0.5$$

I find that in *Coordination 1* (resp. *Coordination 2*), 98% (resp. 96%) of the subjects evaluate $y = X$ as "Appropriate to some extent". This is statistically higher than 50% ($p < 0.0001$ in both cases).
Test 2: Less than half of the subjects evaluate $y = Y$ as "Appropriate to some extent".

For each variant, I conduct the following one-sided paired comparison t-test:

$$H_0 : \tilde{f}_{y=Y}(Appropriate)(Variant) = 0.5$$
$$H_A : \tilde{f}_{y=Y}(Appropriate)(Variant) < 0.5$$

I find that in *Coordination 1* (resp. *Coordination 2*), 25% (resp. 24%) of the subjects evaluate $y = Y$ as "Appropriate to some extent". This is statistically lower than 50% ($p < 0.0001$ in both cases).

C.4.4 Stag hunt game

Test 1: More than half of the subjects evaluate $y = S$ as "Appropriate to some extent".

For each variant, I conduct the following one-sided paired comparison t-test:

$$H_0 : \tilde{f}_{y=S}(Appropriate)(Variant) = 0.5$$
$$H_A : \tilde{f}_{y=S}(Appropriate)(Variant) > 0.5$$

I find that in *Stag Hunt 1* (resp. *Stag Hunt 2*), 93% (resp. 89%) of the subjects evaluate $y = S$ as "Appropriate to some extent". This is statistically higher than 50% ($p < 0.0001$ in both cases).

Test 2: Less than half of the subjects evaluate $y = H$ as "Appropriate to some extent".

For each variant, I conduct the following one-sided paired comparison t-test:

$$H_0 : \tilde{f}_{y=H}(Appropriate)(Variant) = 0.5$$
$$H_A : \tilde{f}_{y=H}(Appropriate)(Variant) < 0.5$$

I find that in *Stag Hunt 1* (resp. *Stag Hunt 2*), 31% (resp. 40%) of the subjects evaluate $y = H$ as "Appropriate to some extent". This is statistically lower than 50% ($p = 0.0001$ and $p = 0.0292$, respectively).

C.4.5 Prisoner's Dilemma

Test 1: More than half of the subjects evaluate $y = C$ as "Appropriate to some extent".

For each variant, I conduct the following one-sided paired comparison t-test:

$$H_0 : \tilde{f}_{y=C}(Appropriate)(Variant) = 0.5$$
$$H_A : \tilde{f}_{y=C}(Appropriate)(Variant) > 0.5$$

I find that in *Prisoners Dilemma 1* (resp. *Prisoners Dilemma 2*), 91% (resp. 88%) of the subjects evaluate $y = C$ as "Appropriate to some extent". This is statistically higher than 50% ($p < 0.0001$ in both cases).

Test 2: Less than half of the subjects evaluate $y = D$ as "Appropriate to some extent".

For each variant, I conduct the following one-sided paired comparison t-test:

$$H_0 : \tilde{f}_{y=D}(Appropriate)(Variant) = 0.5$$
$$H_A : \tilde{f}_{y=D}(Appropriate)(Variant) < 0.5$$

I find that in *Prisoners Dilemma 1* (resp. *Prisoners Dilemma 2*), 35% (resp. 34%) of the subjects evaluate $y = D$ as "Appropriate to some extent". This is statistically lower than 50% ($p = 0.0019$ and $p = 0.0006$, respectively).

D Choice set effects with modified utility function

In this section, I show that with a modified utility $\widetilde{u}(x_i, x_j)$, the choice set effects observed in List (2007) can be explained. That is, (some) individuals choose $x > 0$ in the standard dictator game (over $x = 0$), but choose $x = 0$ in the dictator game with taking option (over $x > 0$). This reversal violates the Weak Axiom of Revealed Preference (WARP) because both $x = 0$ and $x > 0$ are available in the two interactions. It is well-known that this violation cannot be accommodated by complete and transitive preferences, such as *homo moralis*.

Here, $\widetilde{u}(x_i, x_j)$ is defined as follows:

$$\begin{aligned} \widetilde{u}(x_i, x_j) &= 2\frac{\pi(x_i, x_j) - \pi(\overline{x}^M, x_j)}{\pi(\overline{x}^M, x_j) - \pi(\underline{x}^M, x_j)} - 1 \\ &+ \tilde{\gamma}[2\frac{\pi(x_i, x_i) - \pi(\underline{x}^N, \underline{x}^N)}{\pi(\overline{x}^N, \overline{x}^N) - \pi(\underline{x}^N, \underline{x}^N)} - 1], \end{aligned} \tag{90}$$

where:

- $\overline{x}^M \in \arg\max_{x \in X} \pi(x, x_j)$
- $\underline{x}^M \in \arg\min_{x \in X} \pi(x, x_j)$
- $\overline{x}^N \in \arg\max_{x \in X} \pi(x, x)$
- $\underline{x}^N \in \arg\min_{x \in X} \pi(x, x)$

Specifically, $\overline{x}^M$ and $\underline{x}^M$ are the actions that maximize and minimize $\pi(x_i, x_j)$, respectively. In contrast, $\overline{x}^N$ and $\underline{x}^N$ are the actions that maximize and minimize $\pi(x_i, x_i)$. As explained in the main text, since the dictator game is an asymmetric game, one should consider the ex-ante symmetric game where transfers are chosen behind the veil of ignorance. In this case, $\pi(x_i, x_j) = \frac{1}{2}v(x_i) + \frac{1}{2}v(x_j)$, which highlights that x_j does not affect $\overline{x}^M$ and $\underline{x}^M$, and therefore the expression above is well defined.

Standard Dictator game $x \in [0, 10]$**:** In this case, $\overline{x}^M = 0$, $\underline{x}^M = 10$, $\overline{x}^N = 5$, and $\underline{x}^N = 0$.[46] Given that I am only interested in comparing the utility of choosing $x = 0$ and $x > 0$ (and not their exact value), I consider a simplified utility function neglecting terms that are

[46] Note that formally $\underline{x}^N = \{0, 10\}$. Since I am interested in the value $\pi(\underline{x}^N, \underline{x}^N)$, both give the same solution.

constant with $x = 0$ and $x > 0$.[47] More concretely, one can compute $\widetilde{u}(0, x_j)$ and $\widetilde{u}(x, x_j)$ as follows:

$$\begin{aligned} \widetilde{u}(0, x_j) &= \frac{\frac{1}{2}v(10) + \frac{1}{2}v(x_j)}{\frac{1}{2}v(10) - \frac{1}{2}v(x_j) - (\frac{1}{2}v(0) - \frac{1}{2}v(x_j))} + \tilde{\gamma}\frac{\frac{1}{2}v(10) + \frac{1}{2}v(0)}{v(5) - \frac{1}{2}(v(10) + v(0))} \\ &= \frac{v(10) + v(x_j)}{v(10) - v(0)} + \tilde{\gamma}\frac{v(10) + v(0)}{2v(5) - v(10) - v(0)} \end{aligned} \tag{91}$$

and

$$\begin{aligned} \widetilde{u}(x, x_j) &= \frac{\frac{1}{2}v(10 - x) + \frac{1}{2}v(x_j)}{\frac{1}{2}v(10) - \frac{1}{2}v(x_j) - (\frac{1}{2}v(0) - \frac{1}{2}v(x_j))} + \tilde{\gamma}\frac{\frac{1}{2}v(10 - x) + \frac{1}{2}v(x)}{v(5) - \frac{1}{2}(v(10) + v(0))} \\ &= \frac{v(10 - x) + v(x_j)}{v(10) - v(0)} + \tilde{\gamma}\frac{v(10 - x) + v(x)}{2v(5) - v(10) - v(0)}. \end{aligned} \tag{92}$$

There exists $\widetilde{\gamma}_x > 0$ such that if $\gamma > \widetilde{\gamma}_x$, then the individual prefers $x > 0$ over $x = 0$, while the opposite occurs when $\gamma < \widetilde{\gamma}_x$. Thus, to find $\widetilde{\gamma}_x$ I must find the γ such that $\widetilde{u}(0, x_j) = \widetilde{u}(x, x_j)$. With simple algebra, one can find that $\widetilde{\gamma}_x$ is given by:

$$\widetilde{\gamma}_x = \frac{v(10) - v(10 - x)}{v(10 - x) + v(x) - v(10) - v(0)} \times \frac{2v(5) - v(10) - v(0)}{v(10) - v(0)}. \tag{93}$$

Dictator game with Taking option $x \in [-1, 10]$**:** In this case, $\overline{x}^M = -1$, $\underline{x}^M = 10$, $\overline{x}^N = 5$, and $\underline{x}^N = -1$. Therefore, $\overline{x}^M$ and $\underline{x}^N$ vary with the choice set extension, while $\underline{x}^M$ and $\overline{x}^N$ remain unchanged. Following the same reasoning than before, one can characterize $\hat{\gamma}_x$ such that $\widetilde{u}(0, x_j) = \widetilde{u}(x, x_j)$. In this case,

$$\hat{\gamma}_x = \frac{v(10) - v(10 - x)}{v(10 - x) + v(x) - v(10) - v(0)} \times \frac{2v(5) - v(11) - v(-1)}{v(11) - v(0)}. \tag{94}$$

The choice set effect documented in List (2007) occurs when $\hat{\gamma}_x > \widetilde{\gamma}_x$. The reason is that, in this case, any individual with $\gamma \in (\widetilde{\gamma}_x, \hat{\gamma}_x)$ would prefer $x > 0$ over $x = 0$ in the standard dictator game, and $x = 0$ over $x > 0$ in the dictator game with taking options.[48] To compare $\widetilde{\gamma}_x$ and $\hat{\gamma}_x$, I must compare $\frac{2v(5) - v(10) - v(0)}{v(10) - v(0)}$ and $\frac{2v(5) - v(11) - v(-1)}{v(11) - v(0)}$. Hence, $\hat{\gamma}_x >$

[47] The idea is that if I want to compare $\widetilde{u}(0, x_j)$ and $\widetilde{u}(x, x_j)$, the terms (i) $-2\frac{\pi(\underline{x}^M, x_j)}{\pi(\overline{x}^M, x_j) - \pi(\underline{x}^M, x_j)}$, (ii) -1, (iii) $-2\frac{\pi(\underline{x}^N, \underline{x}^N)}{\pi(\overline{x}^N, \overline{x}^N) - \pi(\underline{x}^N, \underline{x}^N)}$, and (iv) $-\tilde{\gamma}$ are present in both $\widetilde{u}(0, x_j)$ and $\widetilde{u}(x, x_j)$, and, therefore, they cancel out.

[48] When $\gamma \in [0, \widetilde{\gamma}_x)$ or when $\gamma > \hat{\gamma}_x$, the individual would prefer $x = 0$ over $x > 0$ or $x > 0$ over $x = 0$ in the two games, respectively.

$\widetilde{\gamma}_x$ occurs when

$$\frac{2v(5)-v(11)-v(-1)}{v(11)-v(0)} > \frac{2v(5)-v(10)-v(0)}{v(10)-v(0)}, \tag{95}$$

or equivalently, when

$$[v(10)-v(0)][2v(5)-v(11)-v(-1)] > [v(11)-v(0)][2v(5)-v(10)-v(0)]. \tag{96}$$

Although I do not prove that this inequality holds for any possible strictly concave v, I show that it does for a general class of functions. More concretely, I assume that

$$v(x) = \frac{(x+1)^{(1-\rho)}-1}{1-\rho}, \tag{97}$$

with $\rho \in (0,1)$. This modified CRRA function has four desirable properties: (i) v is strictly concave (i.e., $v' > 0$ and $v'' < 0$), (ii) $v(0) = 0$, (iii) $v(-1)$ is negative and finite,[49] and (iv) $\frac{-(x+1)v''(x)}{v'(x)} = \rho$. Under (ii), the inequality in (96) simplifies to:

$$2v(5)[v(10)-v(11)] - v(10)v(-1) > 0. \tag{98}$$

Figure 31 shows that this inequality is satisfied for any $\rho \in (0,1)$.

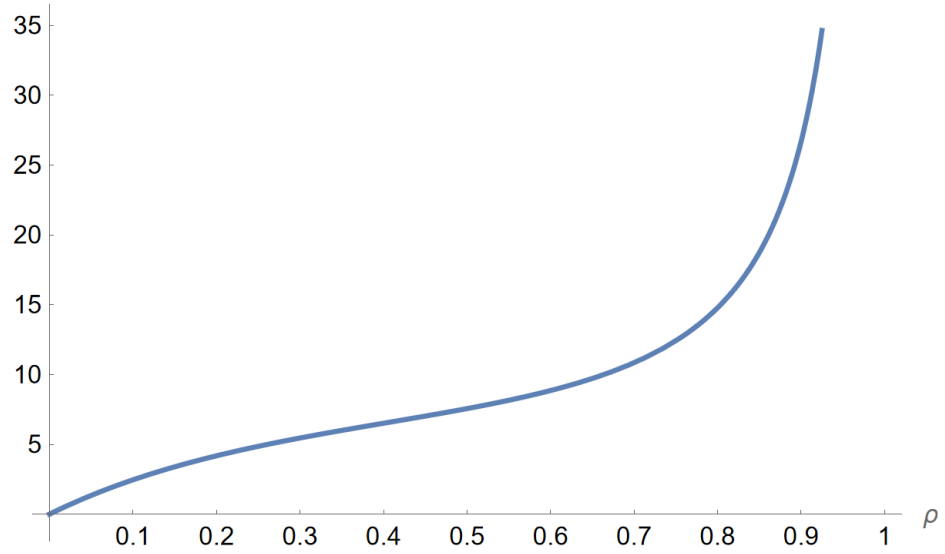


Figure 31: Displays $2v(5)[v(10)-v(11)] - v(10)v(-1)$ when $v(x) = \frac{(x+1)^{(1-\rho)}-1}{1-\rho}$ for any $\rho \in (0,1)$.

[49] The reason of using (97) over $v(x) = \frac{x^{(1-\rho)}}{1-\rho}$ is that with the latter $v(-1)$ is not defined for $0 < \rho < 1$.

References

Alger, I. and J. W. Weibull (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica 81*(6), 2269–2302.

Andreoni, J. and J. Miller (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica 70*(2), 737–753.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity, and social history. *Games and Economic Behavior 10*(1), 122–142.

Cappelen, A. W., A. D. Hole, E. Ø. Sørensen, and B. Tungodden (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review 97*(3), 818–827.

Cherry, T. L., S. Kroll, and J. F. Shogren (2005). The impact of endowment heterogeneity and origin on public good contributions: evidence from the lab. *Journal of Economic Behavior & Organization 57*(3), 357–365.

Güth, W., R. Schmittberger, and B. Schwarze (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization 3*(4), 367–388.

Hofmeyr, A., J. Burns, and M. Visser (2007). Income inequality, reciprocity and public good provision: an experimental analysis. *South African Journal of Economics 75*(3), 508–520.

Kahane, G., J. A. Everett, B. D. Earp, L. Caviola, N. S. Faber, M. J. Crockett, and J. Savulescu (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological review 125*(2), 131.

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

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

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

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American economic review 90*(4), 1072–1091.

Levine, S., M. Kleiman-Weiner, L. Schulz, J. Tenenbaum, and F. Cushman (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences 117*(42), 26158–26169.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy 115*(3), 482–493.

Mollerstrom, J., B.-A. Reme, and E. Ø. Sørensen (2015). Luck, choice and responsibility—an experimental study of fairness views. *Journal of Public Economics 131*, 33–40.

Panizza, F., A. Vostroknutov, and G. Coricelli (2021). The role of meta-context in moral decisions. *Unpublished manuscript). University of Trento and University of Southern California. http://www. vostroknutov. com/pdfs/metacontextPVCnext00. pdf*.

Reuben, E. and A. Riedl (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior 77*(1), 122–137.

Schneeberger, A. and E. L. Krupka (2021). Determinants of norm compliance: Moral similarity and group identification. *Available at SSRN 3969227*.