WORKING PAPERS

N° 1515

March 2024

"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 2024

First Draft: October 2023

Abstract

Recent studies have shown that individuals' behavior is sensitive to their perceptions of socially appropriate behavior. In this paper, I introduce a theory of injunctive norms in which individuals evaluate the social appropriateness of a given behavior using universalization reasoning. The theory allows one to compute the social appropriateness of any behavior without relying on individuals' expectations, preferences, and actual behavior. Furthermore, it can be applied to a wide range of interactions and rationalize several observations unaccounted for by theories of social preferences. I test the theory's predictions with evidence from past studies and new data from a lab 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, Enrico-Mattia Salonia, Sigrid Suetens, Roberto Weber, Jörgen Weibull, and especially 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 TSE's 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 multiple violations of the prediction that all individuals maximize their material payoff. Several alternative preferences have been proposed to account for these observations, 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)). Although these models rationalize several observations, some phenomena remain unexplained.[1]

Several studies have proposed models of norm-dependent preferences as an alternative framework (Cappelen et al. (2007), López-Pérez (2008), Kessler and Leider (2012), Krupka and Weber (2013), Kimbrough and Vostroknutov (2016), Kimbrough and Vostroknutov (2023)). In these models, individuals are driven to maximize their monetary payoff and behave socially appropriately. Recent studies have focused on empirical applications where *injunctive norms* (i.e., the shared belief of how socially appropriate a given behavior is) are elicited with the method introduced in Krupka and Weber (2013). Despite these elicitations' popularity and explanatory power, the current framework neither allows for predicting the injunctive norm nor understanding the reasoning behind individuals' evaluations. This paper proposes a theory to overcome these limitations and tests it in a lab experiment.

Specifically, I introduce a theory that ranks strategies by their social appropriateness and explains how individuals make this evaluation. In the simplest version of the theory, I show that existing models of norm compliance can be reformulated as models where individuals' utility function combines payoff maximization and Kantian morality (Alger and Weibull (2013), Alger and Weibull (2016)). This entails three main benefits. First, it introduces a concrete functional form that ranks strategies by their social appropriateness. Second, the injunctive norm is related to individuals' material payoff and is thus *endogenously* determined by the interaction. Third, the injunctive norm has evolutionary foundations, which is relevant for the close relationship between game theory and social norms (Kandori (1992), Young (2015)).[2]

The theory is grounded in *universalization reasoning*. When evaluating how socially appropriate a given behavior is, individuals examine what their material payoff would be if everyone in the population also chose it. If the strategy is chosen behind the *veil of ignorance* (i.e., before individuals know which role they will play in the interaction), the most socially appropriate strategy is the one that yields the highest expected material payoff if that strategy 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" and Rawls' veil of ignorance (Rawls (1971)) "the principles of justice are chosen behind a veil of ignorance." Using data from the experimental questionnaire, I show that this type of reasoning is of first-order importance in explaining individuals' moral judgments. More concretely, the participants of the experiment evaluated a universalization statement as the most relevant to justify their evaluations.

[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 give less to recipients when they have to work to generate their endowment. List (2007) shows that dictators' transfers decrease when they can take from recipients' endowments.

[2] For the evolutionary foundations of Kantian morality see Alger (2023). For its axiomatic foundations see Salonia (2023).

Four features of the theory are important to emphasize. First, I focus on injunctive norms (i.e., what one should do) rather than *descriptive norms* (i.e., what most people do). Although descriptive norms matter in explaining 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 in individuals' behavior (Krupka and Weber (2013), Kimbrough and Vostroknutov (2016)). Second, the injunctive norm prescribes the social appropriateness of strategies rather than of outcomes. This is in line with previous studies that define injunctive norms at the behavior rather than at the outcome level (Elster (1989), Krupka and Weber (2013)). Third, injunctive norms are homogeneous across individuals and thus do not vary with individuals' willingness to conform to them and their social preferences.[3] Finally, individuals' conformity to injunctive norms comes from the intrinsic utility derived from their attachment to them and not from external punishments or social pressure.

The theory's main strengths are its simplicity and portability. Simplicity is exemplified by its ability to compute the social appropriateness of any strategy available with minimal degrees of freedom. Portability is illustrated by computing the injunctive norm in a large set of symmetric and asymmetric games with an arbitrary number of players. Additionally, the theory can rationalize, in an unifying way, several puzzling 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)).

To test the predictions of the theory, I conduct a pre-registered lab experiment, which consists of seven game protocols.[4] In each protocol, participants are divided into two variants that only differ in one dimension. For example, in the dictator game with earnings, the endowment used in the dictator stage 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 conduct tests where I compare (i) the injunctive norm elicited in a variant with the corresponding prediction of the theory and (ii) how changes in a dimension of a variant affect both the elicited and predicted norms.

I consider three types of interactions. First, I examine symmetric two-player games with two actions in which either (i) both actions may implement a Nash equilibrium, and these are Pareto-ranked (e.g., stag hunt game) or (ii) one of these actions is strictly dominant (e.g., prisoner's dilemma). In both cases, 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 always evaluate the action that may implement the Pareto-dominant Nash equilibrium as more socially appropriate and that selecting a strictly dominant action may be socially inappropriate. The norms elicited in the lab experiment support these two predictions.

Second, I show that the evidence supports most (though not all) qualitative predictions of the theory in the standard dictator game (Forsythe et al. (1994)), dictator games with taking options (List (2007), linear public goods games (Isaac and Walker (1988)) and voluntary games (Diekmann (1985)). In particular, List (2007) finds that dictators decrease their transfers when given the option to take from recipients' endowments.

[3] See Section 6 for an extended version that allows for norm heterogeneity.

[4] The pre-registration and the instructions of the experiment can be found at `https://osf.io/g768h/`

The proposed theory predicts that low positive transfers are more socially appropriate when dictators can take from recipients' endowments. Moreover, the theory can explain the bystander effect, the phenomenon in which individuals are less likely to volunteer when other individuals are present (Latane and Darley (1968)). The theory predicts that it is more socially appropriate to volunteer the lower the number of individuals present.

Third, I study dictator games with production, where individuals exert effort to generate the endowment used in the dictator stage (Konow (2000), Cherry et al. (2002)). I show that the theory predicts that the most socially appropriate transfer is not the equal split but proportional to individuals' efforts. Thus, individuals are more likely to justify inequality when the source of it is effort rather than luck. This is in line with the empirical evidence (Corneo and Grüner (2002), Alesina and La Ferrara (2005), Lefgren et al. (2016)). Importantly, individuals' legitimacy over the endowment is endogenous to the interaction. This contrasts with most prominent models in the literature where fairness ideals are usually taken as exogenous (Cappelen et al. (2007)).

Although the empirical evidence supports most of the qualitative predictions of the theory, it fails to account for several important results. This suggests that universalization reasoning alone is not sufficient to explain all the variation in individuals' evaluations. Specifically, compared to the theory's predictions, participants evaluate too positively strategies that increase others' material payoff despite generating inequalities or inefficiencies. This is exemplified in the two following observations. First, in the standard dictator game, participants evaluate transfers above the equal split as more socially appropriate than the corresponding transfer below the equal split. Second, in the linear public goods game, participants evaluate it as socially appropriate to contribute to the public account, even when doing so is socially inefficient. On the other hand, the theory predicts that (i) complementary transfers in the dictator game are equally appropriate as they induce the same inequality, and (ii) it is only socially appropriate to contribute to the public account when it is socially efficient.

To explain these observations, I build upon recent results in evolutionary game theory (Alger et al. (2020)) and experimental economics (Van Leeuwen and Alger (2024)) to incorporate social concerns into the injunctive norm. In the extended version of the theory, the social appropriateness of a strategy combines Kantian morality and social concerns. I show that this extension can account for most of the results not explained by the simplest version of the theory.[5] Finally, I illustrate how this extension provides a theoretical framework for studying the norm heterogeneity observed in the data. This is important for the growing evidence documenting heterogeneity in norms and preferences across and within countries (Falk et al. (2018), Bursztyn and Yang (2022)).

The remainder of the paper is organized as follows: In the next section, I discuss the related literature. In Section 3, I present the theoretical framework. In Section 4, I describe the experimental design. In Section 5, I consider the theoretical predictions and the experimental results. In Section 6, I introduce the extended theoretical framework. In Section 7, I conclude.

[5] A key consideration when evaluating the predictive power of the theory is its degrees of freedom (see Miettinen et al. (2020) for a great discussion about this topic). While the simplest version of the theory is deliberately kept as simple as possible, the extended version includes additional degrees of freedom to increase its explanatory power.

2 Related Literature

This paper contributes to the literature studying the role of social norms in explaining individuals' decisions. Social norms have long been recognized as important for understanding individuals' behavior (Cialdini et al. (1990), Bicchieri (2005)). However, the difficulty of identifying and measuring them has restricted their use as explanations of outcomes not accounted for by standard models (Fehr and Gächter (2000), Fehr and Fischbacher (2004), Camerer and Fehr (2004), List (2007)). Krupka and Weber (2013) proposed a method for eliciting injunctive norms in an incentive-compatible way, which has been rapidly adopted in several subsequent papers.[6] I contribute to this growing empirical literature in two dimensions. First, I propose a theory to explain the injunctive norms elicited in previous studies and offer a potential explanation for how individuals evaluate what is socially appropriate and what is not. Second, I provide new evidence of the injunctive norms in several interactions.

Several theories of social norms have been proposed. In Cappelen et al. (2007), individuals are motivated by a desire for both income and fairness, and they suffer a disutility when they deviate from their fairness ideal. In López-Pérez (2008), individuals suffer a cost when violating an internalized social norm. In Kessler and Leider (2012), individuals suffer a disutility if they deviate from an agreed amount. In these models, norms are usually exogenous to the interaction, and the focus is on the effect of conformity to the norm rather than on the norm per se.[7] Ellingsen and Mohlin (2023) introduce a theory of social duties in which duties are classified in a 2x2 matrix according to their *conditionality* (i.e., if actions' consequences affect their classification) and their *strictness* (i.e., if they provide a minimum acceptable standard of behavior). The authors present a utility function in which decision-makers care about material outcomes and the fulfillment of duties. The two main differences between the two papers are the following. 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 here injunctive norms are endogenous to the interaction.

The closest paper is Kimbrough and Vostroknutov (2023), which, to the best of my knowledge is the only other paper that introduces a theory of the *content* of the injunctive norms. They propose an axiomatic model where injunctive norms arise endogenously from the set of available outcomes, with the social appropriateness of an outcome being determined by individuals' dissatisfaction with it. 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 strategy level. This distinction leads to differing norms, as Kimbrough and Vostroknutov (2023) assess the appropriateness of outcomes, whereas I evaluate the appropriateness of behaviors.[8] Importantly, my

[6] For other studies using this method see Gächter et al. (2013), Veselỳ (2015), Kimbrough and Vostroknutov (2016), Krupka et al. (2017), Bašić and Verrina (2021), 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.

[7] 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).

[8] To illustrate this difference, consider a two-player prisoner's dilemma where players have to simultaneously choose between (C)ooperate and (D)efect. Kimbrough and Vostroknutov (2023) gives a measure of how socially appropriate each outcome of the interaction is (i.e., (C, C), (C, D), (D, C) and (D, D)). In contrast, the proposed theory evaluates how socially appropriate

approach gives norms prescribing what one *should do*, (i.e., what behaviors are appropriate), which is more in line with previous literature (Elster (1989), Krupka and Weber (2013)). Additionally, I test the predictions of the theory with new data from a lab experiment. For the interested reader, I discuss the main differences between the two theories in Appendix D.

This paper is also related to the interdisciplinary literature on how individuals form moral judgments. Several mechanisms have been proposed, such as utilitarian reasoning (Blackorby et al. (2002)), violations of social norms and rules (Nichols and Mallon (2006), Schultz et al. (2007)), fairness concerns (Cappelen et al. (2007)) and emotions (Greene (2014), Cushman (2013)). In this strand of literature, the most related paper is Levine et al. (2020), a study in psychology that shows that universalization reasoning can explain individuals' moral judgments in threshold games.[9] The present paper generalizes the results from Levine et al. (2020) by introducing a broadly applicable theoretical framework and conducting a laboratory experiment that examines various interactions. Additionally, contrary to this paper's experiment, the experiments in Levine et al. (2020) do not incentivize participants' answers.

Finally, this paper relates to the literature on *homo moralis* preferences derived in Alger and Weibull (2013) by arguing that these preferences can be reinterpreted as norm-dependent preferences.[10] Homo moralis preferences have been used in theoretical studies to explain the behavior of individuals in different settings, such as voting (Alger and Laslier (2022)), taxation (Muñoz Sobrado (2022)), bargaining (Juan-Bartroli and Karagözoğlu (2024)), climate change (Eichner and Pethig (2021)), and team incentives (Sarkisian (2017)). Additionally, they have been tested in laboratory settings (Miettinen et al. (2020), Van Leeuwen and Alger (2024)).

3 The theoretical framework

In this section, I present the simplest version of the theory. In Section 3.1, I describe the framework in Krupka and Weber (2013). In Section 3.2, I introduce the injunctive norm. In Section 3.3, I illustrate how to compute this norm. In Section 3.4, I propose a function to normalize the injunctive norm into $[-1, 1]$.

3.1 The Krupka and Weber (2013) framework

Consider an interaction between two individuals: the decision-maker and the recipient. The decision-maker has to choose an action $a_k \in A = \{a_1, a_2, ..., a_K\}$ that determines his material payoff and the one of the recipient. The decision-maker cares about both the monetary payoff from the selected action (i.e., $\pi(a_k)$) and the degree to which the selected action is collectively perceived as socially appropriate (i.e., $N(a_k)$):

$$u_i(a_k) = V(\pi(a_k)) + \gamma_i N(a_k), \tag{1}$$

selecting each possible action is (i.e., C and D). The theories also give different predictions in games where strategy and outcome sets coincide (e.g., dictator games).

[9] In philosophy, besides the natural connection with Immanuel Kant's work (Kant (1785), Kant (1797)), this paper relates to Sugden (2003), which introduces the concept of team reasoning, where individuals think on what should they do as a group, rather than on their individual decisions.

[10] An alternative way of introducing morality is through the *Kantian equilibrium*, which modifies the equilibrium concept (Roemer (2010), Roemer (2015)).

where $V(\cdot)$ represents the value the individual attaches to the monetary payoff, which is assumed to be increasing in $\pi(a_k)$. The parameter $\gamma_i \geqslant 0$ represents the individual'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)). An individual with $\gamma_i = 0$ maximizes his monetary payoff, while the larger γ_i, the larger the amount of monetary payoff that the individual is willing to give up for behaving in a more socially appropriate manner.[11]

Finally, $N(a_k) \in [-1, 1]$ represents the shared belief on how socially appropriate selecting action a_k is. The larger $N(a_k)$, the more socially appropriate is selecting a_k collectively perceived. The injunctive norm is characterized by the profile of ratings of appropriateness over all the actions available. The main benefit of this framework is that $N(a_k)$ can be elicited in the lab (see Section 4), while $\pi(a_k)$ can be inferred from the payoffs of the interaction.

3.2 The proposed injunctive norm

Consider an interaction with two individuals and let X be the (common) set of pure strategies, which is a non-empty set that can be either discrete or both convex and compact. Let $\pi : X^2 \to R$ be individuals' material payoff function. I assume that individuals have *homo moralis* preferences (Alger and Weibull (2013)). That is, individual i's utility function when playing with individual j is given by

$$u_i(x_i, x_j) = (1 - \kappa_i)\pi(x_i, x_j) + \kappa_i \pi(x_i, x_i), \tag{2}$$

where x_i (resp. x_j) is the strategy selected by the individual i (resp. j), $\pi(x_i, x_j)$ is the material payoff of the individual i under the strategy profile (x_i, x_j). Therefore, $\pi(x_i, x_i)$ represents the individual i's material payoff in the hypothetical case that individual j were to choose the same strategy as himself. Interpreting the last term in (2) as a representation of Kant's categorical imperative (Kant (1785)), I refer to $\kappa_i \in [0, 1]$ as the individual i's degree of morality. The larger κ_i, the larger the weight individual i attaches to $\pi(x_i, x_i)$.[12]

I now show that by normalizing (2), I obtain a utility function of norm compliance (as in (1)) where the shared belief of the social appropriateness of strategy x_i is equal to the individual's (expected) material payoff if his pair were to choose x_i as well. Dividing the expression in (2) by $1 - \kappa_i$ and defining:

- $\gamma_i \equiv \frac{\kappa_i}{1-\kappa_i}$,
- $N(x_i) \equiv \pi(x_i, x_i)$,

I obtain the following expression:

$$\widetilde{u_i}(x_i, x_j) = \pi(x_i, x_j) + \gamma_i N(x_i). \tag{3}$$

Four remarks are important to emphasize. First, $N(x_i)$ applies to the set of available strategies, allowing strategies to vary in their social appropriateness. The profile of appropriateness ratings over all strategies available characterizes the injunctive norm. Second, $N(x_i)$ is homogeneous across individuals, implying that all

[11] An individual may want to adhere to injunctive norms for several reasons, such as to maintain a positive (self-)reputation (Benabou and Tirole (2006)) or to avoid feeling guilt (Charness and Dufwenberg (2006)).

[12] In Section 6, I extend (2) by incorporating social concerns.

individuals share the same injunctive norm, even when they differ in their willingness to conform to it. Third, the injunctive norms elicited with the Krupka and Weber (2013) method are bounded within $[-1, 1]$, while this is not the case for $N(x_i)$. In Section 3.4, I present a function to normalize $N(x_i)$ into $[-1, 1]$ while keeping its ranking prescribed unchanged. Finally, the theory predicts a positive correlation between norm-following and Kantian morality. I show support for this prediction in Section 5.4.[13]

When considering interactions with $N > 2$ individuals, I similarly define the social appropriateness of choosing strategy x_i to be equal to the individuals' expected material payoff in the hypothetical case all N group members were to select strategy x_i. Intuitively, individuals evaluate how socially appropriate a given behavior is by thinking about what would happen if everyone were to choose it as well. Another interpretation is that injunctive norms are based on the ranking of preferred strategies of the most moral individual in society. With the utility function (2), the most moral individual is the *homo kantiensis* (i.e., $\kappa_i = 1$) who maximizes the material payoff if everyone were to choose the same strategy.

Assumption 1. $\pi(x_i, x_j)$ and $\pi(x_i, x_i)$ are strictly concave in x_i.

Assumption 1 is standard in economics. In the context of the theory, it implies that when individuals have to divide a constant sum of money (e.g., in the standard dictator game), the more equal the division, the more socially appropriate it is. Note that this is the only assumption required for computing the ranking of strategies predicted by the theory. This means that the injunctive norm does not depend on other factors such as others' strategies, individuals' beliefs about others' strategies, or the parameters of individuals' utility.

3.3 Two examples

In this section, I exemplify how to compute the injunctive norm in the linear public goods game and in the standard dictator game. To do so, I calculate $N(t) = \pi(t, t)$ for each strategy $t \in X$. Throughout the paper, t^* represents the most socially appropriate strategy (i.e., the one that maximizes $N(t)$), while x_i^* represents individual i's optimal strategy (i.e., the one that maximizes $u(x_i, x_j)$). I discuss the properties of the norms derived in this section in Section 5.

First, I consider a linear public goods game with two individuals $i \in \{1, 2\}$ that decide the amount $x_i \in [0, w]$ they want to deposit to the public account. The material payoff of individual 1 is $\pi(x_1, x_2) = v(w - x_1 + \hat{A}(x_1 + x_2))$, where $\hat{A} \in (0, 1)$ is the marginal per capita return to the public good and v is strictly concave in x_1 (by Assumption 1). The injunctive norm in the two-player linear public goods game is given by

$$N(t) = v(w - t + 2\hat{A}t), \tag{4}$$

with v being strictly concave in t (by Assumption 1). Following the same arguments, the injunctive norm in the N-player linear public goods game is given by $N(t) = v(w - t + N\hat{A}t)$.

Second, I consider the standard dictator game, where dictators decide how to share an endowment of $w > 0$ between themselves and their pair. In asymmetric games, it is necessary to consider the ex-ante symmetric

[13] The purpose of the paper is not to provide a theory of γ. Instead, I interpret the positive correlation between γ_i and κ_i as providing some validity on the normalization conducted to achieve (17).

version of the game, where individuals have the same probability of being assigned to any role and select their strategy behind the veil of ignorance.[14] Otherwise, it would only be possible for some individuals to select some of the available strategies. In the ex-ante symmetric version of the game, each individual $i \in \{1, 2\}$ decides their transfer $x_i \in [0, w]$ in the dictator role and is assigned to either role with equal probability. Therefore, the expected material payoff of the individual 1 is $\pi(x_1, x_2) = \frac{1}{2}v(w - x_1) + \frac{1}{2}v(x_2)$. Thus, the injunctive norm in the dictator game is given by

$$N(t) = \frac{1}{2}v(w - t) + \frac{1}{2}v(t). \tag{5}$$

3.4 The normalization function

As mentioned, the injunctive norms elicited with the Krupka and Weber (2013) method are bounded within $[-1, 1]$. I propose a function $\tilde{z} : X \to [-1, 1]$ to normalize $N(t)$ into $[-1, 1]$. To do so, let $t^* \in \text{argmax}_{t \in X} N(t)$ and $t_* \in \text{argmin}_{t \in X} N(t)$ denote the most and least socially appropriate strategies.

Definition 1. Let $t^* \in \text{argmax}_{t \in X} N(t)$ and $t_* \in \text{argmin}_{t \in X} N(t)$. The **normalization function** $\tilde{z} : X \to [-1, 1]$ is defined as

$$\tilde{z}(t) = 2\frac{N(t) - N(t_*)}{N(t^*) - N(t_*)} - 1. \tag{6}$$

The ranking prescribed by $N(t)$ is maintained by $\tilde{z}(t)$, with $\tilde{z}(t^*) = 1$ and $\tilde{z}(t_*) = -1$.[15] Other functions could be used to normalize $N(t)$. In this paper, I focus on the strategies' relative social appropriateness rather than their absolute social appropriateness. For this purpose, any function that does not change the ranking predicted by $N(t)$ gives the same qualitative predictions.

When comparing the predictions of the theory in two related games $g \in \{A, B\}$, I compute $\tilde{z}^g(t)$ separately for each game by computing $t^*(g) \in \text{argmax}_{t \in X^g} N^g(t)$ and $t_*(g) \in \text{argmin}_{t \in X^g} N^g(t)$ for $g \in \{A, B\}$. This restricts both norms into $[-1, 1]$ while capturing the observation that, in any game, individuals evaluate some strategies as socially appropriate and others as socially inappropriate (see Section 5). Additionally, this aligns with the experimental design, in which participants evaluate the social appropriateness of just one of two related games.

As pre-registered, I do not use the normalization function when individuals only have two actions, as if I were to do so, the normalized appropriateness of these actions could only take values of -1 or 1. This would make comparing actions within the same game and the same action in different games more difficult (see Section 5.1). In this type of interaction, I derive predictions from the absolute value of $N(t)$.

4 The Experimental Design

To test the theory's predictions, I conduct a lab experiment that closely follows the methodology introduced in Krupka and Weber (2013). The experiment consists of seven different situations. Each situation has two variants

[14] Note that considering the ex-ante symmetric game does not mean that the theory can not be applied to asymmetric games without role uncertainty. For example, an individual who is assigned to be dictator (and knows he will not be recipient) may still reason *as if* others were to choose the same transfer as him.

[15] This normalization function is similar to the one proposed in Ferguson and Flynn (2016) in the context of a theory of moral relativism.

that differ in one 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 on a 6-point scale (as in Bašić and Verrina (2021)). 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. They only evaluate the social appropriateness of the actions and do not play in the evaluated games. 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 the following:

I. **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 5.1.1 for more details)

II. **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 5.1.2 for more details)

III. **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 5.1.3 for more details)

IV. **Linear public goods game**. Participants evaluate how socially appropriate they find Person A depositing $y \in \{0, 2, 4, 6, 8, 10\}$€ to the public account. The two variants differ in the return of the public account. (See Section 5.2.2 for more details)

V. **Volunteer's dilemma**. Participants evaluate how socially appropriate they find Person A volunteering with probability $y \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The two variants differ in the group size. (See Section 5.2.2 for more details)

VI. **Dictator game with earnings**. Participants evaluate how socially appropriate they find Person A transferring $y \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$€ in the dictator stage. The two variants differ in whether the dictator or the recipient works to generate the endowment. (See Section 5.3.2 for more details)

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

[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)).

the differences in individuals' contributions are for exogenous or endogenous reasons. (See Section 5.3.3 for more details)

The situations and variants were chosen for three main reasons. First, they include interactions of various natures, such as giving, coordination, 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, to the best of my knowledge, this paper provides new evidence of the injunctive norms in the coordination game with two Pareto-ranked Nash equilibria, stag hunt game, prisoners' dilemma, volunteer's dilemma, dictator game with joint production, and dictator game with earnings.

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, all participants completed a questionnaire that included a norm-following task (Kimbrough and Vostroknutov (2018)) and the Oxford Utilitarian Scale (Kahane et al. (2018)). I describe these two tasks in detail in Section 5.4. Appendix C 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 Euros (including a 5 Euro participation fee), with a minimum of 6 Euros and a maximum of 14 Euros. Participants were primarily French (82%), female (57%), studying economics-related majors (45%) and had 20.78 years old on average.

5 Theoretical predictions and experimental results

In this section, I present the theory's predictions and compare them with the empirical evidence. I divide this section into four parts. In Section 5.1, I consider one-shot symmetric two-player games with two actions. In Section 5.2, I study dictator and public goods games. In Section 5.3, I examine dictator games with production. In Section 5.4, I consider evidence from the experimental questionnaire.

5.1 One-shot symmetric two-player games

I start by computing the injunctive norm in a symmetric two-player game with two actions and arbitrary payoffs. Figure 1 specifies the material payoffs earned by the individuals for the corresponding combinations of actions.

[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 canceled at the beginning of the experiment because of internet issues. I do not consider this session for the analysis of the paper.

I denote by $G = (a, b, c, d)$ the game in Figure 1.

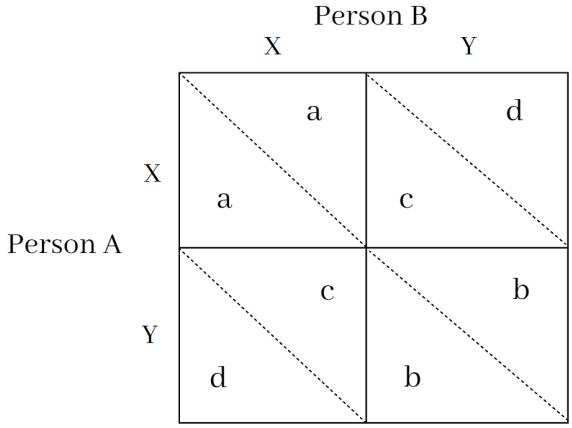


Figure 1: Two-player symmetric game with two actions.

With the proposed universalization reasoning, individuals compute the social appropriateness of selecting an action by considering their material payoff if the other individual also chose it. In line with the lab experiment, I restrict attention to pure strategies. Proposition 1 characterizes the social appropriateness of selecting 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 selecting Y. Second, changes in c and d do not affect the social appropriateness of selecting X and Y. In other words, the payoffs on the counter-diagonal of Figure 1 do not affect the injunctive norm. Finally, the social appropriateness of selecting actions X and Y increases in a and b, respectively.[20]

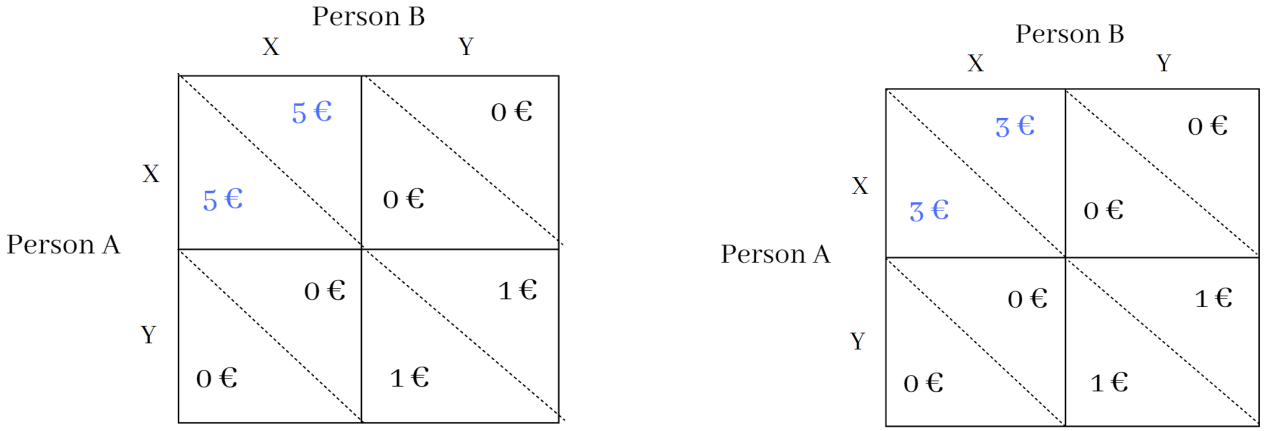
Proposition 1 has two implications. First, when (X, X) and (Y, Y) are Pareto-ranked Nash equilibria, selecting the action that may implement the Pareto-dominant Nash equilibrium is always more socially appropriate. Second, selecting a strictly dominant action may be socially inappropriate. For example, if Y is strictly dominant, selecting X is more socially appropriate when $a > b$. I test these two predictions in the following sections.

5.1.1 Coordination game with two Pareto-ranked Nash equilibria

Consider a game with $a > b > c = d$. This describes 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. Figure 2 shows the two variants

[20] As mentioned before, I do not use the normalization function in games with two actions, as otherwise, this last prediction would not be possible to derive.

evaluated in the experiment. The only difference between these variants is the payoff obtained when both individuals select X.[21]



a) Coordination 1 b) Coordination 2

Figure 2: Coordination game.

Testable predictions The theory has three predictions. First, in both variants, selecting X is more socially appropriate than selecting Y. Second, selecting X is more socially appropriate in *Coordination 1* than in *Coordination 2*. Finally, selecting Y is equally appropriate in both variants.

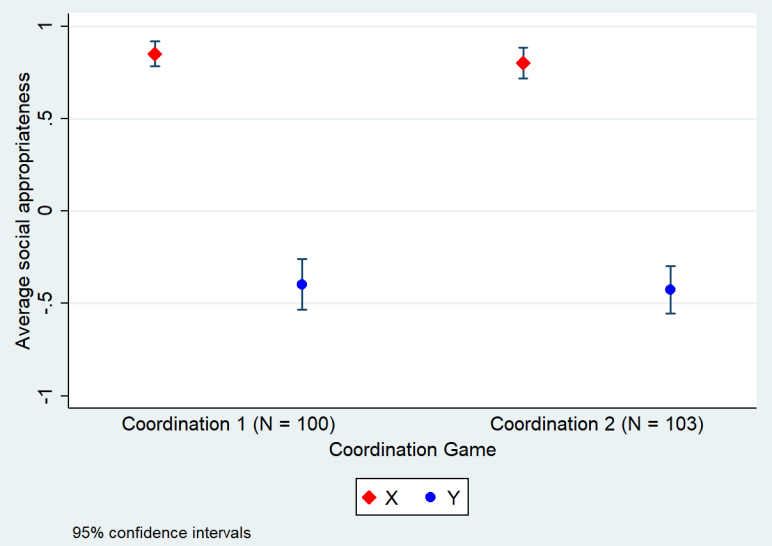


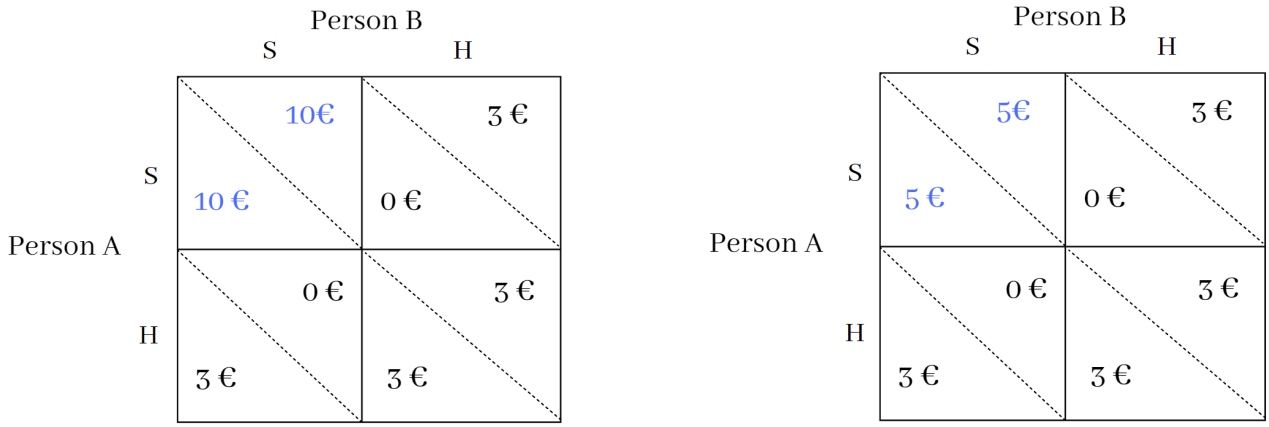
Figure 3: Average social appropriateness ratings in the coordination game.

[21] The magnitude of this difference, as well as those in other situations, was decided taking into account a trade-off between the fact that the magnitude is sufficiently large to be economically significant, but not too large to be unrealistic compared with past studies.

Results Figure 3 gives support for the predictions of the theory. First, in both variants, selecting X is more socially appropriate than selecting Y ($p < 0.0001$ in both cases).[22] Therefore, although 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*. However, this difference is statistically insignificant ($p = 0.17$).[23] Finally, selecting Y is equally appropriate in both variants ($p = 0.75$).

5.1.2 Stag hunt game

Consider a Stag hunt game with $a > b = d > c$. I relabel the actions to S(tag) and H(are). As before, (S, S) and (H, H) are Pareto-ranked Nash equilibria. The main difference from the previously considered game is that individuals can guarantee themselves the payoff obtained in the Pareto-dominated Nash equilibrium by selecting H. Therefore, while (S, S) is "payoff dominant", (H, H) is "risk dominant" (Rydval and Ortmann (2005)). Figure 4 shows the two variants evaluated in the experiment. The only difference between the two variants is the payoff obtained when both individuals select S.



a) Stag Hunt 1 b) Stag Hunt 2

Figure 4: Stag hunt game.

Testable predictions The theory has three predictions. First, for both variants, selecting S is more socially appropriate than selecting H. Second, selecting S is more socially appropriate in *Stag Hunt 1* than in *Stag Hunt 2*. Finally, selecting H is equally appropriate in both variants.

[22] 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.

[23] In several tests of the paper, the theory predicts the correct sign of the difference, but this difference is not statistically significant. This could happen because the true effect is zero or because the sample size is not sufficiently large to detect this difference.

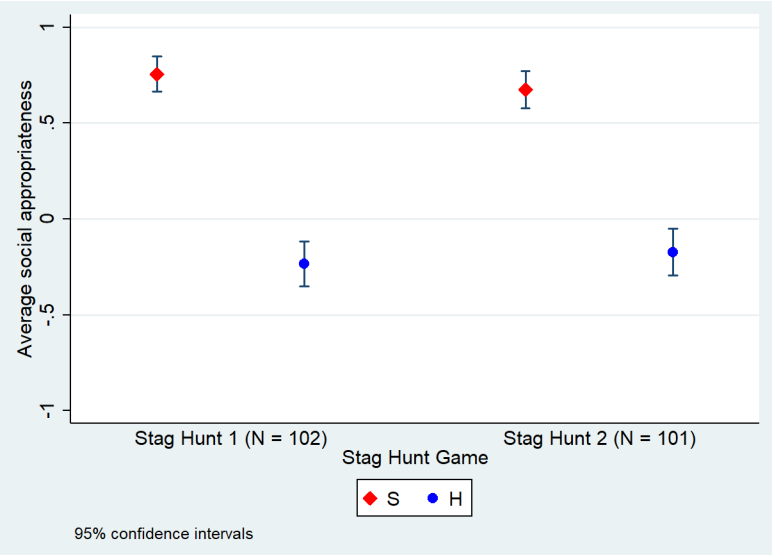
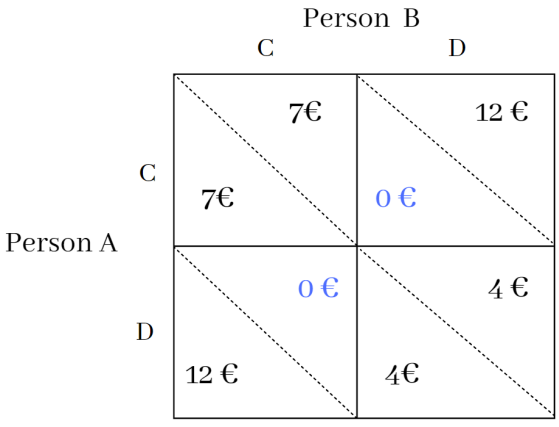


Figure 5: Average social appropriateness ratings in the stag hunt game.

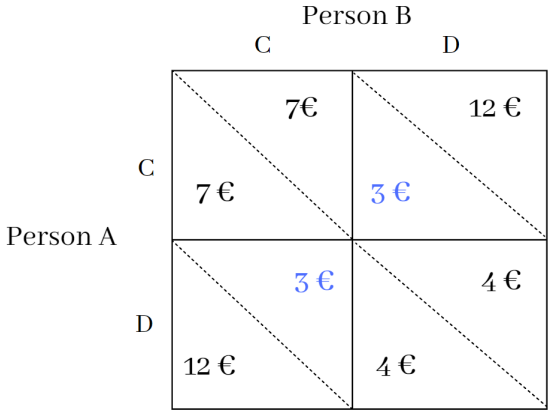
Results Figure 5 shows support for the predictions of the theory. First, for both variants, selecting S is more socially appropriate than selecting H ($p < 0.0001$ for both tests). Second, selecting S is more socially appropriate in *Stag Hunt 1* than in *Stag Hunt 2*. However, this difference is statistically insignificant ($p = 0.11$). Finally, selecting H is equally appropriate in both variants ($p = 0.46$).

5.1.3 Prisoner's dilemma

Consider a Prisoner's dilemma with $d > a > b > c$. This describes situations where, even though individuals would benefit from joint cooperation, each of them has incentives to deviate from it. I relabel the actions to C(operate) and D(efect). In this game, (D, D) is the unique Nash equilibrium, and D is a strictly dominant action. Figure 6 shows the two variants evaluated in the experiment. The only difference between the two variants is the payoff obtained when cooperating and the other individual defects.



a) Prisoner's dilemma 1



b) Prisoner's dilemma 2

Figure 6: Prisoner's dilemma.

Testable predictions The theory predicts that in both variants, selecting C is more socially appropriate than selecting D, and selecting C (D) is equally appropriate in both variants.

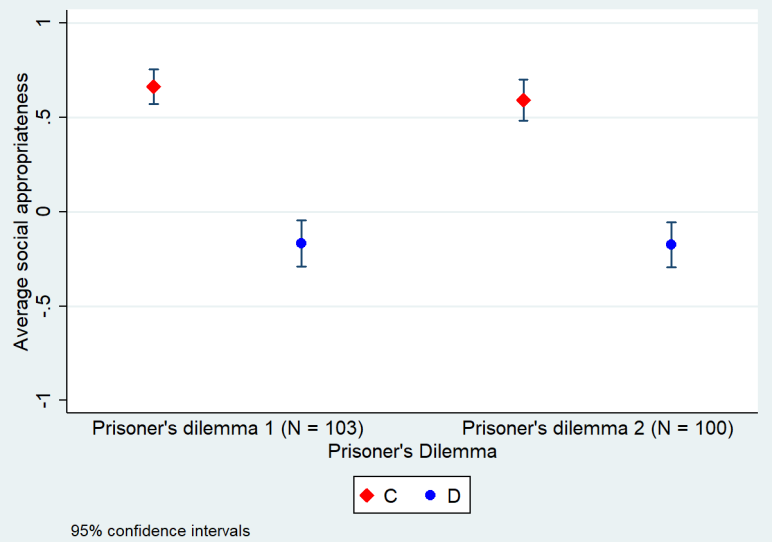


Figure 7: Average social appropriateness ratings in the Prisoner's dilemma.

Results Figure 7 shows support for the predictions of the theory. 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. Finally, selecting C (D) is equally appropriate in both variants ($p = 0.32$ and $p = 0.93$, respectively). This last result supports the prediction that payoffs in the counter-diagonal of the payoff matrix do not affect the social appropriateness of the different actions.

5.2 Dictator and public goods games.

5.2.1 Dictator games

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 to be in either role and select their transfer behind the veil of ignorance. The expected material payoff of individual 1 is

$$\pi(x_1, x_2) = \frac{1}{2}v(w - x_1) + \frac{1}{2}v(x_2), \tag{7}$$

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. Therefore, the injunctive norm in the dictator game is given by

$$N(t) = \frac{1}{2}v(w - t) + \frac{1}{2}v(t). \tag{8}$$

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

Proposition 2 shows that the most socially appropriate transfer is the equal split and that the injunctive norm is symmetric at $\frac{w}{2}$, increasing for transfers below $\frac{w}{2}$, and decreasing for transfers above $\frac{w}{2}$. Figure 8 displays the (normalized) injunctive norm predicted by the theory and the one elicited in Krupka and Weber (2013).[24]

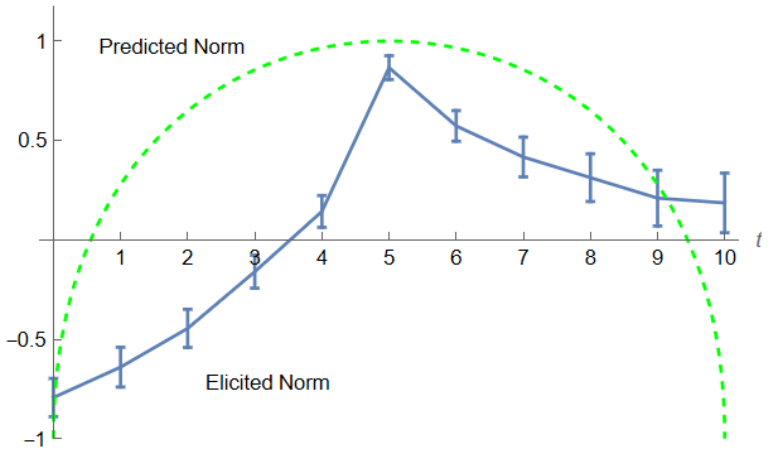


Figure 8: Normalized injunctive norm predicted (in dashed green) and elicited in Krupka and Weber (2013) (in blue) in the 10$ dictator game.

The norm elicited in Krupka and Weber (2013) supports the theory's predictions, except for the prediction of norm symmetry. In Section 6, I show how this result can be explained with the extended version of the theory.

Expanding the dictator's choice set. Recent studies have considered modified dictator games where dictators can take from recipients' endowments. If social preferences explain giving, taking options should not affect dictators' positive transfers. However, List (2007) shows that dictators decrease their transfers when given the option to take 1$ from recipients' endowments. Levitt and List (2007) argue that this result could be explained by a change in social norms resulting from the extension of dictators' choice sets. Proposition 3 shows how the social appropriateness of *existing* alternatives changes by adding a *new* alternative.

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 $t^*(A) \in \operatorname{argmax}_{t \in X} N^A(t)$ and $t_*(A) \in \operatorname{argmin}_{t \in X} N^A(t)$ denote the most and least socially appropriate strategies in game A. Let $\tilde{z}^g(t)$ denote the normalized social appropriateness of strategy t in game g.*

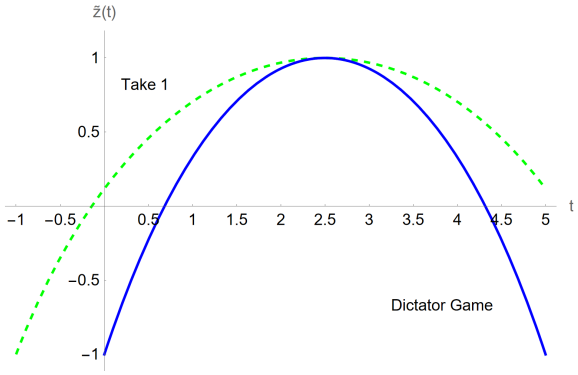
- ***Case 1*** *(x' is neither the most nor the least socially appropriate action): If $N^A(t^*(A)) > N^B(x') > N^A(t_*(A))$, then (i) $\tilde{z}^A(t) = \tilde{z}^B(t)$ $\forall t \in X$ and (ii) $\tilde{z}^B(x') \in (-1, 1)$.*

- ***Case 2*** *(x' is the most socially appropriate action): If $N^B(x') > N^A(t^*(A))$, then (i) $z^A(t) > z^B(t)$ $\forall t \in X$ with $\tilde{z}^A(t) > -1$ and (ii) $\tilde{z}^B(x') = 1$.*

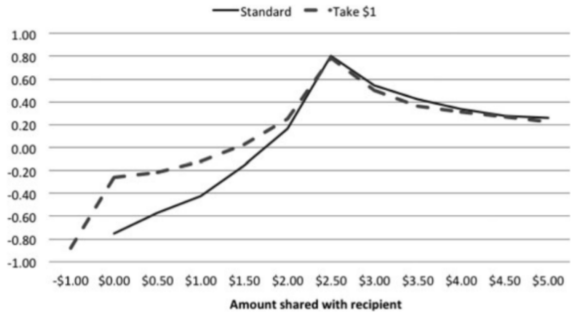
[24] The injunctive norm in the dictator game has been elicited in other studies (e.g., Kimbrough and Vostroknutov (2016), Bašić and Verrina (2021)) with similar qualitative patterns.

- ***Case 3*** *(x' is the least socially appropriate action): If $N^B(x') < N^A(t_*(A))$, then (i) $\tilde{z}^A(t) < \tilde{z}^B(t)$ $\forall t \in X$ with $\tilde{z}^A(t) < 1$ and (ii) $\tilde{z}^B(x') = -1$.*

Thus, the theory predicts that including an additional alternative only affects the social appropriateness of existing alternatives when this alternative is more (or less) socially appropriate than all the existing ones. Note that this result is driven by the normalization function proposed in Section 3.4.



a) Normalized injunctive norms predicted



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

Figure 9: Dictator game and take-1 condition.

In the modified dictator game in List (2007), the theory predicts that taking 1$ from recipients' endowments is the most socially inappropriate action (as it maximizes the inequality between the players while keeping the sum of payoffs constant). Therefore, despite that, in both conditions, the most socially appropriate transfer is the equal split; transferring a low positive amount is more socially appropriate in Take-1 than in the dictator game (see Figure 9a). The norms elicited Krupka and Weber (2013) support the predictions of the theory (see Figure 9b).

5.2.2 Public goods game

Linear public goods game. Individuals decide how to 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 $n \geqslant 2$ group members. Thus, 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.

Therefore, 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:

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

Proposition 4. *The most socially appropriate contribution to the public account is:*

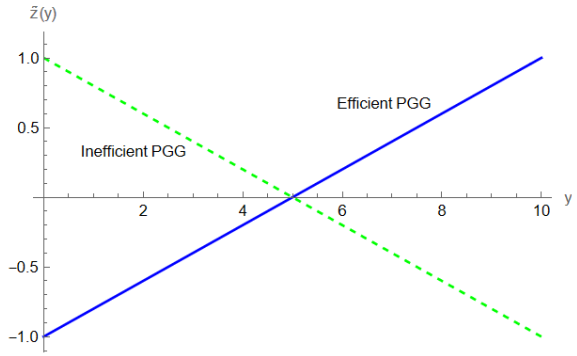
$$t^* = \begin{cases} w & \text{if } \hat{A} \in (\frac{1}{n}, 1) \\ [0, w] & \text{if } \hat{A} = \frac{1}{n} \\ 0 & \text{if } \hat{A} \in (0, \frac{1}{n}) \end{cases}$$

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

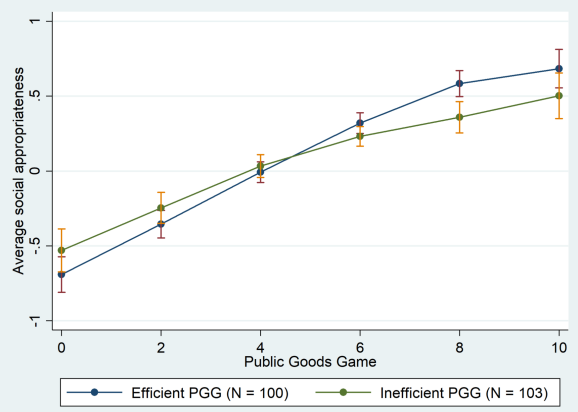
Proposition 4 predicts that individuals evaluate contributions to the public account as socially appropriate only when it is socially efficient. More concretely, it predicts a positive relationship between contributions to the public account and social appropriateness when $\hat{A}n > 1$ and a negative 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 $y \in \{0, 2, 4, 6, 8, 10\}$€ to the public account and $(10 - y)$€ to the private account.[25]

Testable predictions The theory has four predictions (see Figure 10a). First, contributing a low amount to the public account is more socially appropriate in *Inefficient PGG* than in *Efficient PGG*. Second, contributing a high amount to the public account is more socially appropriate in *Efficient PGG* than in *Inefficient PGG*. Third, there is a positive relationship between contributions to the public account and social appropriateness in *Efficient PGG*. Finally, there is a negative relationship between contributions to the public account and social appropriateness in *Inefficient PGG*.



a) Normalized injunctive norms predicted.



b) Average social appropriateness.

Figure 10: Linear public goods game.

Results Figure 10b shows support for the first three predictions of the theory. First, contributing $y = 0$ and $y = 2$ is more socially appropriate in *Inefficient PGG* than in *Efficient PGG*. Second, contributing $y = 8$ and $y = 10$ 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). Third, I find evidence

[25] Kimbrough and Vostroknutov (2016) elicit the *Efficient PGG* condition with similar qualitative patterns. On the other hand, Abbink et al. (2017) elicit a modified *Inneficient 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.

of a positive relationship between contributions and social appropriateness in *Efficient PGG*. On the other hand, I do not find evidence of a negative relationship between contributions and social appropriateness in *Inefficient PGG*.[26] In Section 6, I show that this result can be rationalized with the extended version of the theory.

The volunteer's dilemma. Individuals are assigned to a group of size $n \geqslant 2$ and decide whether to volunteer or not. When at least one of the individuals 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 to volunteer themselves, compared to no one volunteering, but they prefer someone else to do so.

Let $x_i \in [0, 1]$ denote the probability that individual i volunteers. From i's perspective, the probability that at least one of the other individuals volunteer is $1 - \prod_{j \neq i}(1 - x_j)$. Individual i's expected material payoff under action profile $(x_1, ..., x_n)$ is given by

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

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.[27] This effect has been documented both in the lab and the field (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 that multiple individuals volunteer simultaneously (which is socially inefficient). The injunctive norm is given by

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

Proposition 5. *The most socially appropriate volunteering probability is* $t^* = 1 - (\frac{c}{bn})^{\frac{1}{n-1}} \in (0, 1)$. *Additionally, (i)* $\frac{\partial t^*}{\partial n} \leqslant 0$ *and (ii)* $\frac{\partial N(t)}{\partial t} > 0$ *for any* $t \in [0, t^*)$ *and* $\frac{\partial N(t)}{\partial t} < 0$ *for any* $t \in (t^*, 1]$.

Proposition 5 predicts that the most socially appropriate probability of volunteering decreases with group size. This result is related to the *diffusion of responsibility principle*, which states that individuals tend to divide their responsibility to help others by the number of individuals present (Latane and Darley (1968)). The larger the number of individuals, the lower the responsibility one has for the group, and consequently, the less socially appropriate it is to volunteer.

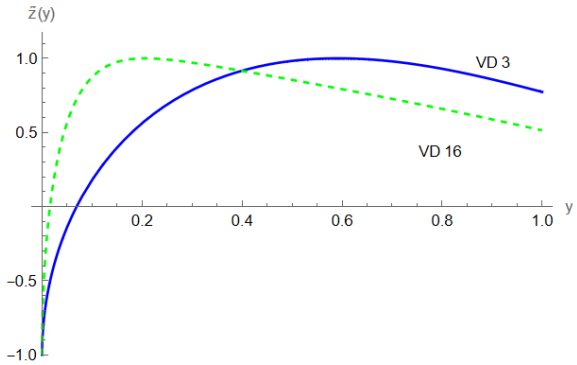
Experimental details Participants evaluated a situation in which Person A is in a group with either $N = 3$ (*VD 3*) or $N = 16$ (*VD 16*). Individuals have to simultaneously choose between volunteering or not. If no individual volunteers, all individuals earn 0€. If at least one individual volunteers, all individuals earn 10€ at

[26] The regression tables of these two last results are in Appendix C. The results are robust to participants' answers to the comprehension questions, which suggests that they can not be attributed to participants' misunderstanding of the task.

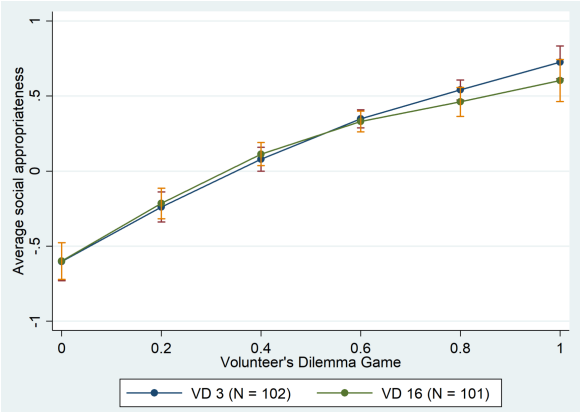
[27] Note that when all individuals maximize their material payoff, the unique (symmetric) equilibrium probability of volunteering is decreasing in n (i.e., $x_{NE} = 1 - (\frac{c}{b})^{\frac{1}{n-1}}$), while the probability that no one volunteers is increasing in n.

a cost of 5€ for the volunteers. Participants evaluate how socially appropriate they find Person A volunteering with probability $y \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$.[28]

Testable predictions The theory has four predictions (see Figure 11a): First, volunteering with low probability is more socially appropriate in *VD 16* than in *VD 3*. Second, volunteering with high probability is more socially appropriate in *VD 3* than in *VD 16*. Third, in both variants, volunteering with certainty is more socially appropriate than volunteering with zero probability. Finally, the most socially appropriate probability of volunteering is lower in *VD 16* than in *VD 3*.



a) Normalized injunctive norms predicted.



b) Average social appropriateness.

Figure 11: Volunteer's dilemma.

Results Figure 11b shows partial support for the predictions of the theory. First, selecting $y = 0$ and $y = 0.2$ is more socially appropriate in *VD 16* than in *VD 3*. However, these differences are small and not statistically significant ($p = 0.48$ and $p = 0.37$, respectively). Second, selecting $y = 0.8$ and $y = 1$ is more socially appropriate in *VD 3* than in *VD 16* ($p = 0.08$ in both cases). This provides some evidence of the diffusion of responsibility effect predicted by the theory. Additionally, it contradicts the predictions of a utilitarian norm where individuals care about helping the maximum number of individuals. Third, in both variants, volunteering with certainty is more socially appropriate than volunteering with null probability ($p < 0.0001$ in both cases).

On the other hand, I do not find evidence that the most socially appropriate probability of volunteering in *VD 16* is lower than in *VD 3*, as it is equal to 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 gives suggestive evidence supporting this fourth prediction. I discuss more about this result in Section 6.

[28] To explain mixed strategies, I follow the experimental instructions in Hillenbrand and Winter (2018), where participants were asked to use mixed strategies in a volunteer's dilemma by depositing balls of different colors into an urn.

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

5.3 Dictator games with production

In this section, I consider dictator games with production. In Section 5.3.1, I derive the theory's predictions in the general game. I apply these predictions to the dictator game with earnings (Section 5.3.2) and joint production (Section 5.3.2).

5.3.1 General game

The dictator games with production consist of two different stages. In the first stage, individuals are matched in pairs, and one (or both) works to generate an endowment. In the second stage, dictators divide the endowment generated at $t = 1$. In this game, individual i chooses (behind the veil of ignorance) a strategy $x_i = ({e_1}^i, {e_2}^i, t_i({e_1}^i, {e_2}^j))$ 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., $t_i({e_1}^i, {e_2}^j)$). 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 (i.e., ${c_k}^i({e_k}^i) = c({e_k}^i)$ for any $i \in \{1, 2\}$ and $k \in \{1, 2\}$). 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 (x_i, x_j) is the following:

$$\pi(x_i, x_j) = \frac{1}{2}\underbrace{[v(w({e_1}^i, {e_2}^j)) - t_i({e_1}^i, {e_2}^j) - c({e_1}^i)]}_{\text{Individual } i \text{ is dictator}} + \frac{1}{2}\underbrace{[v(t_j({e_1}^j, {e_2}^i) - c({e_2}^i)]}_{\text{Individual } i \text{ is recipient}}. \tag{13}$$

Therefore, the social appropriateness of choosing a strategy $x = (e_1, e_2, t(e_1, e_2))$ is given by

$$N(x) = \frac{1}{2}v(w(e_1, e_2) - t(e_1, e_2) - c(e_1)) + \frac{1}{2}v(t(e_1, e_2) - c(e_2)). \tag{14}$$

Proposition 6. *The most socially appropriate transfer is* $t^*(e_1, e_2) = \min\{\max\{\frac{w(e_1,e_2)}{2} + \frac{c(e_2)-c(e_1)}{2}, 0\}, w(e_1, e_2)\}$. *Additionally,* $\frac{\partial N(t)}{\partial t} > 0$ *for any* $t \in [0, t^*)$ *and* $\frac{\partial N(t)}{\partial t} < 0$ *for any* $t \in (t^*, w(e_1, e_2)]$.

Proposition 6 shows that the most socially appropriate transfer allocates a larger share of the resources to the individual that exerts a higher 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 by receiving a higher share of the aggregate resources. This is consistent with the evidence documenting that most individuals hold meritocratic views in games with joint production (Konow (2000), Cappelen et al. (2010), Luhan et al. (2019)). Note that when individuals evaluate the social appropriateness of the transfer at $t = 2$, they consider the efforts exerted at $t = 1$. This is in line with the evidence documenting that individuals take into account *sunk costs* when deciding present and future behavior (Arkes and Blumer (1985), Friedman et al. (2007), Buchheit and Feltovich (2011), Ronayne et al. (2021)).

5.3.2 Dictator game with earnings

In the 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*, efforts as recipients are restricted to zero. Thus, the injunctive norm is given by

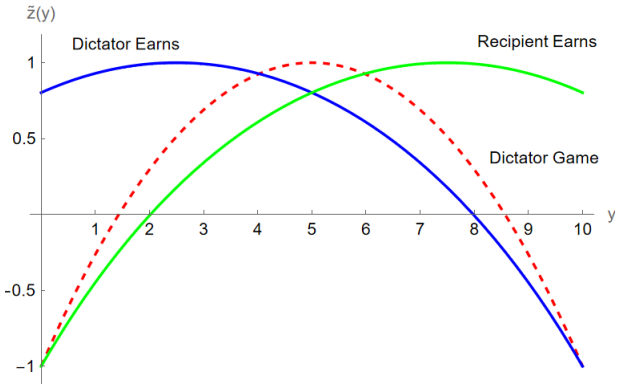
$$N(x) = \frac{1}{2}v(w(e_1, 0) - t(e_1, 0) - c(e_1)) + \frac{1}{2}v(t(e_1, 0)). \tag{15}$$

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

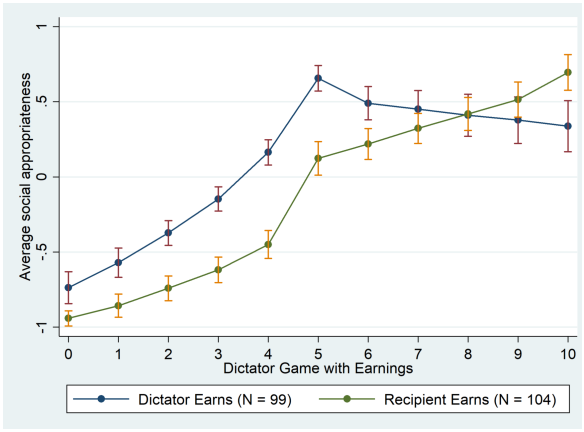
Therefore, in *Dictator Earns*, the most socially appropriate transfer is below the equal split. This implies that, conditional on the same endowment at $t = 2$, the most socially appropriate transfer in *Dictator Earns* is lower than in the standard dictator game. Similar arguments can be applied to derive the injunctive norm in *Recipient Earns*.

Experimental details Participants evaluated a situation in which either Person A or Person B 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 $y \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$€ to Person B at $t = 2$.[30]

Testable predictions The theory gives four predictions (see Figure 12a). First, giving less than the equal split is more socially appropriate in *Dictator Earns* than in *Recipient Earns*. Second, giving more than the equal split is more socially appropriate in *Recipient Earns* than in *Dictator Earns*. Third, the most socially appropriate transfer in *Recipient Earns* is above the equal split. Finally, the most socially appropriate transfer in *Dictator Earns* is below the equal split.



a) Normalized injunctive norms predicted.



b) Average social appropriateness.

Figure 12: Dictator game with earnings.

Results Figure 12b shows support for the first three predictions of the theory. First, giving $y = 0$, $y = 1$ or $y = 2$ 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 $y = 9$ or $y = 10$ is significantly more socially appropriate in *Recipient Earns* than in *Dictator Earns* ($p = 0.0788$ and $p = 0.0003$, respectively), while this is not the case for $y = 8$ ($p = 0.45$). Third, the most socially appropriate transfer in *Recipient Earns* is strictly above the equal split ($p < 0.0001$).

On the other hand, I do not find evidence that the most socially appropriate transfer in *Dictator Earns* is strictly below the equal split. This could be explained by the elicitation method's reliance on coordination at

[30] Ellingsen and Mohlin (2023) elicit the *Dictator Earns* condition with similar qualitative patterns. Additionally, Kassas and Palma (2019) and Bašić and Verrina (2021) elicit a modified dictator game where individuals' roles are decided with a contest.

focal points. In the experiment, it was not specified how costly it was for participants to solve 20 tables. This may make it difficult for participants to determine which transfer below the equal split others would perceive as most socially appropriate. On the other hand, the equal split has been shown to be a strong fairness norm (Andreoni and Bernheim (2009)).

5.3.3 Dictator game with joint production

In the dictator game with joint production, both individuals work to generate the endowment, and they may have different productivities. As shown in Proposition 6, the theory can account for the meritocratic divisions implemented by third parties in experiments using dictator games with joint production. I illustrate this in the setting proposed in Konow (2000).

In Konow (2000), individuals are matched in pairs and fold letters for mailing. Each individual generates a production of $q_i = p_i \cdot l_i$, 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 production between two individuals who exerted the same effort are due to their assigned piece-rates. Under this assumption, the production of the pair is $w({e_1}^i, {e_2}^j) = p_i l({e_1}^i) + p_j l({e_2}^j)$. Thus, the social appropriateness of choosing strategy $x = (e_1, e_2, t(e_1, e_2))$ is given by

$$N(x) = \frac{1}{2}v(p_i l(e_1) + p_j l(e_2) - t(e_1, e_2) - c(e_1)) + \frac{1}{2}v(t(e_1, e_2) - c(e_2)). \tag{16}$$

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

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 the individuals. On the other hand, when individuals fold different amounts of letters, the most socially appropriate division assigns the individual with the most letters produced a higher share of the total production.

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.[31] Figure 13 shows the divisions of third parties when individuals have the same (left) or different (right) piece-rates.

[31] 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 $N(t)$.

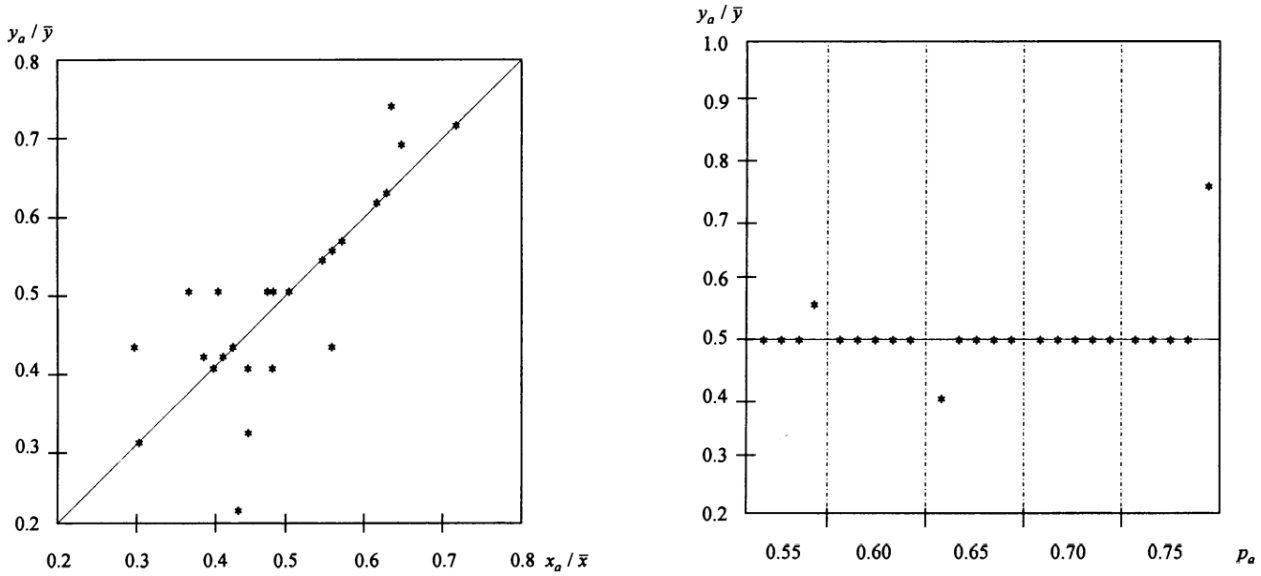


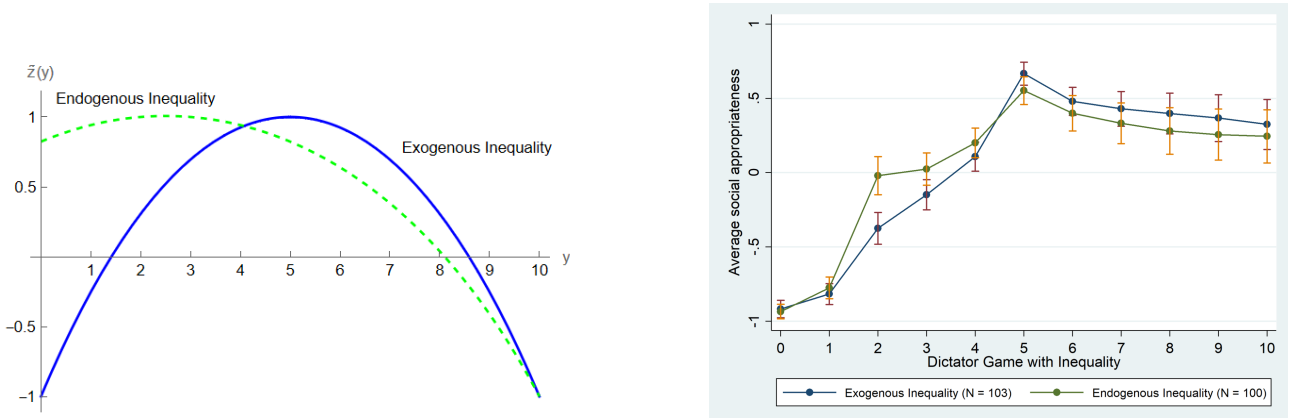
Figure 13: Left: Condition with the same piece-rate. x-axis: Share of letters produced by individual A. y-axis: Share of the total production assigned to individual A. Right: Condition with different piece-rates. x-axis: Piece-rate assigned to individual A (with (i) $p_A + p_B = 1$ and (ii) $p_A > p_B$). y-axis: Share of the total production assigned to individual A. Extracted from Konow (2000).

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.

I also test the theory's predictions with the norms elicited in the lab.

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*, individuals exerted the same effort, but Person A was assigned a higher piece-rate. In *Endogenous Inequality*, individuals are assigned the same piece-rate, but Person A has exerted more effort. In both variants, Person A contributes 8€ to the joint endowment while Person B contributes 2€.

Testable predictions The theory has four predictions (see Figure 14a): First, transferring a low amount is more socially appropriate in *Endogenous Inequality* than in *Exogenous Inequality*. Second, transferring an intermediate and high amount is more socially appropriate in *Exogenous Inequality* than in *Endogenous Inequality*. Third, in *Exogenous Inequality*, the most socially appropriate transfer is the equal split. Finally, in *Endogenous Inequality*, the most socially appropriate transfer is below the equal split.



a) Normalized injunctive norms predicted.

b) Average social appropriateness.

Figure 14: Dictator game with joint production.

Results Figure 14b shows some support for the first three predictions of the theory. First, giving $y = 2$ is significantly more socially appropriate in *Endogenous Inequality* than in *Exogenous Inequality* ($p < 0.001$). However, this is not the case for $y = 0$ ($p = 0.67$) and $y = 1$ ($p = 0.20$). Second, giving $y \geqslant 5$ is always more socially appropriate in *Exogenous Inequality* than in *Endogenous Inequality*. However, this difference is only statistically significant in the case of $y = 5$ ($p = 0.03$). Third, for *Exogenous Inequality*, the most socially appropriate transfer is the equal split ($p < 0.001$).

On the other hand, I do not find evidence that the most socially appropriate transfer in *Endogenous Inequality* is below the equal split. One possible explanation of this result is that in contrast with the third-party case in Konow (2000), Person A is both working and dividing the money. Thus, Person A may be perceived as generous, giving more than he should from a meritocratic perspective. The extended version of the theory in Section 6 partially addresses this motivation.

5.4 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 5.4.1, I show that a universalization statement was the most relevant for participants to justify their evaluations. In Section 5.4.2, I demonstrate that individual's degrees of Kantian morality and norm-following are positively correlated.

5.4.1 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 3, I assumed 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 15 shows that the participants deemed the universalization statement most relevant to justify their decisions in the experiment.[32]

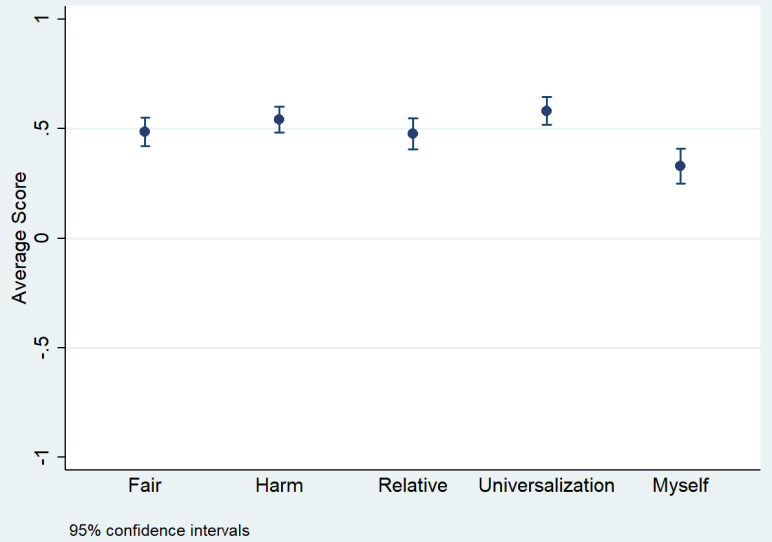


Figure 15: Average score of the five statements.

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

[32]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.

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.[33]

5.4.2 Correlation between Kantian morality and norm-following

Here, I exhibit the validity of the normalization conducted in (17) by showing that γ_i and κ_i are positively correlated. I measure the participant's degree of norm-following with the task introduced in Kimbrough and Vostroknutov (2018).[34] 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 Utilitarian Scale (Kahane et al. (2018)) to measure participants' degree of Kantian morality. 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 Kantian morality 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). This follows the spirit of Kant's categorical imperative, and, for instance, his famous quote: "But a lie is a lie, and in itself intrinsically evil, whether it be told with good or bad intents" (Kant (1797), Kant (1963)).

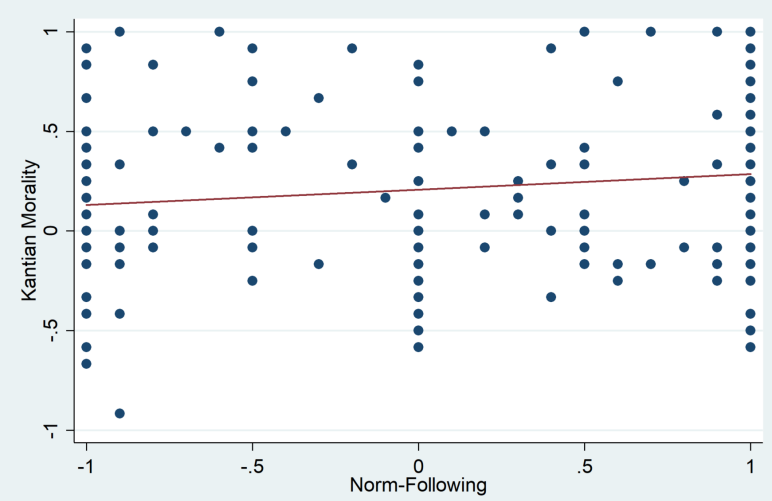


Figure 16: Relationship between norm-following and Kantian morality.

[33] 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.

[34] For other studies using this task, see Kimbrough and Vostroknutov (2018), Schneeberger and Krupka (2021) and Panizza et al. (2021)

Figure 16 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 ($r = 0.1533$, $p = 0.029$).[35]

6 The extended theoretical framework

In this section, I extend the theoretical framework proposed in Section 3. 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 3 does not explain the following three observations:

- **Standard dictator game**: Any transfer $t \in (\frac{w}{2}, w]$ is perceived as more socially appropriate than the complementary transfer $w - t$ (see Figure 8).
- **Linear public goods game**: There is a positive relationship between contributions to the public account and their perceived social appropriateness, even when these contributions are socially inefficient (see Figure 10b).
- **Volunteer's dilemma**: Volunteering with certainty is viewed as the most socially appropriate probability of volunteering (see Figure 11b).

To account for these observations, I extend the utility function in (2) by assuming that individuals' utility function combines universalization and social concerns (Alger et al. (2020)). More concretely, individual i's utility function when playing with individual j is given by

$$\begin{aligned} u_i(x_i, x_j) &= (1 - \kappa_i)\pi(x_i, x_j) \\ &- \alpha_i \max[\pi(x_j, x_i) - \pi(x_i, x_j), 0] \\ &- \beta_i \max[\pi(x_i, x_j) - \pi(x_j, x_i), 0] \\ &+ \kappa_i \pi(x_i, x_i). \end{aligned} \tag{17}$$

Here, $\pi(x_j, x_i)$ is the material payoff of the individual j under the strategy profile (x_i, x_j). Equation (17) is motivated by two main arguments. First, it nests several prominent preferences as special cases. For example, 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 $\alpha_i \geqslant \beta_i > 0$), spitefulness (i.e., $\kappa_i = 0$ and $\alpha_i = -\beta_i$ for a $\beta_i \in (-1, 0)$) and homo moralis (i.e., $\alpha_i = \beta_i = 0$

[35] 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).

and $\kappa_i \in (0,1]$). Second, Van Leeuwen and Alger (2024) use experimental data to structurally estimate (17) and find that it has a larger explanatory power than models with either social preferences or Kantian concerns.[36][37]

To obtain the extended injunctive norm, I first decompose (17) as

$$\begin{aligned} u_i(x_i,x_j) &= (1-\kappa_i)\pi(x_i,x_j) - (1-\tilde{\beta}_i)(\alpha_i \max[\pi(x_j,x_i)-\pi(x_i,x_j),0] + \beta_i \max[\pi(x_i,x_j)-\pi(x_j,x_i),0]) \\ &- \tilde{\beta}_i(\alpha_i \max[\pi(x_j,x_i)-\pi(x_i,x_j),0] + \beta_i \max[\pi(x_i,x_j)-\pi(x_j,x_i),0]) + \kappa_i\pi(x_i,x_i), \end{aligned} \tag{18}$$

where $\tilde{\beta}_i \in [0,1]$. Intuitively, a fraction $\tilde{\beta}_i$ of the social concerns is now embedded inside the injunctive norm. Building on the experimental evidence in Section 5, I assume that $\pi(x_j,x_i)$ enters positively inside the injunctive norm. This is equivalent to assuming $\alpha_i = -\beta_i$ with $\beta_i \in [0,1]$ in the second line of (18), which leads to:[38]

$$\begin{aligned} u_i(x_i,x_j) &= (1-\kappa_i-\tilde{\beta}_i\beta_i)\pi(x_i,x_j) - (1-\tilde{\beta}_i)(\alpha_i \max[\pi(x_j,x_i)-\pi(x_i,x_j),0] + \beta_i \max[\pi(x_i,x_j)-\pi(x_j,x_i),0]) \\ &+ \tilde{\beta}_i\beta_i\pi(x_j,x_i) + \kappa_i\pi(x_i,x_i). \end{aligned} \tag{19}$$

By dividing the expression in (19) by $1-\kappa_i-\tilde{\beta}_i\beta_i$ and defining:

- $\widetilde{V}(\pi(x_i,x_j)) \equiv \pi(x_i,x_j) - \frac{1-\tilde{\beta}_i}{1-\tilde{\beta}_i\beta_i-\kappa_i}(\alpha_i \max[\pi(x_j,x_i)-\pi(x_i,x_j),0] + \beta_i \max[\pi(x_i,x_j)-\pi(x_j,x_i),0])$,
- $\tilde{\gamma}_i \equiv \frac{1}{1-\tilde{\beta}_i\beta_i-\kappa_i}$,
- $\widetilde{N}(x_i,x_j) \equiv \tilde{\beta}_i\beta_i\pi(x_j,x_i) + \kappa_i\pi(x_i,x_i)$,

I obtain the following expression:

$$\widetilde{u}_i(x_i,x_j) = \widetilde{V}(\pi(x_i,x_j)) + \tilde{\gamma}_i\widetilde{N}(x_i,x_j). \tag{20}$$

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

$$\widetilde{N}(x_i,x_j) = (1-\tau_i)N(x_i) + \tau_i\pi(x_j,x_i). \tag{21}$$

Here, $\tau_i \in [0,1]$ represents the weight that individual i attaches to individual j's material payoff in evaluating the social appropriateness of selecting x_i. When $\tau_i = 0$, $\widetilde{N}(x_i,x_j)$ coincides with the universalization norm, while it coincides with $\pi(x_j,x_i)$ when $\tau_i = 1$. Therefore, the social appropriateness of a strategy is now the weighted sum of the universalization norm and the material payoff of the other individual. I refer to this latter concern as a *kindness* type of motive.[39]

[36] Alger et al. (2020) shows that (17) has strong evolutionary foundations, in the sense that evolution by natural selection favors preferences that not only combine self-interest and Kantian morality but also social concerns, when preferences are expressed at the level of (trivial) material payoffs. The main difference between Alger and Weibull (2013) and Alger et al. (2020) is that the former does not distinguish between reproductive success and material payoffs.

[37] In a related result, Carpenter and Robbett (2022) structurally estimate a model with both social preferences and conformity to injunctive norms (similar to (20)) finding that both motivations are necessary to account for individuals' behavior.

[38] Naturally, one could also make this assumption for the social concerns embedded inside individuals' material payoff. The extended injunctive norm proposed would be the same in both cases.

[39] 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).

In N-player interactions, I similarly define the *extended norm* as

$$\tilde{N}(x_i, \tilde{x}_{-i}) = (1 - \tau_i)N(x_i) + \tau_i g(x_i, \tilde{x}_{-i}), \tag{22}$$

with $g(x_i, \tilde{x}_{-i}) = \sum_{j \neq i} \pi_j(\tilde{x}_{-i}, x_i)$ being the *kindness norm* and $\pi_j(\tilde{x}_{-i}, x_i)$ denoting individual j's material payoff. Intuitively, $g(x_i, \tilde{x}_{-i})$ measures the impact of individual i's strategy on others' material payoff, which relates it to altruism (Becker (1976)) and efficiency concerns (Charness and Rabin (2002)).[40]

In Section 3, I assumed 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)). One possible explanation for this heterogeneity is that even if (22) is the correct specification, individuals may have incorrect beliefs about the weight other individuals attach to the two concerns. Indeed, a considerable amount of empirical evidence documents that misperceptions about others are widespread across individuals and domains (Bursztyn and Yang (2022)).

In the following sections, I compute $\tilde{N}(t, \tilde{t})$ in several interactions and study how it varies with τ. The notation of the interactions in the following sections is the same as the one used in Section 5.

6.2 Standard dictator game

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

$$\tilde{N}(t, \tilde{t}) = (1 - \tau) \underbrace{\frac{1}{2}(v(w - t) + v(t))}_{\text{Universalization norm}} + \tau \underbrace{\frac{1}{2}(v(t) + v(w - \tilde{t}))}_{\text{Kindness norm}}, \tag{23}$$

where $\tilde{t} \in [0, w]$ represents the transfer of the other individual. Note that $\frac{\partial \tilde{N}(t,\tilde{t})}{\partial t}$ does not depend on $\tilde{t}$ implying that $\tilde{t}$ does not affect the ranking of transfers prescribed by $\tilde{N}(t, \tilde{t})$. Proposition 7 characterizes the main properties of the extended injunctive norm in the standard dictator game.

Proposition 7. *Let $\hat{t}$ be such that $\left.\frac{\partial \tilde{N}(t,\tilde{t})}{\partial t}\right|_{t=\hat{t}} = 0$. Then, there exists $\overline{\tau} = 1 - \frac{v'(w)}{v'(0)} \in (0, 1)$ such that the most socially appropriate transfer in the dictator game is:*

$$t^* = \begin{cases} \hat{t} \in [\frac{w}{2}, w) & \text{if } \tau < \overline{\tau} \\ w & \text{if } \tau \geqslant \overline{\tau} \end{cases} \tag{24}$$

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

[40] Note that $g(x_i, \tilde{x}_{-i})$ (and consequentially $\tilde{N}(x_i, \tilde{x}_{-i})$) may vary with others' strategies as individual i's impact on others may depend on others' strategies.

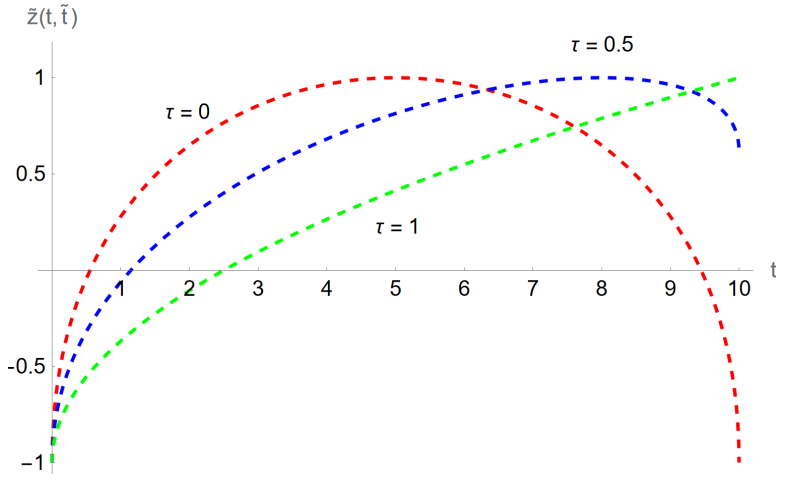


Figure 17: Normalized extended injunctive norm in the 10€ dictator game when $\tau = 0$ (dashed red), $\tau = 0.5$ (dashed blue) or $\tau = 1$ (dashed green).

Figure 17 displays the extended injunctive norm for different values of τ. When $t \in [0, \frac{w}{2})$, 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 $t \in (\frac{w}{2}, w]$, 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 $s_0 \geqslant \underline{s}$ and $\tau \in (0, 1]$, $t = \frac{w}{2}$ is (on average) the most socially appropriate transfer.*

Despite this being a quantitative question, this seems in line with the evidence from Figure 14, where the injunctive norm elicited in *Exogenous Inequality* has the equal split as the most socially appropriate transfer.[41] 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) evaluations that can not explained by the theory (17 participants) or (ii) a mix of the $\tau = 0$ and $\tau > 0$ cases (14 participants).[42]

In sum, 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.

[41] Note that the theory makes the same predictions for *Exogenous Inequality* and the standard dictator game. Therefore, Proposition 7 and Corollary 3 also apply to the former.

[42] Participants are classified as $\tau = 0$ if their elicited norm is represented by the red dashed line in Figure 17. 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 17). 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.

6.3 Linear public goods game

For simplicity, I restrict attention to strategy profiles $\tilde{x}_{-i} \in [0,w]^{n-1}$ with all other $n-1$ individuals contributing the same amount $y \in [0,w]$. In this case, the extended norm is given by

$$\widetilde{N}(t,\tilde{t}) = (1-\tau)\underbrace{v(w-t+\hat{A}nt)}_{\text{Universalization norm}} + \tau\underbrace{(n-1)v(w-y+(n-1)\hat{A}y+\hat{A}t)}_{\text{Kindness norm}}. \tag{25}$$

Note that the kindness norm always increases in t as contributing to the public account increases others' material payoff, even if doing so is socially inefficient. Proposition 8 characterizes the main properties of the extended injunctive norm in the linear public goods game.

Proposition 8. *Let $\hat{t}$ be such that $\left.\frac{\partial \tilde{N}(t,\tilde{t})}{\partial t}\right|_{t=\hat{t}} = 0$. 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**: $\hat{A}n \geqslant 1$ (Socially efficient case)*

$$t^* = w \qquad \forall \tau \in [0,1] \tag{26}$$

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

$$t^* = \begin{cases} 0 & \text{if } \tau \in [0,\underline{\tau}] \\ \hat{t} \in (0,w) & \text{if } \tau \in (\underline{\tau},\overline{\tau}) \\ w & \text{if } \tau \in [\overline{\tau},1] \end{cases} \tag{27}$$

Additionally, (i) $\frac{\partial \hat{t}}{\partial \tau} \geqslant 0$, (ii) when $t^* = 0$, $\frac{\partial \tilde{N}(t,\tilde{t})}{\partial t} < 0$ for $t \in (0,w]$, (iii) when $t^* = w$, $\frac{\partial \tilde{N}(t,\tilde{t})}{\partial t} > 0$ for $t \in [0,w)$, and (iv) when $t^* \in (0,w)$, $\frac{\partial \tilde{N}(t,\tilde{t})}{\partial t} > 0$ for $t \in [0,t^*)$ and $\frac{\partial \tilde{N}(t,\tilde{t})}{\partial t} < 0$ for $t \in (t^*,w]$.

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

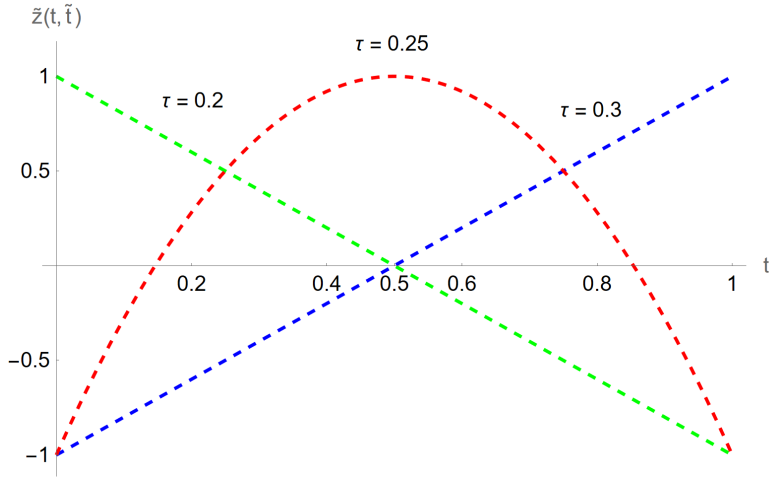


Figure 18: Normalized extended injunctive norm when $\hat{A} = 0.2$, $n = 4$, $w = 1$, $y = 0.5$, $v(\cdot) = \sqrt{(\cdot)}$, and $\tau = 0.2$ (dashed green), $\tau = 0.25$ (dashed red) or $\tau = 0.3$ (dashed blue).

Figure 18 displays the extended injunctive norm when $\hat{A} = 0.2$ and $n = 4$ 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 Conclusion

Recent studies have shown the explanatory power of injunctive norms in explaining individuals' behavior. However, past literature has been restricted to empirical settings where injunctive norms are elicited with the method proposed in Krupka and Weber (2013). In this paper, I introduce a theory of injunctive norms where individuals evaluate the social appropriateness of a given behavior based on a counterfactual scenario where all individuals in the population choose that behavior. I illustrate the theory's tractability and portability by deriving its predictions in a large set of interactions. Additionally, I show that the evidence from past studies and a new lab experiment supports a large number of the predictions of the theory. Finally, I extend the theory by incorporating social concerns into the injunctive norm. This provides a framework to study norm heterogeneity and helps to explain several results not accounted for by the simplest version of the theory.

One natural question is how to apply the theory outside the lab, where norms are elicited from real-life behaviors (e.g., Lane et al. (2023)). To do so, one could follow, for example, Muñoz Sobrado (2022) and Alger and Laslier (2022) which use homo moralis preferences 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 it with the corresponding elicitations. I leave this type of comparison for future research.

References

Abbink, K., L. Gangadharan, T. Handfield, and J. Thrasher (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.

Alesina, A. and E. La Ferrara (2005). Preferences for redistribution in the land of opportunities. *Journal of Public Economics 89*(5-6), 897–931.

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

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

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

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

Alger, I., J. W. Weibull, and L. Lehmann (2020). Evolution of preferences in structured populations: Genes, guns, and culture. *Journal of Economic Theory 185*, 104951.

Andre, P., T. Boneva, F. Chopra, and A. Falk (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*, 891–904.

Andreoni, J. and B. D. Bernheim (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica 77*(5), 1607–1636.

Arkes, H. R. and C. Blumer (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes 35*(1), 124–140.

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

Bašić, Z. and E. Verrina (2021). 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.

Benabou, R. and J. Tirole (2006). Belief in a just world and redistributive politics. *The Quarterly Journal of Economics 121*(2), 699–746.

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

Blackorby, C., W. Bossert, and D. Donaldson (2002). Utilitarianism and the theory of justice. *Handbook of Social Choice and Welfare 1*, 543–596.

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

Buchheit, S. and N. Feltovich (2011). Experimental evidence of a sunk-cost paradox: A study of pricing behavior in bertrand–edgeworth duopoly. *International Economic Review 52*(2), 317–347.

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

Bursztyn, L. and D. Y. Yang (2022). Misperceptions about others. *Annual Review of Economics 14*, 425–452.

Camerer, C. F. and E. Fehr (2004). Measuring social norms and preferences using experimental games: A guide for social scientists. *Foundations of Human Sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies 97*, 55–95.

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

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.

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

Carpenter, J. and A. Robbett (2022). Measuring Socially Appropriate Social Preferences. *Working Paper*.

Charness, G. and M. Dufwenberg (2006). Promises and partnership. *Econometrica 74*(6), 1579–1601.

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

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

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

Cialdini, R. B., R. R. Reno, and C. A. Kallgren (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.

Corneo, G. and H. P. Grüner (2002). Individual preferences for political redistribution. *Journal of Public Economics 83*(1), 83–107.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review 17*(3), 273–292.

Di Mauro, C. and M. Finocchiaro Castro (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 G. Kirchsteiger (2004). A theory of sequential reciprocity. *Games and Economic Behavior 47*(2), 268–298.

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

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

Ellingsen, T. and E. Mohlin (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.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics 133*(4), 1645–1692.

Fehr, E. and U. Fischbacher (2004). Social norms and human cooperation. *Trends in Cognitive Sciences 8*(4), 185–190.

Fehr, E. and S. Gächter (2000). Cooperation and punishment in public goods experiments. *American Economic Review 90*(4), 980–994.

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

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

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

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

Friedman, D., K. Pommerenke, R. Lukose, G. Milam, and B. A. Huberman (2007). Searching for the sunk cost fallacy. *Experimental Economics 10*(1), 79–104.

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

Greene, J. (2014). *Moral Tribes: Emotion, reason, and the gap between us and them*. Penguin.

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

Hillenbrand, A. and F. Winter (2018). Volunteering under population uncertainty. *Games and Economic Behavior 109*, 65–81.

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

Isaac, R. M. and J. M. Walker (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 E. Karagözoğlu (2024). Moral preferences in bargaining. *Economic Theory*, 1–24.

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.

Kandori, M. (1992). Social norms and community enforcement. *The Review of Economic Studies 59*(1), 63–80.

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

Kant, I. (1797). On a supposed right to lie from philanthropy. *Practical philosophy 612*.

Kant, I. (1963). Lectures on Ethics. *Indianapolis, IN: Hackett*.

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

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

Kettrey, H. H. and R. A. Marx (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. 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.

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

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

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

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

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

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

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

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

Lefgren, L. J., D. P. Sims, and O. B. Stoddard (2016). Effort, luck, and voting for redistribution. *Journal of Public Economics 143*, 89–97.

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.

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

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

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

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.

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

Nichols, S. and R. Mallon (2006). Moral dilemmas and moral rules. *Cognition 100*(3), 530–542.

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

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*.

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

Rawls, J. (1971). A Theory of Justice. *Belknap Press/Harvard University Press*.

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

Roemer, J. E. (2015). Kantian Optimization: A microfoundation for cooperation. *Journal of Public Economics 127*, 45–57.

Ronayne, D., D. Sgroi, and A. Tuckwell (2021). Evaluating the sunk cost effect. *Journal of Economic Behavior & Organization 186*, 318–327.

Rydval, O. and A. Ortmann (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.

Schneeberger, A. and E. L. Krupka (2021). Determinants of Norm Compliance: Moral Similarity and Group Identification. *Available at SSRN 3969227*.

Schultz, P. W., J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science 18*(5), 429–434.

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

Van Leeuwen, B. and I. Alger (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.

Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics 7*(1), 359–387.

8 Appendix A: Other interactions.

In this section, I consider interactions not included in the main text. More concretely, I study the trust game (Section 8.1), the ultimatum game (Section 8.2), the dictator game with different prices of giving (Section 8.3), the dictator game with deserving recipient (Section 8.4), and the linear public goods game with heterogeneous endowments (Section 8.5). Additionally, I consider the volunteer's dilemma with the extended norm in Section 8.6.

8.1 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' material payoff is $w - x + r$, while trustees' material 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

$$N(t) = \frac{1}{2}v(w - t_1 + f(t_2)) + \frac{1}{2}v(w + mt_1 - f(t_2)). \tag{27}$$

Note that t_1 and $f(t_2)$ can be interpreted as measures of efficiency and equality. The most socially appropriate strategy $t^* = (t_1{}^*, f^*(t_2))$ consists of $t_1{}^* = w$ and $f^*(t_2) = \frac{(m+1)t_2}{2}$. Note that trustees' decision is analogous to dictators' decision with $w = (m+1)t_2$. On the other hand, given that $f^*(t_2) = \frac{(m+1)t_2}{2}$, the most socially appropriate amount sent is $t_1{}^* = w$.

When trustors' endowments (i.e., w_R) are larger than trustees' endowments (i.e., w_P), the injunctive norm is given by

$$N(t) = \frac{1}{2}v(w_R - t_1 + f(t_2)) + \frac{1}{2}v(w_P + mt_1 - f(t_2)). \tag{28}$$

As before, the most socially appropriate return function $f^*(t_2)$ 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 $t_1^* = w_R$ and $f^*(t_2) = max[g^*(t_2), 0]$, where $g^*(t_2) \equiv \frac{w_P - w_R}{2} + \frac{(m+1)t_2}{2}$.

Finally, note that $f^*(t_2) = \frac{(m+1)t_2}{2}$ is increasing in m, as the higher m, the higher the inequality generated by t_2. On the other hand, $t_1^* = w$ if $m > 1$ and $t_1^* = 0$ if $m < 1$.

8.2 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

$$\pi(x, y) = v(0) + \frac{1}{2} \cdot \mathbf{1}_{\{x_1 \geqslant y_2\}} \cdot [v(w - x_1) - v(0)] + \frac{1}{2} \cdot \mathbf{1}_{\{y_1 \geqslant x_2\}} \cdot [v(y_1) - v(0)]. \tag{29}$$

Thus, the social appropriateness of strategy $t = (t_1, t_2)$ is:

$$N(t) = v(0) + \frac{1}{2} \cdot \mathbf{1}_{\{t_1 \geqslant t_2\}} \cdot [v(t_1) + v(w - t_1) - 2v(0)]. \tag{30}$$

As $v(t_1) + v(w - t_1) - 2v(0) \geqslant 0$, $N(t)$ is maximized when ${t_1}^* = \frac{w}{2}$ and ${t_2}^* \in [0, \frac{w}{2}]$. 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.

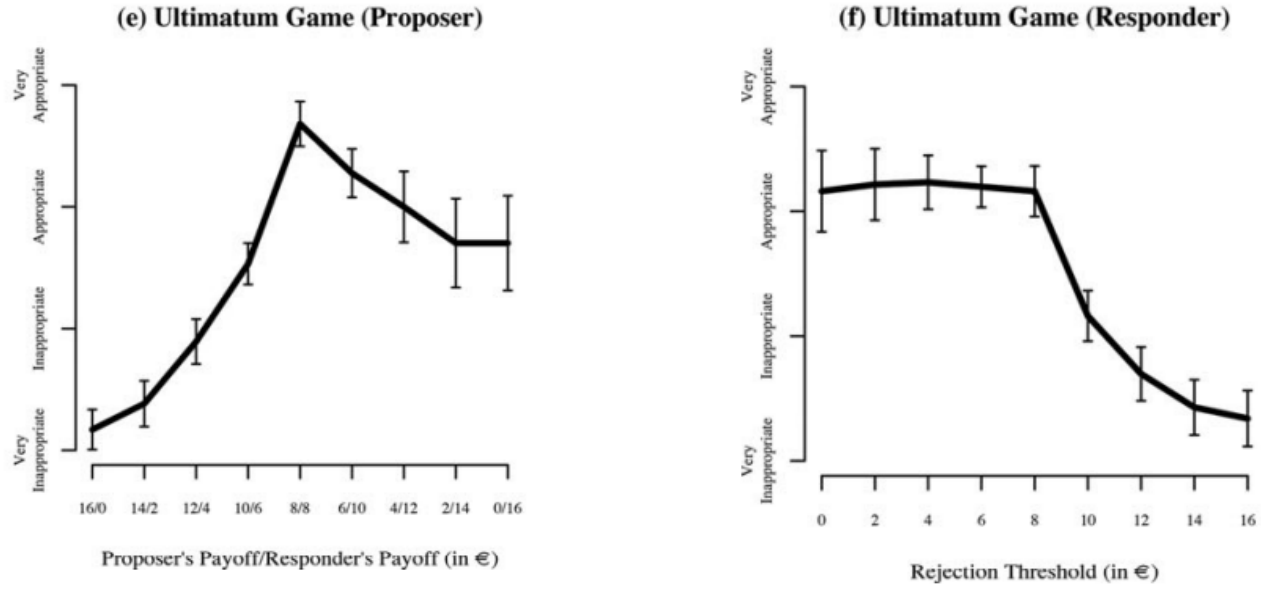


Figure 19: Injunctive norms elicited in the ultimatum game (Kimbrough and Vostroknutov (2016)).

Figure 19 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.

8.3 Dictator game with different prices of giving

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

$$N(t) = \frac{1}{2} v(w - t) + \frac{1}{2} v(mt). \tag{31}$$

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., $RRA(t) = \frac{-tv''(t)}{v'(t)}$).

Proposition 9. *Let $RRA(t) = \frac{-tv''(t)}{v'(t)}$ denote the relative risk aversion of function v at t. The most socially appropriate transfer t^* satisfies:*

$$\frac{\partial t^*}{\partial m} \begin{cases} > 0 & \text{if } RRA(mt^*) < 1 \\ = 0 & \text{if } RRA(mt^*) = 1 \\ < 0 & \text{if } RRA(mt^*) > 1 \end{cases} \tag{32}$$

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

To exemplify Proposition 10, let $v(c) = \frac{c^{(1-\rho)}-1}{1-\rho}$, where ρ represents the (constant) relative risk aversion of v. In this case, the injunctive norm is given by

$$N(t) = \frac{1}{2}\frac{(w-t)^{(1-\rho)}-1}{1-\rho} + \frac{1}{2}\frac{(mt)^{(1-\rho)}-1}{1-\rho}, \tag{33}$$

which leads to

$$t^* = \frac{m^{\frac{1}{\rho}}}{m + m^{\frac{1}{\rho}}} w \tag{34}$$

$$\frac{\partial t^*}{\partial m} = -\frac{\rho - 1}{\rho}\frac{w m^{\frac{1}{\rho}}}{(m + t^{\frac{1}{\rho}})^2}. \tag{35}$$

Thus, the sign of $\frac{\partial t^*}{\partial m}$ 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 $\frac{w}{2}$.

8.4 Dictator Game with Deserving Recipient

Several studies have shown that dictators' transfers depend on recipients' income and social status.[39] An explanation for these results is that transferring a low amount is more socially inappropriate when the recipient is deserving. To capture this motivation, I assume that individuals consider their wealth outside the lab. For example, if recipients are from a poorer country, dictators (from a rich country) can infer that their wealth is higher than that of the recipients. This leads to a modified injunctive norm that depends on the wealth of the two individuals.

$$N(t) = \frac{1}{2}v(w_d + w - t) + \frac{1}{2}v(w_r + t), \tag{36}$$

[39] For example, Brañas-Garza (2006) finds that dictators transfer two-thirds of their endowment when they are told that their recipient is poor, while they give ten percent in the standard no-information condition. Eckel and Grossman (1996) show that dictators' transfers increase when the recipient is a charity. Korenok et al. (2008) find that dictators' transfers are decreasing in recipients' endowments.

where w_d and w_r represent the dictator's and recipient's wealth outside the lab. The most socially appropriate transfer can be characterized as follows:

$$t^* = \begin{cases} 0 & \text{if } w_d - w_r < -w \\ \frac{w}{2} + \frac{w_d - w_r}{2} & \text{if } w_d - w_r \in [-w, w] \\ w & \text{if } w_d - w_r > w \end{cases} \tag{37}$$

Three remarks are important to emphasize. First, when both individuals have the same wealth, the most socially appropriate transfer is the equal split. This is the case when individuals are students at the same university. Second, when dictators have higher wealth, the most socially appropriate transfer is above the equal split and increases in the wealth difference between the individuals. Finally, when dictators have lower wealth, the most socially appropriate transfer is below the equal split and decreases in the wealth difference between the individuals.

8.5 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 of heterogeneous endowments (Cherry et al. (2005) and 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) and Kingsley (2016)).

For simplicity, I consider the case with two individuals and two endowment levels (i.e., $w_H > w_L > 0$). The results for a larger number of individuals and income levels follow without complication. 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 he has a high (resp. low) endowment. The individual 1's (expected) material payoff is given by

$$\pi(x_1, x_2) = \frac{1}{2} \underbrace{v(w_H - x_1^H + \hat{A}(x_1^H + x_2^L))}_{\text{Individual 1 has high endowment}} + \frac{1}{2} \underbrace{v(w_L - x_1^L + \hat{A}(x_1^L + x_2^H))}_{\text{Individual 1 has low endowment}}. \tag{38}$$

Therefore, the injunctive norm is the following:

$$N(t) = \frac{1}{2} v(w_H - t_H + \hat{A}(t_H + t_L)) + \frac{1}{2} v(w_L - t_L + \hat{A}(t_L + t_H)). \tag{39}$$

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

8.6 Voluntary game with extended norm

As in the linear public goods game, I restrict attention to strategy profiles $\tilde{x}_{-i} \in [0,1]^{n-1}$ with all other $n-1$ individuals volunteering the same probability $y \in [0,1]$. In this case, the extended norm is given by

$$\widetilde{N}(t, \tilde{t}) = (1-\tau) \underbrace{v(b(1-(1-t)^n) - ct)}_{\text{Universalization norm}} + \tau \underbrace{(n-1)v(b(1-(1-y)^{n-1}) + bt(1-y)^{n-1} - cy)}_{\text{Kindness norm}}. \tag{40}$$

As in the linear public goods game, the kindness norm is always increasing in t as volunteering always increases others' (expected) material payoff, although it may be socially inefficient.

Proposition 10. *Let $\hat{t}$ be such that $\frac{\partial \tilde{N}(t,\tilde{t})}{\partial t}\Big|_{t=\hat{t}} = 0$. Then, there exists $\overline{\tau} \in (0,1]$ such that the most socially appropriate volunteering probability in the volunteer's dilemma is:*

$$t^* = \begin{cases} \hat{t} \in [1 - (\frac{c}{bn})^{\frac{1}{n-1}}, 1) & \text{if } \tau < \overline{\tau} \\ 1 & \text{if } \tau \geqslant \overline{\tau} \end{cases} \tag{41}$$

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

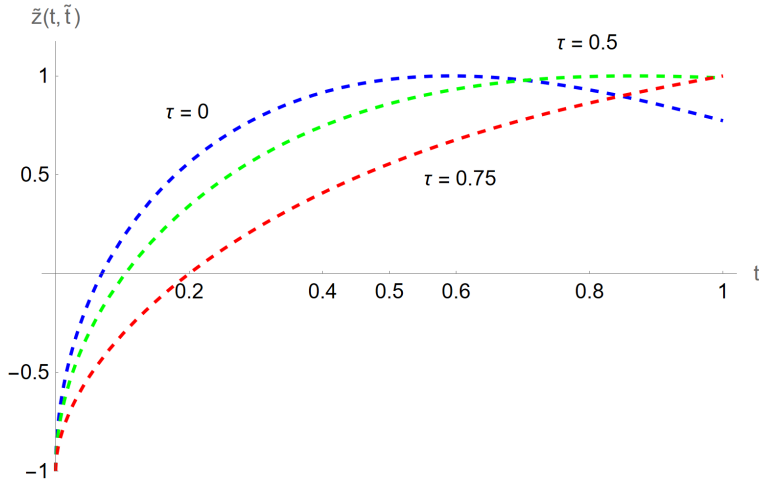


Figure 20: Normalized injunctive norms when $n = 3$, $b = 10$, $c = 5$, $y = 0.5$, $v(\cdot) = \sqrt{(\cdot)}$, and $\tau = 0$ (dashed blue), $\tau = 0.5$ (dashed green) or $\tau = 0.8$ (dashed red).

Figure 20 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.

9 Appendix B: Mathematical proofs.

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

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

$$\frac{\partial N(t)}{\partial t} = -\frac{1}{2}v'(w - t) + \frac{1}{2}v'(t), \tag{42}$$

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

Proof. Proposition 3: I distinguish between three cases: (i) $N^A(t^*(A)) > N^B(x') > N^A(t_*(A))$, (ii) $N^B(x') > N^A(t^*(A))$ and (iii) $N^B(x') < N^A(t_*(A))$.

In (i), x' is neither the most nor the least socially appropriate strategy. Then, $t^*(A) = t^*(B)$ and $t_*(A) = t_*(B)$, which implies that $\tilde{z}^A(x) = \tilde{z}^B(x)$ for all $x \in X$. To see this, it is sufficient to note that $\tilde{z}^g(x)$ only depends on $N(t^*(g))$, $N(t_*(g))$ and $N(x)$. Finally, $\tilde{z}^B(x') \in (-1, 1)$ when $N^A(t^*(A)) > N^B(x') > N^A(t_*(A))$ and $\tilde{z}^B(x') = 1$ (resp. $\tilde{z}^B(x') = -1$) when $N^B(x') = N^A(t^*(A))$ (resp. $N^B(x') = N^A(t_*(A))$).

In (ii), x' is the least appropriate strategy of B. Thus, $N(t^*(A)) = N(t^*(B))$, $N(t_*(A)) > N(x')$, $\tilde{z}^A(t^*(A)) = \tilde{z}^B(t^*(B)) = 1$ and $\tilde{z}^A(t_*(A)) = \tilde{z}^B(x') = -1$. On the other hand, $\tilde{z}^A(x) < \tilde{z}^B(x)$ for all $x \in X$ with $\tilde{z}^g(x) < 1$. To see this, let $N(x') = N(t_*(A)) - k$ with $k > 0$. Then,

$$\begin{aligned} \tilde{z}^B(x) &= 2\frac{N(x) - N(x')}{N(t^*(B)) - N(x')} - 1 \\ &= 2\frac{N(x) - (N(t_*(A)) - k)}{N(t^*(A)) - (N(t_*(A)) - k))} - 1 \\ &= 2\frac{N(x) - N(t_*(A)) + k}{N(t^*(A)) - N(t_*(A)) + k} - 1 \\ &> \tilde{z}^A(x) \end{aligned} \tag{43}$$

for any $k > 0$.

In (iii), x' is the most appropriate strategy of B. Thus, $N(t^*(A)) < N(x')$, $N(t_*(A)) = N(t_*(B))$, $\tilde{z}^A(t^*(A)) = \tilde{z}^B(x') = 1$ and $\tilde{z}^A(t_*(A)) = \tilde{z}^B(t_*(B)) = -1$. On the other hand, $\tilde{z}^A(x) > \tilde{z}^B(x)$ for all $x \in X$ with $\tilde{z}^g(x) > -1$. To see this, let $N(x') = N(t^*(A)) + k$ with $k > 0$. Then,

$$\begin{aligned} \tilde{z}^B(x) &= 2\frac{N(x) - N(t_*(B))}{N(x') - N(t_*(B))} - 1 \\ &= 2\frac{N(x) - N(t_*(A))}{N(t^*(A)) + k - N(t_*(A))} - 1 \\ &< \tilde{z}^A(x) \end{aligned} \tag{44}$$

for any $k > 0$. □

Proof. Proposition 4: As $v' > 0$ and $w - t + \hat{A}nt$ is linear in t, t^* can be computed by determining for which values of $\hat{A}n$ is $w - t + \hat{A}nt$ increasing in t. Thus, $t^* = w$ ($w - t + \hat{A}nt$ is strictly increasing in t) when $\hat{A}n > 1$, $t^* = [0, w]$ ($w - t + \hat{A}nt$ is constant in t) when $\hat{A}n = 1$, and $t^* = 0$ ($w - t + \hat{A}nt$ is strictly decreasing in t) when $\hat{A}n < 1$.

□

Proof. Proposition 5: As $N(t)$ is a strictly concave function, I compute t^* by taking the first order condition of $N(t)$ with respect to t.

$$\frac{\partial N(t)}{\partial t} = v'(b(1 - (1 - t)^n) - ct)(bn(1 - t)^{n-1} - c), \tag{45}$$

which implies that

$$t^* = 1 - (\frac{c}{bn})^{\frac{1}{n-1}}. \tag{46}$$

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

$$\frac{\partial t^*}{\partial n} = \frac{(\frac{b}{vn})^{\frac{1}{n-1}}(n - 1 + n\log(\frac{c}{bn}))}{n(n-1)^2}. \tag{47}$$

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

$$n[\log(c)-\log(b)]-n[\log(n)-1]-1. \tag{48}$$

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

□

Proof. Proposition 6: The injunctive norm is given by

$$N(x)=\frac{1}{2}v(w(e_1,e_2)-t(e_1,e_2)-c(e_1))+\frac{1}{2}v(t(e_1,e_2)-c(e_2)), \tag{49}$$

where $x=(e_1,e_2,t(e_1,e_2))$, $e_1 \geqslant 0$, $e_2 \geqslant 0$ and $t(e_1,e_2) \in [0,w(e_1,e_2)]$. Differentiating $N(x)$ with respect to e_1, e_2 and $t(e_1,e_2)$ gives $t^*(e_1,e_2)=\frac{w(e_1,e_2)}{2}-\frac{c(e_1)-c(e_2)}{2}$. □

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

$$\tilde{N}(t,\tilde{t})=\frac{1}{2}v(t)+\frac{1}{2}(1-\tau)v(w-t)+\tau\frac{1}{2}v(w-\tilde{t}), \tag{50}$$

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

$$\frac{1}{2}v'(\hat{t})-\frac{1}{2}(1-\tau)v(w-\hat{t})=0. \tag{51}$$

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

$$\frac{\partial \hat{t}}{\partial \tau}=\frac{-\frac{1}{2}v'(w-\hat{t})}{\frac{1}{2}v''(\hat{t})+\frac{1}{2}(1-\tau)v''(w-\hat{t})} \geqslant 0, \tag{52}$$

as both numerator and denominator are negative. This implies that $\hat{t} \in [\frac{w}{2},w)$ (as $t^*=\frac{w}{2}$ when $\tau=0$). On the other hand, a sufficient condition for having $t^*=w$ is that $\left.\frac{\partial \tilde{N}(t,\tilde{t})}{\partial t}\right|_{t=w} \geqslant 0$, as if the derivative at $t=w$ is non-negative, it means that the derivative at any $t<w$ is also non-negative (for the strict concavity of $\tilde{N}(t,\tilde{t})$). This implies that (51) can not be satisfied. This condition can be expressed as:

$$v'(w)-(1-\tau)v'(0) \geqslant 0, \tag{53}$$

which is equivalent to

$$\tau \geqslant 1-\frac{v'(w)}{v'(0)} \equiv \overline{\tau}. \tag{54}$$

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

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

$$\tilde{N}(t,\tilde{t})=(1-\tau)v(w-t+\hat{A}nt)+\tau(n-1)v(w-y+(n-1)\hat{A}y+\hat{A}t). \tag{55}$$

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

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

$$(1-\tau)v'(w-\hat{t}+\hat{A}n\hat{t})(-1+\hat{A}n)+\tau(n-1)v'(w-y+(n-1)\hat{A}y+\hat{A}\hat{t})\hat{A}=0. \tag{56}$$

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

$$\frac{\partial\hat{t}}{\partial\tau}=\frac{v'(w-\hat{t}+\hat{A}n\hat{t})(\hat{A}n-1)-v'(w-y+(n-1\hat{A}y+\hat{A}\hat{t}))\hat{A}(n-1)}{v''(w-\hat{t}+\hat{A}n\hat{t})(\hat{A}n-1)^2(1-\tau)+v''(w-y+(n-1\hat{A}y+\hat{A}\hat{t}))\hat{A}^2(n-1)\tau}\geqslant 0, \tag{57}$$

as both numerator and denominator are negative. On the other hand, the sufficient conditions for having $t^*=0$ and $t^*=w$ are (i) $\left.\frac{\partial\tilde{N}(t,\tilde{t})}{\partial t}\right|_{t=0}\leqslant 0$ and (ii) $\left.\frac{\partial\tilde{N}(t,\tilde{t})}{\partial t}\right|_{t=w}\geqslant 0$. In the first case $\tilde{N}(t,\tilde{t})$ is decreasing in t and $t^*=0$, while in the second case $\tilde{N}(t,\tilde{t})$ is increasing in t and $t^*=w$.

The first condition can be expressed as:

$$(1-\tau)v'(w)(-1+\hat{A}n)+\tau(n-1)v'(w-y+(n-1)\hat{A}y)\hat{A}\leqslant 0, \tag{58}$$

which is equivalent to:

$$\tau\leqslant\frac{v'(w)(1-\hat{A}n)}{v'(w)(1-\hat{A}n)+(n-1)\hat{A}v'(w-y+(n-1)\hat{A}y)}\equiv\underline{\tau}. \tag{59}$$

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

$$(1-\tau)v'(\hat{A}nw)(-1+\hat{A}n)+\tau(n-1)v'(w-y+(n-1)\hat{A}y+\hat{A}w)\hat{A}\geqslant 0, \tag{60}$$

which is equivalent to:

$$\tau\geqslant\frac{v'(\hat{A}nw)(1-\hat{A}n)}{v'(\hat{A}nw)(1-\hat{A}n)+(n-1)\hat{A}v'(w-y+(n-1)\hat{A}y+\hat{A}w)}\equiv\overline{\tau}. \tag{61}$$

In this case $v'(\hat{A}nw)(1-\hat{A}n)$ and $(n-1)\hat{A}v'(w-y+(n-1)\hat{A}y+\hat{A}w)>0$ 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 $B\equiv(n-1)\hat{A}v'(w-y+(n-1)\hat{A}y+\hat{A}w)$ and $\underline{\tau}=\frac{C}{C+D}$ with $C\equiv v'(w)(1-\hat{A}n)$ and $D\equiv(n-1)\hat{A}v'(w-y+(n-1)\hat{A}y)$. 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

$$N(t)=\frac{1}{2}v(w-t)+\frac{1}{2}v(mt), \tag{62}$$

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

$$\frac{1}{2}v'(w-t^*)-\frac{1}{2}v'(mt^*)m=0. \tag{63}$$

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

$$-v''(w-t^*)\frac{\partial t^*}{\partial m}-v''(t^*m)(t^*+\frac{\partial t^*}{\partial m}m)m-v'(mt^*)=0, \tag{64}$$

which gives

$$\frac{\partial t^*}{\partial m} = \frac{-(mt^* v''(mt^*) + v'(mt^*))}{(m^2 v''(mt^*) + v''(w - t^*))}. \tag{65}$$

As $v'' < 0$, the sign of $\frac{\partial t^*}{\partial m}$ depends on the sign of $mt^* v''(mt^*) + v'(mt^*)$. Using simple algebra, I obtain the result:

$$mt^* v''(mt^*) + v'(mt^*) = \tag{66}$$

$$= \frac{v'(mt^*)}{v'(mt^*)}(v''(mt^*)mt^* + v'(mt^*)) = \tag{67}$$

$$= v'(mt^*)(1 + \frac{mt^* v''(mt^*)}{v'(mt^*)}) = \tag{68}$$

$$= v'(mt^*)(1 - RRA(mt^*)), \tag{69}$$

where $RRA(mt^*) \equiv -\frac{mt^* v''(mt^*)}{v'(mt^*)}$. Therefore, t^* is increasing in m when $RRA(mt^*) < 1$, decreasing in m when $RRA(mt^*) > 1$ and constant in m when $RRA(mt^*) = 1$. □

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

$$\begin{aligned} \tilde{N}(t, \tilde{t}) &= (1-\tau)v(b(1-(1-t)^n) - ct) \\ &+ \tau(n-1)v(b(1-(1-y)^{n-1}) + bt(1-y)^{n-1} - cy). \end{aligned} \tag{70}$$

First, I consider three polar cases. When $\tau = 0$ or when $y = 1$, $t^* = 1 - (\frac{c}{bn})^{\frac{1}{n-1}}$ (see Proposition 5), while when $\tau = 1$, $t^* = 1$.

When $\tau \in (0,1)$ and $y \in [0,1)$, $\tilde{N}(t, \tilde{t})$ is strictly concave in t. Thus, t^* can be interior (i.e., $t \in (0,1)$) or in the boundary (i.e., $t = 1$). When $t^* = \hat{t} \in (0,1)$, $\hat{t}$ satisfies $\left.\frac{\partial \tilde{N}(t,\tilde{t})}{\partial t}\right|_{t=\hat{t}} = 0$, or equivalently

$$\begin{aligned} &(1-\tau)v'(b(1-(1-\hat{t})^n) - c\hat{t})(nb(1-\hat{t})^{n-1} - c) \\ + \quad &\tau(n-1)v'(b(1-(1-\hat{t})(1-y)^{n-1}) - cy)b((1-y)^{n-1}) = 0. \end{aligned} \tag{71}$$

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

$$\frac{\partial \hat{t}}{\partial \tau} = \frac{v'(b(1-(1-\hat{t})^n)-c\hat{t}) - (n-1)b(1-y)^{n-1}v'(b(1-(1-\hat{t})(1-y)^{n-1})-cy)}{(1-\tau)v''(b(1-(1-\hat{t})^n)-ct)(nb(1-\hat{t})^{n-1}-c)^2 - (1-\tau)bnv'(b(1-(1-\hat{t})^n)-c\hat{t}) + \tau(n-1)v'(b(1-(1-\hat{t})(1-y)^{n-1})-cy)(b\hat{t}(1-y)^{n-1})^2}. \tag{72}$$

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

$$v'(b(1-(1-\hat{t})^n) - c\hat{t}) - (n-1)b(1-y)^{n-1}v'(b(1-(1-\hat{t})(1-y)^{n-1}) - cy). \tag{73}$$

As $\hat{t}$ is an interior solution, (71) must be satisfied, implying that

$$\begin{aligned} &-\frac{\tau}{1-\tau}(n-1)b(1-y)^{n-1}v'(b(1-(1-\hat{t})(1-y)^{n-1}) - cy) \\ &-(n-1)b(1-y)^{n-1}v'(b(1-(1-\hat{t})(1-y)^{n-1}) - cy) \leqslant 0. \end{aligned} \tag{74}$$

Thus, $\frac{\partial \hat{t}}{\partial \tau} \geqslant 0$, which implies that $\hat{t} \in [1 - (\frac{c}{bn})^{\frac{1}{n-1}}, 1)$ (as $t^* = 1 - (\frac{c}{bn})^{\frac{1}{n-1}}$ when $\tau = 0$). On the other hand, a sufficient condition for having $t^* = 1$ is that $\left.\frac{\partial \tilde{N}(t,\tilde{t})}{\partial t}\right|_{t=1} \geqslant 0$, as if the derivative at $t = 1$ is non-negative,

the derivative at any $t < 1$ is also non-negative (for the strict concavity of $\tilde{N}(t, \tilde{t})$). Therefore, (71) cannot be satisfied. This condition can be expressed as:

$$-(1-\tau)v'(b(1-c)(c) + \tau(n-1)v'(b-cy)b((1-y)^{n-1}) \geqslant 0, \tag{75}$$

which is equivalent to

$$\tau \geqslant \frac{v'(b-c)c}{v'(b-c)(c) + (n-1)v'(b-cy)b((1-y)^{n-1})} \equiv \overline{\tau}. \tag{76}$$

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

Proof. Corollary 3: Let $z_\tau(t) \in [-1,1]$ be the normalized appropriateness of strategy t for an individual with type $\tau \in [0,1]$. This is computed with the extended norm $\tilde{N}(t,\tilde{t})$ in (23) and the normalization function in (6). Corollary 3 is equivalent to show that for any $t \neq \frac{w}{2}$ and $s_0 \geqslant \underline{s}$

$$\underbrace{s_0 z_0(\frac{w}{2}) + (1-s_0)z_\tau(\frac{w}{2})}_{\text{Average normalized appropriateness of } \frac{w}{2}} \geqslant \underbrace{s_0 z_0(t) + (1-s_0)z_\tau(t).}_{\text{Average normalized appropriateness of } t} \tag{77}$$

This is evident for any $t \neq \frac{w}{2}$ with $z_\tau(t) \leqslant z_\tau(\frac{w}{2})$.[40] Thus, I restrict attention to the $t \neq \frac{w}{2}$ with $z_\tau(t) > z_\tau(\frac{w}{2})$. Given that $z_0(\frac{w}{2}) = 1$, (77) is given by

$$s_0 \geqslant \frac{z_\tau(t) - z_\tau(\frac{w}{2})}{1 - z_0(t) + z_\tau(t) - z_\tau(\frac{w}{2})} \equiv s(t,\tau). \tag{78}$$

Then, I define $\underline{s}(\tau) \equiv \text{argmax}_{t\in[0,w]}\, s(t,\tau)$ and show that $\underline{s}(\tau) \in (0,1)$ for any $\tau \in (0,1]$. To do so, note that both numerator and denominator of (78) are positive (as (i) $z_\tau(t) > z_\tau(\frac{w}{2})$, and (ii) $z_0(t) < 1$). On the other hand, $1 - z_0(t) > 0$ implies

$$1 - z_0(t) + z_\tau(t) - z_\tau(\frac{w}{2}) > z_\tau(t) - z_\tau(\frac{w}{2}). \tag{79}$$

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

[40] If types τ find $\frac{w}{2}$ more socially appropriate than t, then the average appropriateness of $\frac{w}{2}$ is larger than the one of t for any $s_0 \in [0,1]$.

10 Appendix C: Lab Experiment

10.1 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 21: Dictator game with inequality

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 22: 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 23: 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 24: 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 25: 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 26: 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 27: Prisoner's dilemma

10.2 Secondary and Robustness Tests

I this section, I conduct secondary tests detailed in the pre-registration not included in the main text and several robustness checks.

10.2.1 Linear public goods game

Regression Tables

Dependent Variable: Social appropriateness				
	(1)	(2)	(3)	(4)
	Efficient PGG	Efficient PGG	Inefficient PGG	Inefficient PGG
Action	0.143***	0.143***	0.103***	0.105***
	(0.011)	(0.011)	(0.014)	(0.014)
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***	-0.455***	-0.944***
	(0.061)	(0.139)	(0.072)	(0.280)
Observations	600	594	618	612

Standard errors clustered at the individual level (in parentheses)

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

10.2.2 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).

10.2.3 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).

10.2.4 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).

10.2.5 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).

11 Appendix D: Comparison with Kimbrough and Vostroknutov (2023)

In this section, I compare the proposed theory with the one introduced in Kimbrough and Vostroknutov (2023) (hereafter KV). I first describe their theory and then discuss several differences. Two considerations are important to emphasize. First, I only provide a summarized description of their theory. For a more detailed version, I refer the reader to the corresponding paper. Second, I do not intend to conduct a horse race between the theories but rather show that they have different predictions in several settings.

11.1 KV's theory

The theory introduced in KV grounds injunctive norms in the psychology of dissatisfaction. The dissatisfaction with a particular outcome is defined relative to all feasible outcomes. Intuitively, an individual is dissatisfied with the outcome $x \in C$ if he could have obtained a higher utility with another feasible outcome. The authors define an outcome's social appropriateness to be inversely proportional to its aggregate dissatisfaction.

To obtain the injunctive norm proposed in KV, let

$$d_i(u_i(x), u_i(y)) = \max\{u_i(y) - u_i(x), 0\}. \tag{80}$$

represent individual i's dissatisfaction with consequence x because of possibility y.

Therefore, given a set of consequences C, individual i's aggregate dissatisfaction with a consequence x is given by

$$D_i(x|C) = \sum_{y \in C \setminus \{x\}} d_i(u_i(x), u_i(y)). \tag{81}$$

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

$$D(x|C) = \sum_{i \in N} w_i D_i(x|C), \tag{82}$$

where w_i represents individual i's social weight. Given the previous definitions, the authors propose the following injunctive norm:

Definition 1. 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 the 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, $\eta(x|C)$ is inversely proportional to $D(x|C)$ 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.

11.2 Different predictions with KV's theory

As I mentioned in the introduction, KV's theory focuses on outcomes' social appropriateness, while the proposed theory does so at the strategy level. Therefore, this section mainly focuses on dictator and allocation games, where both outcomes and strategies coincide.

Consider a situation where individual A has to choose between several feasible allocations $C = \{C_1, C_2, ..., C_k\}$, where $C_i = (a_i, b_i)$ determines the material payoffs of individuals A and B. All these allocations have a constant efficiency (i.e., $a_i + b_i = a_j + b_j$ for any $i \in \{1, ..., k\}$ and $j \in \{1, ..., k\}$ with $i \neq j$).

KV's theory predicts the most socially appropriate consequence is the *midpoint* consequence: the consequence that has an equal number of better and worse consequences for both individuals (see Proposition 6 in KV). On the other hand, the proposed theory predicts that when allocations have the same payoff efficiency, the most socially appropriate allocation is the one with the lowest difference in material payoffs between the two individuals (see Section 5.2.1).

This has important implications for the predictions of both theories in dictator and allocation games. For example, in the standard dictator game with $w = 10$ and $x \in [0, 10]$. Both theories predict that $t^* = 5$. However, KV's prediction is driven by the availability of the transfers above (and below) the equal split. 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 allocation is $t^* = 2.5$ (resp. $t^* = 7.5$), while the proposed theory still predicts $t^* = 5$ in both cases. To my best knowledge, this prediction remains untested.

The same occurs if we consider the modified dictator game in List (2007) (see Section 5.2.1). In this case, KV's theory predicts (as does the proposed theory) 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 $t^* = 2$ for the modified dictator game (see Figure 6 in KV). However, the elicited norms in Krupka and Weber (2013) show that the most socially appropriate transfer is $t^* = 2.5$ (see Figure 9). Similarly, Ellingsen and Mohlin (2023) show that the equal split is the most socially appropriate action in all the variations of the dictator game with taking option they consider (see Figure 4 in Ellingsen and Mohlin (2023)). This is in line with the proposed theory, but not with the one in KV.

Following the same reasoning, one could derive different predictions between the two theories in dictator games where (i) allocations have different efficiency or (ii) there is an earnings phase.

As mentioned in the introduction, the comparison between the two theories in other situations is challenging as outcomes and strategies do not coincide. In this case, one could try to distinguish between the theories by studying if the social appropriateness of the different strategies changes with the decision makers' beliefs with respect to others' behavior. While the proposed theory predicts that the injunctive norm is unaffected by the decision-maker's beliefs, this could not be the case in KV. I leave this as an open avenue of research.

References

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.

Brañas-Garza, P. (2006). Poverty in dictator games: Awakening solidarity. *Journal of Economic Behavior & Organization 60*(3), 306–320.

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.

Eckel, C. C. and P. J. Grossman (1996). Altruism in anonymous dictator games. *Games and Economic Behavior 16*(2), 181–191.

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

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.

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 (2023). A theory of injunctive norms. *Working Paper*.

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.

Korenok, O., E. L. Millner, and L. Razzolini (2008). Experimental evidence on inequality aversion: dictators give to help the less fortunate. *Virginia Commonwealth University 28*.

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

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

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.