Norms Make Preferences Social*

Erik O. Kimbrough† **Alexander Vostroknutov**‡

August 2013

Abstract

We explore a unifying explanation for prosocial behavior in which people care not about others' payoffs *per se*, but whether their own behavior accords with social norms. Individuals who are sensitive to norms will adhere to them so long as they observe others doing the same. A model formalizing this generates both prosociality (without relying on explicit distributional preferences) *and* well-known context effects (for which simple distributional preferences cannot account). A simple experiment allows us to measure individual-level norm-sensitivity and to show that norm-sensitivity explains heterogeneity in prosociality in public goods, dictator, ultimatum, and trust games.

JEL classifications: C91, C92, D03
Keywords: experimental economics, norms, social preferences, reciprocity

*This manuscript benefited immeasurably (title included) from repeated discussions with Ryan Oprea. We also thank Lucas Coffman, Dan Houser, Taylor Jaworski, Krishna Pendakur, Arno Riedl, Jared Rubin, Vernon Smith, Bart Wilson, participants in seminars at the University of British Columbia, University of Arkansas, University of Alabama, and conference participants at the 2011 North American meetings of the Economic Science Association and at the 2012 meetings of the Association for the Study for Religion, Economics, and Culture for helpful comments and gratefully acknowledge funding from Maastricht University's METEOR research school and the European Union Marie Curie FP7 grant program. All figures and data analysis produced using R: A Language and Environment for Statistical Computing (2013). The data are available from the authors upon request. Any remaining errors are our own.

†Corresponding Author: Department of Economics, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada, e-mail: ekimbrou@sfu.ca

‡Department of Economics, Maastricht University, P.O. Box 616, Maastricht 6200 MD, The Netherlands. e-mail: a.vostroknutov@maastrichtuniversity.nl

1 Introduction

> Man is as much a rule-following animal as he is a purpose-seeking one.
>
> (Friedrich Hayek, 1973, *Law Legislation and Liberty, Vol 1: Rules and Order*, p. 11)

> Without this sacred regard to general rules, there is no man whose conduct can be much depended upon. It is this which constitutes the most essential difference between a man of principle and honor and a worthless fellow.
>
> [...]
>
> that *reverence for the rule* which past experience has impressed upon him, checks the impetuosity of his passion, and helps him to correct the too partial views which self-love might otherwise suggest, of what was proper to be done in his situation.
>
> (Adam Smith, 1759, *The Theory of Moral Sentiments*, §3.5.2, *italics* added)

Over the last thirty years, economists have uncovered robust evidence of human sociality in simple, anonymous laboratory games. Subjects display systematic tendencies towards egalitarian outcomes, cooperative strategies and reciprocal behavior, often in violation of the predictions of selfish profit maximization.

How to interpret these observations remains a lingering puzzle. Economists have made great progress in understanding sociality by modifying the utility function of individual players to incorporate explicit distributional preferences (i.e. preferences over others' payoffs).[1] But such models have a blind spot – there is a long chain of evidence showing that minute changes to context in experiments can radically alter the nature and degree of sociality observed in the lab.[2]

In this paper we experimentally examine a unifying explanation for pro-social behavior that also provides a framework for understanding context effects. The idea is that sociality is driven not by preferences over payoffs of others, but rather preferences for following well-established social rules (be they written rules or informal norms). When people judge behavior, they compare it to an external, socially defined normative standard, and we argue that individuals internalize this process, judging their own behavior according to its conformity to convention. Because many norms are prosocial, behavior that ultimately results from a desire to follow social norms finds a proximate explanation in social preferences. It is in this sense that 'norms make preferences social' - though they needn't always.

[1] See e.g. Rabin (1993); Levine (1998); Fehr and Schmidt (1999); Bolton and Ockenfels (2000); Charness and Rabin (2002); Cox, Friedman, and Sadiraj (2008); Halevy and Peters (2009).

[2] A small sample of this literature includes Hoffman and Spitzer (1985); Forsythe, Horowitz, Savin, and Sefton (1994); Hoffman, McCabe, Shachat, and Smith (1994); Andreoni (1995b); Hoffman, McCabe, and Smith (1996); Eckel and Grossman (1996); Burnham, McCabe, and Smith (2000); Goeree and Holt (2001); Cherry, Frykblom, and Shogren (2002); McCabe, Rigdon, and Smith (2003); List (2007); Levitt and List (2007); Bardsley (2008); Falk, Fehr, and Fischbacher (2008); Andreoni and Bernheim (2009); Smith and Wilson (2013). One recent exception that finds limited effects of context in dictator games is Dreber, Ellingsen, Johannesson, and Rand (2012).

To clarify our hypotheses, our paper first formalizes the idea that people suffer a disutility from violating norms, and importantly, that people differ in their sensitivity to own norm violations.[3] The model implies that norm-sensitivity is correlated with the magnitude of prosocial behavior in a variety of games. We develop a simple experimental technique that allows us to measure individual-level norm-sensitivity, and we find support for the model in public goods, trust, dictator and ultimatum games.

A few examples illustrate the intuition for why norm-sensitivity (or norm-dependent preferences) can explain prosocial behavior: in the model, subjects cooperate in public goods games because they bring cooperative norms into the lab and expect others to do the same; when these expectations are dashed by evidence of the weakness of the norm in their group, subjects abandon the norm, generating the well-known pattern of cooperative decay in public goods experiments. Likewise dictator giving and ultimatum game rejections are driven by strong egalitarian norms related to windfall gains.

Changing contexts can alter these predictions because they alter the norms subjects lean on when judging behavior. For example, if control over resources is assigned non-randomly, norms related to windfall gains no longer apply, and subjects (e.g. dictators who have earned their role pre-play) are inclined to behave more selfishly. Similarly, the presence of an audience emphasizes to a subject that she ought to be concerned with others' views on what is appropriate in her present circumstances, making norms more salient.[4] In this sense, it is not an "error" or "irrationality" if people exhibit prosocial behavior in some contexts and not in others; it is a natural consequence of the fact that people care about norms and that norms are fundamentally context-dependent.

Thus, to the extent that an individual suffers disutility from violating norms, he may in some contexts take prosocial actions that are inconsistent with a narrow view of self-interest. Observed heterogeneity in sociality *across contexts* is due to the fact that norms are context-dependent, but, crucially for our story, observed heterogeneity *within a context* is just a product of the fact that individuals differ in the degree to which they care about norms.

In order to reveal this relationship between norm-sensitivity and prosociality, we develop a unique individual decision task (called the Rule-Following, or RF, Task) that measures subjects' preferences for following established rules and norms, in a context that has nothing to do with social interaction or distributional concerns. Specifically, we tell subjects to follow an arbitrary rule when doing so provides them with no monetary benefits and instead imposes explicit monetary costs proportional to the time spent following the rule. Under these circumstances, only

[3]This idea is well-known outside of economics, e.g. in psychology and sociology where early examples include Sherif (1936) and Merton (1957). Adam Smith's *Theory of Moral Sentiments*, which captures this idea nicely, predates this literature by nearly 200 years, and there has been a renewed interest in norms among economists (See e.g. Elster, 1989; Young, 1998; Krupka and Weber, 2009). The model introduced here and developed more fully in Kimbrough, Miller, and Vostroknutov (2013) is most closely related to López-Pérez (2008); Kessler and Leider (2012) and Krupka and Weber (2013).

[4]These are implications of the model, but both claims are consistent with evidence cited in footnote 2 above.

those individuals who are intrinsically motivated to adhere to rules and norms will follow the rule. Thus, the RF task provides us with a continuous measure of individual rule-following proclivity, and the model implies that those who suffer more disutility from violating rules and norms will be more likely to engage in certain other behaviors that are consistent with social norms. We compare individual behavior in the RF Task to behavior in some of the most important games of sociality in the literature: public goods games, trust games, dictator games and ultimatum games. The model has a central testable implication: a preference for following norms carries over from context to context, even if these contexts are unrelated.

In our main treatment, we divide subjects (without their knowledge) into groups based on the strength of their rule-following preferences, as measured by the RF task, and have them play standard repeated VCM public goods games. The model predicts that groups composed of individuals with high levels of norm-sensitivity will be able to sustain cooperation over time, and indeed, we find that high rule-following groups sustain cooperation with no evidence of decay, while low rule-following groups exhibit swift cooperative decay. Our RF task, which measures rule-following in an unrelated context, therefore strongly predicts cooperation in public goods games. A variety of robustness treatments validate this account of our results.

In further treatments we use RF Task measurements to understand behavior in three other standard games. We find that reciprocity (but not trust) is significantly higher among assortatively matched groups of rule-followers in trust games. We also find that giving in dictator games and rejections (but not offers) in ultimatum games are higher among rule-followers. All of these findings are natural implications of a model with norm-sensitive preferences.

Our framework allows us to understand the sources of heterogeneity in prosocial behavior across individuals, contexts and cultures. As we detail in the discussion, through the lens of norms and norm-dependent preferences, many apparently mystifying phenomena appear commonplace. Moreover, understanding prosocial behavior as norm-driven can help to account for extensive cooperation observed *outside* the lab since it implies the availability of simple screening mechanisms similar to our RF task that allow third-parties to identify norm-following types. For example, such a mechanism has been proposed to explain religious strictures regarding consumption: imposing costly restrictions on behavior may screen out insincere prospective members, facilitating cooperation among the remaining members (Iannaccone, 1992; Aimone, Iannaccone, Makowsky, and Rubin, forthcoming).

Observing that many people trade off own and other's payoffs so that other-regarding behavior is sensitive to its "price," economists have argued that the tools of demand theory license the inference that other-regarding actions are the result of distributional preferences or preferences for giving (see e.g. Andreoni and Miller, 2002; Fisman, Kariv, and Markovits, 2007). In light of our interpretation of the evidence, we view such explanations as useful, but reduced form. In particular, we would argue that distributional preference models confuse proximate and ultimate (or at least, nearer ultimate) cause. Under our interpretation, measured prosocial-

ity is driven by social norms that are salient in these games. As we show below, such a model implies behavior that is consistent with distributional preference explanations in many cases, but it can also account for a number of observations that are anomalous with respect to those models.

Crucially, in contrast to Gächter, Nosenzo, and Sefton (2013), we do not view social norms and distributional preferences as mutually exclusive explanations because any model of distributional preferences implicitly imports some norm, or norms, into the utility function – a model of fairness assumes a broad, *a priori* agreement about what constitutes (un)fairness, and models of reciprocity contain normative assumptions about which behaviors merit a particular reciprocal response. In a second sense, this is how norms make preferences social. The occasional failure of such models to correctly predict behavior then stems from the fact that sometimes the implicit norms are not appropriately matched to the context. When norm-triggering contexts change, so does behavior, as in the dictator games reported in Cherry, Frykblom, and Shogren (2002) or List (2007), in which measured aversion to inequity all but disappears.

Of course, our model is also reduced form in an important sense, and two crucial questions might arise after reading the paper: "Where do norms come from?" and "How do individuals recognize the norm in a given setting?" These are exactly the questions that, in our opinion, should be the topics of future research (some efforts along these lines include Axelrod, 1986; Ellickson, 1989; Young, 1993; Skyrms, 2004; Greif, 2006; Kimbrough, Smith, and Wilson, 2008; Acemoglu and Jackson, 2012; Gächter, Nosenzo, and Sefton, 2013; Krupka and Weber, 2013). One of the goals of this paper is to focus the discussion about prosociality on the importance of norms rather than individual preferences. If the reader accepts our argument that preferences are "made social" by reference to norms, then future research on prosociality should focus on explaining the creation, adoption and evolution of norms. A second important strand of research would seek to explain the origins of norm-sensitive preferences. One recent example providing an evolutionary foundation for one kind of norm-following can be found in Alger and Weibull (Forthcoming). We return to these issues below.

Finally, a word on one alternative interpretation of our results: one might argue that we are simply measuring the strength of experimenter demand effects in the RF task. Under this interpretation, our experiment shows that much of the sociality in lab experiments is driven by a desire to satisfy the experimenter on the part of the subjects. This interpretation is reasonable, though we prefer a different interpretation. We view susceptibility to experimenter demand effects as an instance of norm-sensitivity (where the norm might be giving the experimenter what he or she has paid for, following clear authority in hierarchies, etc). Our results show at least that following this norm is related to following others.

The next section briefly reports the model, section 3 reports the details of our experimental design, section 4 reports the results our experiments and a series of robustness checks, and section 5 summarizes our findings and concludes.

2 The Model

In this section we discuss the implications of a model with norm-dependent utility in repeated public goods games. For the sake of expositional ease we describe a modified game and equilibria without presenting the general model for all games. In Appendix A we describe the implications of the model for one-shot Dictator, Ultimatum and Trust games. However, it should be noted that all the game-specific definitions can be derived from a single model of norm-dependent utility for general extensive form games with observable actions reported in Kimbrough, Miller, and Vostroknutov (2013).[5] Before we describe the repeated game, we introduce some preliminary concepts.

In the context of games repeated once (dictator, ultimatum, trust, one-shot public goods) we model the *norm* (η) as a strategy profile. That is, a norm provides a description of the socially appropriate choices of each player in each information set. Thus, a norm defines the "right" choice in each possible contingency, independent of any possible future decisions made by others.[6] To demonstrate the intuition behind this definition consider a norm of equity in the ultimatum game. The equity norm prescribes accepting relatively equal divisions of the pie and rejecting all sufficiently unequal offers. Thus, the norm provides a full description of socially appropriate actions in all potential choice nodes of the Responder. For the Proposer a norm might prescribe dividing the pie equally.

The next ingredient of the model is the norm-dependent utility function. Our work here builds on the model presented in Kessler and Leider (2012), though there are other ways to model norm-dependence that generate similar conclusions (see footnote 3 above). Briefly, norm-dependence implies a utility function in which individual utility is increasing in own payoff and decreasing in the deviation between own action and the normative action.

To give the definition let us make several observations first. All (non-repeated) games that we consider in Appendix A have a specific structure: 1) each player moves only once and 2) the final payoff of player i in each end node (a_i, a_{-i}) of the game, defined by a single choice of action by each player a_i and a_{-i}, can be written as $u_i(a_i, a_{-i}) = u_{i,i}(a_i) + u_{i,-i}(a_i, a_{-i})$, where $u_{i,i}(a_i)$ is the part of the payoff that player i "chooses for himself" by choosing action a_i in the appropriate information set, for example, the share x of the pie in the ultimatum game, and $u_{i,-i}(a_i, a_{-i})$ is the part of the payoff "chosen" for player i by the other player(s) (in the ultimatum game we would have Proposer's utility $u_{i,i}(x) = x$; in case of rejection we would have $u_{i,-i}(x, R) = -x$ and $u_{i,-i}(x, A) = 0$ in case of acceptance, A). Notice that even if player i moves first and all other players move afterwards, the part of the utility $u_{i,i}(a_i)$ is defined unambiguously for all nodes and *does not depend on the subsequent moves*.[7] Given these preliminaries, let us define the

[5] A preliminary draft is available at: https://www.dropbox.com/s/f05ofcfwvjxa9hu/FramingKMV.pdf.

[6] This view is consistent with that of Elster (1989) in which norms are "either unconditional or, if conditional, not future-oriented."

[7] The idea of dividing the payoffs into pieces "chosen" by different players is generalized in a definition of a *frame*

norm-dependent utility for player i as

$$U_i(a_i, a_{-i}) = u_i(a_i, a_{-i}) - \phi_i g(\|u_{i,i}(a_i) - u_{i,i}(\eta_i)\|)$$

Here, $g : [0,1] \to [0,1]$ is a strictly convex increasing function with $g(0) = 0$ and $g(1) = 1$ that represents the disutility of deviating from the norm; ϕ_i is a parameter indicating the sensitivity of player i to deviations from the norm; η_i is the action of player i which is prescribed by the norm in the information set where a_i is available[8]; and $\|\cdot\|$ is an appropriately normalized absolute value. This normalization $\|\cdot\|$ is necessary to make the disutility of deviating from the norm comparable in all information sets. The absolute value of the payoff difference is divided by the maximal possible absolute value of payoff difference in each node. If $\phi = 0$, the agent is a standard selfish utility maximizer, and if $\phi \to \infty$, the agent will always follow the norm, independent of the payoff consequences.

We use this model to generate predictions for three of the games that we study in our experiments: Ultimatum, Dictator and Trust (see Appendix A for details).[9] We will describe the implications of the model for each of those games in the hypotheses section. Here we illustrate how the model can be extended to repeated games.

Repeated Public Goods Game. Suppose we have an n-player repeated, linear, voluntary contributions public goods game with T periods. $x_{it} \in [0,1]$ is the action chosen by player i in period t. The material payoff to player i in period t is

$$\pi_i(x_{it}, x_{-it}) = 1 - x_{it} + \alpha \sum_{j=1..n} x_{jt} = 1 - (1-\alpha)x_{it} + \alpha \sum_{j \neq i} x_{jt}$$

where $\alpha < 1$ and $\alpha n > 1$ (i.e. the payoff from full cooperation is larger than the payoff from full defection). Here the action set $[0,1]$ has a natural interpretation: 0 is the selfish action, corresponding to keeping the entire endowment in the private account; 1 is the cooperative action. The distance between actions is defined as the distance between corresponding real numbers.

of the game in Kimbrough, Miller, and Vostroknutov (2013). In general games, where players can move multiple times and possibly simultaneously, the components of $u_i(\cdot)$ at each node can be chosen in accordance with the context in which the game is played. This choice constitutes a frame of the game. In the general case, the optimal choice, given fixed norm-dependent utility, changes with different frames and thus allows us to model behavior influenced by framing effects.

[8] η_i is uniquely defined for each player i in each information set where player i moves since we consider only games with a single move for each player.

[9] While we do not discuss this class of games here, we note that the general formulation of the model in Kimbrough, Miller, and Vostroknutov (2013) also allows us to explain behavior in "distribution games" such as those studied in Engelmann and Strobel (2004), where own payoff is unaffected by own actions; in such a setting, agents with $\phi > 0$ will choose the distribution that accords with the norm, while agents with $\phi = 0$ are indifferent between all distributions. It should also be clear that the model can account for third-party punishment such as that observed in Fehr and Fischbacher (2004), where costly punishment trades off own costs against the disutility of violating a punishment norm.

The utility of player i in period 1 is

$$u_{i1}(x_{i1}, x_{-i1}) = \pi_i(x_{i1}, x_{-i1}) - \phi_i g(\|\eta - x_{i1}\|).$$

Here ϕ_i and g are the same objects as in the previous section and η represents the action considered a norm by all players.[10]

Now we come to an important difference between repeated and one-shot games. We hypothesize that when deciding whether to continue adhering to a norm in any period, individuals tend to consider the prior play of others. Thus, we introduce a notion of *norm-reciprocity* to formalize the intuition that people are not blind adherents to norms, who will continue to follow them under any circumstances. Instead, individuals may adapt their norm-following behavior to the behavior of others – following norms when others do so, and ceasing to when they do not.

To model this, first, we have to define how players determine whether the norm has been followed. To do this, we suppose that after each period, each player bases his evaluation on the average contribution of all players $m_t = \frac{1}{n}\sum_{j\in N} x_{jt}$.[11] We assume that as long as players have *on average* followed the norm, the player considers the norm satisfied. Next, we must alter the utility function to introduce dependence on prior actions. We define the utility of player i in period $t > 1$ as

$$u_{it}(x_{it}, x_{-it}) = \pi_i(x_{it}, x_{-it}) - \phi_i p(m_{t-1}) g(\|\eta - x_{it}\|).$$

We have added a new element to the utility function: the norm-reciprocity response $p(m_{t-1})$. This is a number in $[0, 1]$ which depends on the average actions of all players in the previous period m_{t-1}. In the simplest case, consider $p(m) = 1$ if $m = \eta$ and $p(m) = 0$ otherwise. This is the most basic reciprocal response which can be interpreted as follows: if, in the previous period, the average contribution to the public good was consistent with the norm, then in the current period the player will take the norm into account when choosing how much to contribute. If, however, on average the players contributed less than the norm prescribes, then the player reciprocates by ceasing entirely to care about the norm in the current period. This case is extremely simple, but, it captures the major intuition behind the model.[12]

The overall utility of player i in the repeated game is given by the sum of utilities in all

[10]Notice that now the norm is an *action* rather than a strategy profile. In any one-shot game with a continuous action space and payoff, separable in own and others' actions, which is strictly monotonic in own action (like in our Public Goods game), this formulation is equivalent to ours. In what follows we will treat repeated games in this way because of their special structure. This allows us to introduce a notion of norm-reciprocity (see below and appendix B).

[11]This is consistent with the information available to players in the standard VCM public goods game (and thus with our experimental design) in which players know n and learn the total contribution at the end of each period but do not know each individual's contribution.

[12]Note that we assume that contributions greater than prescribed by the norm are also "punished" by abandoning the norm. We assume this to be consistent with the general setup. However, it is easy to alter the $p(\cdot)$ function so that players do not negatively reciprocate when others contribute more than prescribed by the norm, and all results that follow would remain unchanged.

periods: $U_i = \sum_{t=1..T} u_{it}$. This describes an extensive game with observable actions Γ which, unfortunately, does not have a repeated game structure anymore, because payoffs between the repetitions are not independent. Our goal is to characterize Nash Equilibria of this game.

Consider the following strategy s_i^0 of player i:

- In period 1 choose $x_i^* := \text{argmax}_{x\in[0,1]} \ -(1-\alpha)x - \phi_i g(\|\eta - x\|)$;
- After history h contribute $x_i^*(h) := \text{argmax}_{x\in[0,1]} \ -(1-\alpha)x - \phi_i p(m(h)) g(\|\eta - x\|)$.

Here $m(h)$ is the average contribution after the last period of history h. This strategy corresponds to each player maximizing her utility in each period separately. Notice that $x_i^* \in [0, \eta]$, since for the values $x \geq \eta$ the function $-(1-\alpha)x - \phi_i g(\|\eta - x\|)$ is strictly decreasing in x. Also, $x_i^*(\phi_i)$ weakly increases in ϕ_i with $x_i^*(\phi_i) = 0$ for $\phi_i \leq \underline{\phi}$ and $x_i^*(\phi_i) = \eta$ for $\phi_i \geq \overline{\phi}$. $x_i^*(\phi_i)$ is strictly increasing on the interval $[\underline{\phi}, \overline{\phi}]$.[13]

Before we present our Propositions let us introduce some notation and a Lemma. For proofs of the Lemma and all Propositions below, see Appendix B. Let $\Sigma := \sum_{j\in N} x_j^*$ and $\Sigma_{-i} := \sum_{j\neq i} x_j^*$.

Lemma 1. Suppose $\frac{1}{n} < \eta$ and consider any subgame of Γ staring in period $t \in 2..T$ such that $m_{t-1} \neq \eta$. Then s^0, restricted to this subgame, is a NE.

In the NE of the subgames described in Lemma 1 all players choose to contribute zero. It should be noted, however, that, for some values of the parameters, this NE is not unique: another NE exists in which all players contribute η until the penultimate period. We consider such an equilibrium in subgames hardly plausible as it requires full coordination of all players on contributing according to the norm *after* some of them have already deviated from it. Nevertheless, the existence of such cooperative NE points to the possibility that deviators *can* return to norm-following given some commonly observed signal, like, for example, a message that focuses attention on norm following.[14]

Proposition 1. Suppose $\frac{1}{n} < \eta$ then the following is true

1.1 If $\Sigma_{-i} < n\eta - 1$ for all $i \in N$ then the strategy profile $s^0 = (s_i^0)_{i\in N}$ is a NE of Γ.

1.2 If $n\eta - 1 \leq (1-\alpha)n\eta$ and $\Sigma_{-i} \leq (1-\alpha)n\eta$ for all $i \in N$ then s^0 is a NE of Γ.

1.3 If $n\eta - 1 > (1-\alpha)n\eta$ and $\Sigma_{-i} \geq n\eta - 1$ for some $i \in N$ then s^0 is not a NE of Γ.

The strategy s^0, generates a NE in which players maximize their utility in each period, resulting in a collapse of cooperation after the first period. Notice that conditions of Proposition 1

[13] For certain g functions it is possible that $x_i^*(\phi_i)$ does not have either a flat 0 or flat η part (or both).

[14] Engel and Kurschilgen (2013) study public goods games in which they ask questions regarding the beliefs of the subjects before play. They find that asking subjects about the "norm" that should be followed makes them more cooperative, which is in line with our hypothesis.

roughly say that for this equilibrium to obtain, the ϕ_i parameters for all players should be low enough (this is expressed in terms of x_i^* which are increasing in ϕ_i). The important thing to notice is that this equilibrium *cannot* be sustained if some players have sufficiently high ϕ.

Next we identify two other NE of the game in which cooperation is sustained. Consider the following strategy s_i^1 of player i:

- In period 1 contribute η;
- After history h in period $t < T$ if $m(h) = \eta$ contribute η;
- After history h in period $t < T$ if $m(h) \neq \eta$ contribute $x_i^*(h)$;
- After history h in period T contribute $x_i^*(h)$.

Proposition 2. Suppose $\frac{1}{n} < \eta$. If for all i it is true that $\Sigma_{-i} \geq \eta\,[1/\alpha - 1]$ then the strategy profile $s^1 = (s_i^1)_{i \in N}$ constitutes a NE of Γ.

Proposition 2 says that if all players have sufficiently high ϕ, then there is an equilibrium in which, on the equilibrium path, cooperation is sustained at the level of the norm until the penultimate period.

Next we give an intermediate result for partial cooperation. Consider strategy $s_i^{2\ell}$ of player i:

- In periods 1 to $\ell < T - 1$ contribute η;
- After all histories h not on the path generated by the above rule contribute $x_i^*(h)$.

Proposition 3. Suppose $\frac{1}{n} < \eta$. If for all i it is true that $\Sigma_{-i} \geq \eta\,[1/\alpha - 1]$ and conditions of Proposition 1.1 or 1.2 are satisfied then the strategy profile $s^{2\ell} = (s_i^{2\ell})_{i \in N}$ constitute NE of Γ.

Proposition 3 indicates that if the parameters of the model are such that both cooperative and non-cooperative equilibria can be sustained, then there is also a collection of other equilibria in which cooperation can be sustained for any fixed number of periods.

In addition, the previous two Propositions indicate that we should expect to observe increasing cooperation as α increases in the interval [0,1). Note that this is consistent with extensive evidence from public goods games despite being inconsistent with the standard utility maximizing model (e.g. Isaac and Walker, 1988).

Now we briefly describe two Subgame Perfect equilibria that can be constructed using Propositions 1 and 2.

Proposition 4. Suppose that $\frac{1}{n} < \eta$ and the assumptions of Proposition 1.1 or 1.2 hold. Then strategy profile s^0 is a SPNE of Γ.

Proposition 5. Suppose $\frac{1}{n} < \eta$. If for all i it is true that $\Sigma_{-i} \geq \eta\,[1/\alpha - 1]$ then strategy profile s^1

is a SPNE of Γ.

Finally, we discuss the effect of incomplete information in the repeated public goods game. In appendix B.1 we show that the cooperative and non-cooperative SPNE just described also constitute Perfect Bayesian Nash Equilibria with essentially no additional assumptions on the uncertainty over ϕ. The reason for this result is that incomplete information does not really impact players' payoff maximizing choices, since they are independent of the characteristics of others, which enter expected utility only indirectly through others' actions.

Result 1. *Propositions 1, 2 and 3 describe equilibrium behavior in the repeated public goods game:*

- *For low enough norm sensitivity parameters $(\phi_i)_{i\in N}$, which satisfy the inequalities $\sum_{j\neq i} x_j^* < \min\{\eta\,[1/\alpha - 1]\,, n\eta - 1\}$ for all $i \in N$, only the non-cooperative equilibrium exists in which on the equilibrium path players contribute their optimal one-shot game amounts in the first period and then contribute 0 in all other periods;*

- *If parameters $(\phi_i)_{i\in N}$ are sufficiently high and satisfy $\sum_{j\neq i} x_j^* > \max\{(1-\alpha)n\eta, n\eta - 1\}$ for all $i \in N$ only the fully cooperative equilibrium exists in which all players contribute η in all periods but the last, where they maximize the one-shot game payoff;*

- *For the intermediate values of parameters $(\phi_i)_{i\in N}$ cooperative equilibria exist in which cooperation can last from 1 period to $T-1$ periods with later-period defection to 0 as in the non-cooperative equilibrium.*

Thus, the SPNE in which players follow the norm by contributing η only obtains if ϕ_i's and the resulting x_i^* are sufficiently high. In other words, in groups where all members are sufficiently concerned about social norms, there is a SPNE in which the normative action is sustained in the repeated game for all but the final period. Thus, if we can identify norm-sensitive (high-ϕ) and norm-*insensitive* (low-ϕ) types *ex ante* and assortatively match them, we should observe sustained cooperation over time in the high-ϕ groups, even without punishment.

More generally, taking the results here and those in Appendix A, the model predicts that heterogeneity in ϕ can explain heterogeneity in prosocial behavior. That is, individuals with different levels of concern about norms will exhibit differences in other-regarding behavior. Thus, we design an experimental task (described in Section 3) to measure a proxy for the parameter ϕ, and we test the predictions of the model in four games.

3 Experimental Design

The experiment consists of two decision-making stages and a questionnaire. In stage 1, which we call the Rule Following task (RF), subjects control a stick figure walking across the computer screen. Each subject makes 5 decisions concerning the amount of time they wait at a sequence of

red traffic lights, each of which will turn green 5 seconds after their arrival. Figure 1 shows the screen that the subjects see.

At the beginning of the RF task, the stick figure is standing at the left border of the screen, and all traffic lights are red.[15] Subjects initiate the RF task by pressing the START button. At this moment, the stick figure starts walking towards the first traffic light. Upon reaching the first red light, the stick figure automatically stops. The light turns green 5 seconds after the stick figure stops; however, subjects are free to press a button labeled 'WALK' any time after the stick figure stops. When a subject presses WALK, the stick figure continues walking to the next red light before stopping again, and subjects must once again press WALK to continue to the next light. Throughout the RF task, the WALK button is shown in the middle of the screen. Subjects can press the WALK button at any time during the RF task. However, it becomes functional only when the stick figure stops at a traffic light.

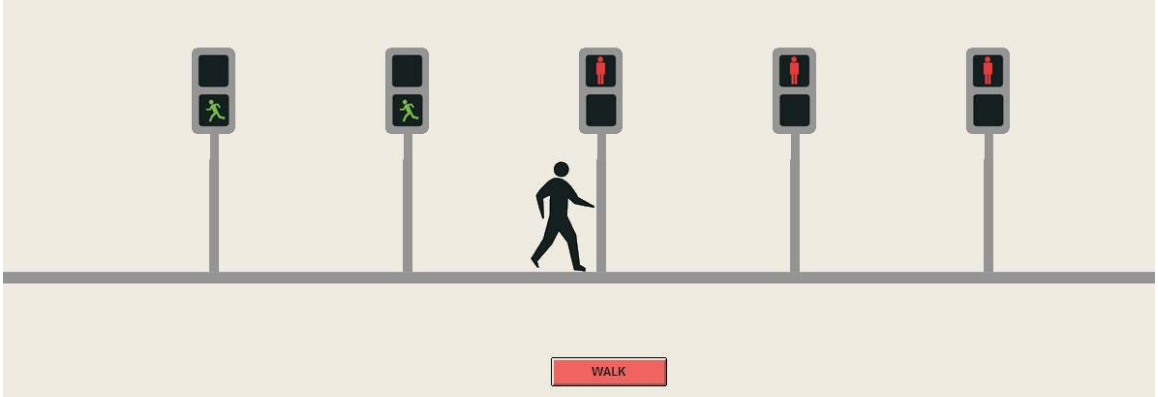

Figure 1: Screen shot of the Rule Following (RF) task.

Subjects receive an endowment of 8 Euro, and they are told that for each second they spend in the RF task they will lose 0.08 Euro. It takes 4 seconds to walk between each traffic light, and 4 seconds from the final light to the finish. Therefore, all subjects lose around 2 Euro walking, and if a subject waits for green at all 5 traffic lights, she will lose an additional 2 Euro waiting. Thus the most a subject can earn in the RF task is 6 Euro (if she spends no time waiting at traffic lights), and the most she can earn if she waits is 4 Euro (if she waits exactly 5 seconds at each light). In the instructions for the RF task (see Appendix C) subjects are told: "The rule is to wait at each stop light until it turns green." No other information, apart from the payment scheme and a general description of the walking procedure, is provided in the instructions.[16]

The rule following task creates a situation, familiar to most subjects, in which they are asked to follow an arbitrary rule at some cost to themselves. Waiting at a stoplight when there are no other vehicles or individuals in sight is an example of seemingly 'irrational' obedience, in the

[15]Before subjects start the task, they see a short cartoon in which the traffic lights blink from red to green. This ensures that subjects understand that the lights can turn green.

[16]If subjects asked what would happen if they pass through the red light, one of the experimenters explained that all information relevant to the experiment is given in the instructions.

sense that (barring the presence of traffic cameras) there is no cost to breaking the rule. In such circumstances, the usual justification for obeying traffic law—ensuring the safety of drivers and pedestrians—has no bite because there are no other drivers or pedestrians to protect or be protected from. Yet in our experience, it is quite common for people to stop and wait impatiently at traffic lights, even in the middle of the night. We argue that norm-dependent preferences provide the explanation. Individuals who care about norms (or rules) will wait; their disutility from violating social expectations is greater than the utility from quickly getting to the destination, and others who are not so concerned (or who face large opportunity costs of waiting) will run the light.

In the laboratory we control the opportunity cost of obedience, and by observing individual willingness to follow our arbitrary and costly rule, we can measure a proxy for the parameter ϕ. By explicitly stating that "the rule is ...," we induce a common prior that others will follow the rule. This allows us to eliminate possible norm-reciprocal responses among the participants. Thus, when we observe the extent to which an individual follows the rule, this reveals information about his/her ϕ.

The choice data important for our hypotheses is the number of seconds that each subject spends waiting at the traffic lights. This number is directly proportional to the amount of money that the subject gains by not waiting (she can gain between 0 and 2 Euros). Suppose that the norm prescribes to wait at all lights (to gain 0 Euros). Then, the optimal waiting time can be computed from the maximization problem $x^* = \text{argmax}_{x\in[0,2]}\, x - \phi g(x/2)$.[17] The optimal x^* solves $g'(x^*/2) = 2/\phi$. Thus, since g is strictly convex and increasing we can conclude that x^* (which we observe) and ϕ are monotonically related, which allows us to use x^* as a proxy for ϕ.

Another way to view obedience in our RF task is as a pure "experimenter demand" effect. Under that interpretation, we are simply using demand effect sensitivity as a proxy for ϕ. This has the nice feature that a long-time bogeyman of experimenters turns out to be an ally. We are sympathetic to this view, but we would argue that any experimenter demand effect is actually a manifestation of the norm-dependence underlying the model, else why should individuals be concerned about the demands of the experimenter?[18] In our RF task, though, we also induce a familiar context – we were concerned that otherwise our 'rule' would be ignored. Because of the induced context it is not immediately clear how much of our observed obedience is due to experimenter demand and how much to norms associated with traffic lights. For this reason we conducted an additional "No Rule" treatment, which we describe below (see Section 4.3).

Given the model in Section 2, the parameter ϕ should be closely related to behavior in a variety of games. Specifically, high-ϕ individuals will be more likely to follow norms of cooperation. Thus, after identifying high- and low-ϕ individuals, we test this hypothesis in a repeated

[17] x is divided by 2 in order to have the argument of g to change in the range from 0 to 1.

[18] Levitt and List (2007) suggest that demand effects may be responsible for much of the apparently anomalous heterogeneity in prosocial behaviors that the model purports to explain (e.g. context effects); in these cases too, we argue that sensitivity to a demand effect actually reveals norm-dependent preferences.

VCM public goods games with fixed matching (PG) treatment, and we examine some competing explanations with some additional diagnostic experiments. Then, to test the robustness of the model for social behavior more generally, we also test it in repeated trust games with random rematching (TG) as well as one-shot dictator (DG) and ultimatum games (UG).

Stage 1 of each treatment is the Rule Following task as described above. In stage 2 of the PG treatment subjects play 10 periods of a repeated public goods game with a voluntary contributions mechanism in fixed groups of 4 (similar to Isaac and Walker, 1988). In stage 2 of the TG treatment subjects play a trust game 6 times (Berg, Dickhaut, and McCabe, 1995). In particular, each subject plays the game twice with each other subject in the group, once as a first mover and once as a second mover. The order is randomized, and subjects receive no identifying information about their partner. In stage 2 of the DG treatment, subjects play one round of the dictator game due to Forsythe, Horowitz, Savin, and Sefton (1994), where one-half are randomly assigned to be dictators. In stage 2 of the UG treatment, subjects play one round of the ultimatum game due to Güth, Schmittberger, and Schwarze (1982), where we elicit responder decisions via the strategy method.[19]

Before making decisions in the RF task, subjects only receive instructions for that stage. They are aware that the experiment will consist of several stages, but they know neither what they will do in the next stage(s) nor the connection between the RF task and consecutive stages. In particular, subjects see a label that reads "Part 1" at the top of the rule following instructions (see Appendix C). In previous dictator game experiments, knowledge of the existence of an unspecified second-stage has been shown to alter subjects' behavior by making them more cooperative in expectation that their first-stage behavior may influence their second-stage reputation (Smith, 2008). If subjects are concerned for their reputation and thus wait longer than they might in a treatment without an implicit 'shadow of the future' (or, similarly, with a double-blind protocol), this could only bias behavior in one direction. Any such 'strategic' rule-following would only yield false positives, diluting the information content of the RF task and strengthening any results we obtain that confirm our predictions.

In our main Public Goods treatment each subject receives an endowment of 50 tokens at the beginning of each of the 10 periods (1 token = 1 Euro cent), and she must choose how to divide her tokens between a group account and a private account. In each period, each subject earns the sum of the amount placed in the private account plus the individual return from the group account, which is (0.5 * (sum of all contributions)). Thus, it is individually optimal to contribute nothing to the group account and Pareto optimal for all subjects to contribute their entire endowments. After each period, subjects learn their earnings in that period, the sum of group account contributions from all members of their group, and their total earnings through

[19]Our model also predicts differences in rejection behavior in the standard extensive form version of the game, but pilot sessions yielded a sufficiently low rejection rate that it would have been very costly to collect enough data to detect differences using that method.

that period. Subjects are informed only that they will participate in 'several' periods of decision-making.

Crucially, unknown to the subjects, their decisions in the RF task determine with whom they are grouped in the PG stage. First, we randomly divide subjects into groups of 8 (sessions consisted of 16, 24 or 32 subjects). Second, within each group of 8, we rank subjects according to the total time they spent waiting in the RF task – at least 25 seconds for those subjects who waited for the green light at all traffic lights and close to 0 seconds for those who did not wait at any light. Then, in each group of 8, we separate the top 4 subjects (rule-followers) and the bottom 4 subjects (rule-breakers) into two groups for stage 2. After we match subjects, there is no interaction between any groups of 4. Subjects are not informed about the matching procedure, and they are told only that they will now interact with a fixed group of three other participants (see Appendix G).[20]

In each period of the TG treatment, we divide each group of 4 into pairs. During the 6 periods, pairs are re-matched so that no pair ever interacts in two consecutive periods. Each subject participates 3 times in the role of first mover (blue person) and 3 times as a second mover (red person, see Appendix F). Subjects are informed only that they will make several decisions, but they are aware that they will participate in both roles.[21] We employ an identical matching procedure to that in the PG treatment, and again subjects are not informed of that fact.

Each subject receives an endowment of 80 tokens in each period (1 token = 1 Euro cent). The first mover chooses to send any amount between 0 and 80 tokens, knowing that the amount sent will be multiplied by 3 and given to the second mover. The second mover then chooses to send back to the first mover any amount between 0 and the amount received. In each period the earnings of the first mover are (80 tokens - tokens sent to the second mover + tokens sent back from the second mover). The earnings of the second mover are (80 tokens + tokens received from the first mover - tokens sent back to the first mover). After each period subjects observe the amounts sent, received and returned as well as their total earnings through that period.

In the DG and UG treatments, the RF task has no bearing on the second stage. Subjects are randomly assigned to be Proposers who are allocating a pie worth 16 Euro. In the DG treatment, the Proposer's offer (x) is final: the Proposer receives $16 - x$ and the Responder receives x. In the UG treatment, while the Proposer is choosing how much to offer to the Responder (y), the Responder also chooses the minimum offer he would be willing to accept (y^*). If $y \geq y^*$, the Proposer receives $16 - y$ and the Responder receives y; otherwise, both receive nothing.

After stages 1 and 2, subjects answered the Moral Foundations Questionnaire, which was designed to measure the extent of an individual's concern for certain fundamental moral issues

[20]Note that we did not deceive our subjects. None of the statements in the instructions are false or misleading. It is a separate, and also interesting, question whether subjects' behavior would change if they had knowledge of the sorting procedure, but our purpose was to use isolated rule-following behavior to identify subjects' types.

[21]Burks, Carpenter, and Verhoogen (2003) find that telling subjects that they will be playing both roles reduces both trust and reciprocity relative to a treatment in which they are unaware.

(Graham, Haidt, and Nosek, 2008, see Appendix H). Then subjects received cash equal to the sum of money earned in stages 1 and 2. We substituted earnings from the RF task for a formal show-up payment. The experiments were conducted at Maastricht University's BEELab between May 2011 and February 2013. Overall 72 subjects participated in the PG treatment (18 groups of 4), 96 subjects participated in the TG treatment (24 groups of 4), 134 subjects participated in the DG treatment, and 138 subjects participated in the UG treatment.

3.1 Hypotheses

Our rule-following task allows us to classify subjects according to ϕ by observing the extent to which they incur costs in order to follow an arbitrary rule. Our model predicts that agents with high values of ϕ will be more inclined to behave in accordance with social norms of generosity and reciprocity than those with lower values of ϕ.

In our experiments, we do not allow subjects to discuss strategies, and we do not provide contextual cues in the instructions meant to induce particular norms. Thus, the success of our screening mechanism relies on the subsidiary hypothesis that subjects import norms of behavior from outside of the lab that influence their decision-making. Suffice to say that this idea has a long history in experimental economics. For example, in dictator games, Hoffman, McCabe, and Smith (1996); Cherry, Frykblom, and Shogren (2002); and List (2007) have all interpreted their results in terms of norms imported from daily life, and Krupka and Weber (2013) have developed a method of measuring these norms directly. Thus our main treatment tests the following hypothesis, derived from the model in Section 2:

Hypothesis 1: (PG) groups of rule-followers will sustain higher contributions than groups of rule-breakers.

As robustness checks meant to rule out some alternative hypotheses, we ran the following additional diagnostic treatments, which we will discuss in sections 4.1.1 and 4.3:

1. A NoSort-PG treatment in which subjects first performed the rule-following task and then played the public goods game with 3 randomly chosen individuals (64 subjects, 16 groups of 4)

2. A Reverse-PG treatment in which the public goods game was played first with random matching into groups of 4, followed by the Rule Following task and the questionnaire (48 subjects, 12 groups of 4)

3. A NoRule-PG treatment in which the phrase "The rule is to wait at each stop light until it turns green" in the instructions for the RF task was replaced by "5 seconds after the stick figure reaches a stop light, it will turn from red to green" (24 subjects, 6 groups of 4)

4. A NoRule-Reverse-PG treatment combining (2) and (3) (24 subjects, 6 groups of 4)

The NoSort and Reverse treatments, which use unsorted groups, provide a baseline against which we compare the path of public goods contributions when groups *are* sorted according to our proxy for ϕ; this allows us to test norm-reciprocity more directly by observing mixed-type groups. The NoRule treatments, on the other hand, allow us to determine what portion of observed rule-following is due to the the statement that "the rule is..." and what portion is due to the induced context.

Then, we performed additional treatments to test the hypothesis that our RF task predicts sociality in three other classic games from the social preferences literature. Appendix A details the model's implications for Trust, Dictator and Ultimatum games. Specifically, the model generates the following hypotheses:

Hypothesis 2: (TG) groups of rule-followers will exhibit more reciprocity than groups of rule-breakers, *but* there will be no difference in trust.

Hypothesis 3: (DG) rule-followers will give more than rule-breakers.

Hypothesis 4: (UG) rule-followers will have higher rejection thresholds than rule-breakers, *but* there will be no difference in offers.

Some brief intuition may be useful. Hypothesis 2 stems from the fact that a norm of reciprocity (or equity) will lead second movers who care about norms to return a larger portion of what they receive. However, since the first mover does not know the second-mover's type, we cannot expect differences in first mover behavior - even if there is a social norm of sending high amounts. Under complete information, all else equal, high-ϕ types would send amounts closer to the norm, but there are also strategic reasons for low-ϕ types to send a significant amount if they believe the second mover is a high type and would reject low offers. The reason for Hypothesis 3 should be obvious; to the extent the the norm specifies higher contributions, people who care more about norms should give more. The intuition for hypothesis 4 is that second movers who care about norms of equity will reject unequal offers. However, as in the trust game, since the first mover does not know the second mover's type, there may be strategic reasons for low-ϕ first movers to send large amounts. Note that with respect to sender behavior in trust and ultimatum games, the model makes distinctive predictions that differ sharply from the predictions of models with purely distributional preferences. See Appendix A for details.

As we discuss in our conclusion, it is a separate (though similarly interesting) question whether our measure of norm-sensitivity also reveals itself in within-subject behavioral sensitivity to context manipulation. This is not our research question here, and we leave this for future work.

We also ran but do not report our first TG session, which fell prey to a software error, and two extensive form ultimatum game sessions (as opposed to the reported sessions which employ the strategy method) which we mentioned in footnote 19 above. These data are available from the

Treatments:	PG	NoSort-PG	NoRule-PG	Reverse-PG	NoRuleReverse-PG	TG	DG	UG
`Stage 1`	RF	RF	NoRule	PG	PG	RF	RF	RF
`Stage 2`	PG	PG	PG	RF	NoRule	TG	DG	UG
`Post Experiment`	Moral Foundations Questionnaire							
`Group Size`	4	4	4	4	4	4	2	2
`Sorted (Y/N)`	Y	N	Y	N	N	Y	N	N
`# of Subjects`	72	64	24	48	24	96	134	138
`# of Groups`	18	16	6	12	6	24	67	69
`# of Obs. per Group`	10	10	10	10	10	6	1	1

RF - Rule-Following Task, NoRule - RF task with no rule

DG - Dictator Game, UG - Ultimatum Game, TG - Trust Game, PG - Voluntary contributions PG

Table 1: Summary of Experimental Design

authors. No other data were collected for this experiment either in the form of pilots or other sessions/treatments. All experiments were programmed in z-Tree (Fischbacher, 2007). Table 1 summarizes our experimental design.

4 Experimental Findings

In this section we analyze our main public goods treatment, and then we turn to additional diagnostic treatments meant to compare and rule-out competing explanations. Next, we report the results of our DG, UG and TG treatments to test the robustness of the model, and after describing the data from each game and summarizing our results, we review the data from the rule-following task and finish with a discussion of the implications of our findings.

4.1 Public Goods Treatment

Table 2 displays average public goods contributions and waiting times for individuals in rule-following and rule-breaking groups. Relatively high waiting times among 'rule-breaking' groups are explained by the fact that many individuals classified as rule-breakers (because they were in the bottom four in their group of eight in the RF task) nevertheless followed the rule. As we discuss below, only 37.5% of subjects broke the rule at all, and much of the data is clustered close to the 25 second cutoff for rule-following, so there is considerable noise in our measurement.[22] Despite that noise, on average, rule-following groups contribute 17 percentage points more of their endowment to the public good than rule-breaking groups over the entire experiment, and

[22]In our *No Rule* treatments, the proportion of those breaking the rule increases to 87.5%, which suggests a strong effect of our statement that "The rule is…." We return to this below.

the difference is even larger in the second half of the experiment. Figure 2 displays time series of mean total contributions and associated standard errors in rule-following and rule-breaking groups. From the figure, it is clear that contributions decline over time only among rule-breaking groups, and we find statistical support in Table 3 which reports Wilcoxon rank-sum tests of the hypothesis of equality of mean group-wise contributions by group type for each period.

Variables	Rule-Following Groups	Rule-Breaking Groups
`Percent Contributed (All Periods)`	63.84 (2.020)	46.24 (1.890)
`Percent Contributed (Periods 1-5)`	65.01 (2.742)	55.92 (2.569)
`Percent Contributed (Periods 6-10)`	62.67 (2.972)	36.56 (2.585)
`Waiting Time (Seconds)`	27.19 (0.090)	20.39 (0.438)

Standard errors in parentheses.

Table 2: Mean Public Good Contributions and Waiting Time by Group Type

Period	1	2	3	4	5	6	7	8	9	10
`Test Statistic` ($W_{9,9}$)	40	47	61.5	55.5	63	64	60	60.5	61	67
`p-value`	0.53	0.30	**0.035**	0.100	**0.026**	**0.021**	**0.047**	**0.042**	**0.039**	**0.011**

Bolded entries statistically significant with p-value < 0.05. One-sided tests.
$W_{n,m}$ indicates the Wilcoxon test statistic with n and m observations per group type.

Table 3: Wilcoxon Tests of Mean Group Contribution, μ, by Period

In 7 out of 10 periods, we reject the null hypothesis of equal mean contributions in favor of the alternative that rule-following groups contribute more to the public good. Furthermore, comparing average group contributions over the first 5 periods and last 5 periods, additional Wilcoxon tests indicate that mean group contribution is significantly higher in rule-following groups than in rule-breaking groups in both early periods ($W_{9,9} = 61$, p-value = 0.039, one-sided test) and late ($W_{9,9} = 65$, p-value = 0.017, one-sided test). If we also include the 6 groups from our No-Rule treatment in the analysis, which were also sorted according to waiting time, the results remain essentially unchanged.[23]

[23]See figure I1 in Appendix I showing time series including sorted NoRule sessions in the computation of rule-following and rule-breaking means.

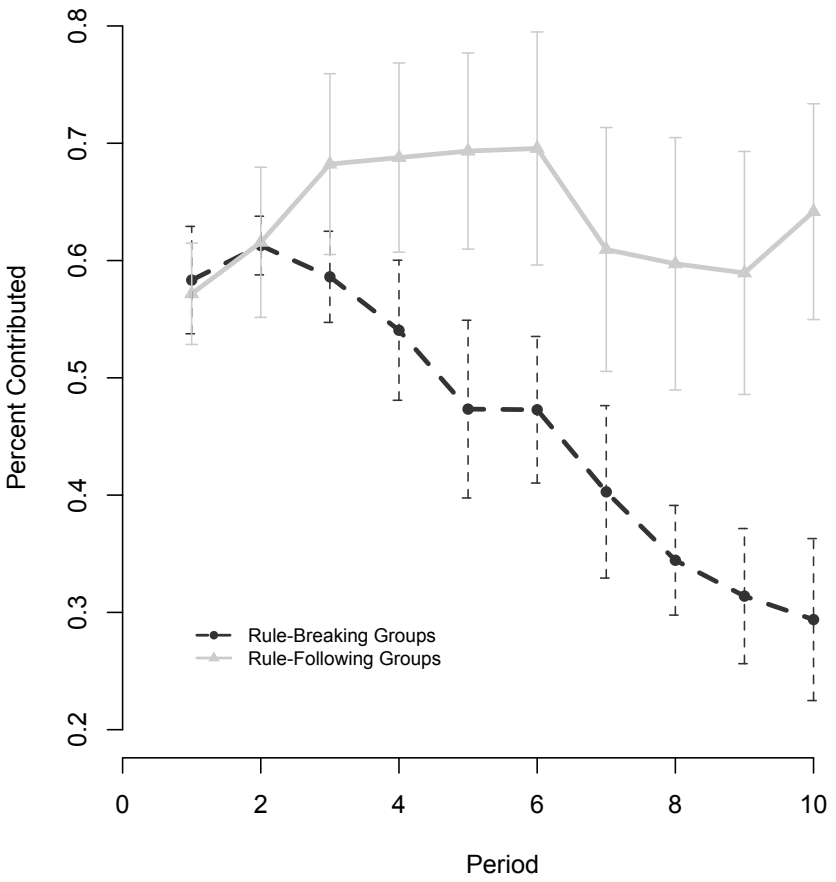

Figure 2: Time series of mean percent of endowment contributed $+/-$ 2 SEs, for rule-following and rule-breaking groups in the PG treatment (9 independent observations each).

Finding 1: *In accordance with hypothesis 1, rule-followers sustain significantly higher contributions than rule breakers in the VCM public goods game.*

4.1.1 Robustness

Having confirmed our main hypothesis, we now report treatments designed to test its robustness and compare it to other competing explanations. For example, one potential concern is that our rule-following task simply measures other-regarding preferences - people who follow rules are altruists. If this is true, then it is no surprise that rule-followers are more cooperative in Public Goods games. To distinguish these hypotheses, we can exploit our NoSort and Reverse treatments in which subjects were not sorted into groups.

If our rule-following task captures other-regarding preferences—instead of norm-dependence and norm reciprocity as we hypothesize—then the contributions of rule-breakers should systematically differ over time from those of rule-followers, even in the absence of sorting. We test this hypothesis using our data from 16 groups of 4 in the NoSort-PG treatment and 12 groups of 4 subjects in the Reverse-PG treatment (in which the order of the two stages was reversed).

Figure 3 shows mean group contribution by period for the NoSort-PG and Reverse-PG treatments as well as for Rule-Followers and Rule-Breakers in the PG treatment. When subjects are matched randomly into groups, the well-known pattern of cooperative decay reappears. In period 1 of the Reverse-PG and NoSort-PG treatments, the mean contribution is 60% of the endowment; whereas, in the PG treatment both rule-followers and rule-breakers average 58%. However, by period 10, Reverse-PG mean contributions decline to 41% of the endowment and NoSort-PG contributions fall to 43%, while rule-followers contribute 64% and rule-breakers contribute 29%. A Wilcoxon test rejects the null hypothesis of equal mean contributions in period 10 between rule-followers and the pooled NoSort-PG and Reverse-PG groups in favor of the alternative that contributions are higher among rule-following groups ($W_{9,28} = 181$, p-value = 0.027, one-sided test), but we cannot reject the null hypothesis of equal mean contributions between rule-breakers and NoSort-PG and Reverse-PG groups ($W_{9,28} = 159$, p-value = 0.257, two-sided test). These results are essentially unchanged if we perform separate tests for the NoSort- and Reverse-PG treatments.

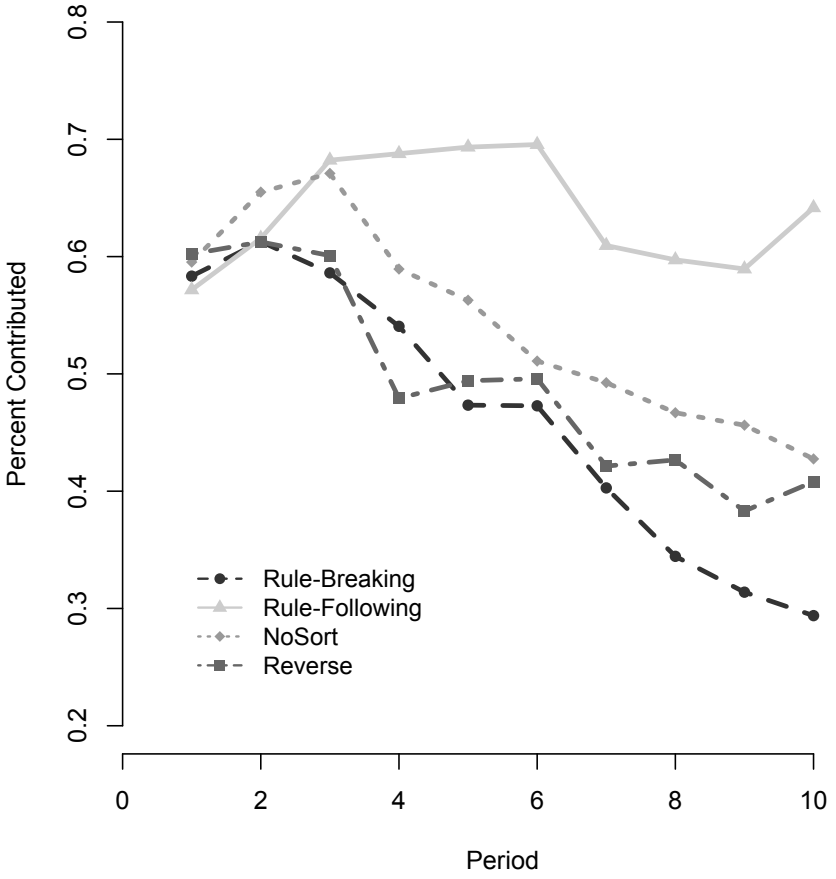

Figure 3: Time Series of Mean Group Public Good Contributions by Treatment

Moreover, we find no significant relationship between individual RF scores and contributions to the public good in either of the treatments without sorting. Pooling the data from the

NoSort- and Reverse-PG treatments, we estimate a mixed-effects panel regression of individual contributions on RF task waiting time and a period trend; we include random effects for each group and nested individual-in-group to control for repeated measures. The regression (output available upon request) reveals no significant relationship between individual RF behavior and contributions, though the period term is negative and significant. Therefore, we conclude that the sorting procedure in the PG treatment eliminates cooperative decay in rule-following groups. This provides support for a model in which norm-following behavior, even among those with high values of ϕ, is *conditional* on norm-following by others. Recall the condition for sustained cooperation in section 2: rule-following individuals, when matched with other individuals who are sufficiently concerned with adhering to social norms, are willing to conform to norms of high contribution, but in the absence of other rule-followers, such norms are unsustainable as rule-followers respond to low contributions by others by decreasing their own.

This interpretation is supported directly by additional mixed-effects panel regression analysis. Restricting our attention to the NoSort-PG and Reverse-PG treatments, we regress individual contributions on a constant term, a period trend, the total time the individual spent waiting in the RF task, the one-period lagged average of others' contributions to the public good, a reverse treatment dummy, and an interaction between waiting time and lagged others' contributions. We include random effects for each group and subject-in-group to control for repeated measures and within group correlations, and we estimate the model using restricted maximum likelihood. Our estimates of the conditionality of cooperation are contained in the coefficient on lagged others' contributions and its interaction with RF task waiting time. Regression output is reported in Table 4. A positive and significant estimated coefficient on the interaction between lagged others' contributions and waiting time provides clear evidence of norm reciprocity among rule-followers – that is, contributions made by individuals who wait longer in the RF task are more responsive to the contributions made by others.

PG Contribution	Coef.	Std. Err.	z value	Pr($>\|z\|$)
(Intercept)	36.288	4.900	7.41	0.000***
Period	-1.377	0.178	-7.71	0.000***
Reverse	-1.992	3.326	-0.60	0.549
Time Waited	-0.259	0.173	-1.50	0.134
Mean Others' Contributions$_{t-1}$	-0.029	0.117	-0.25	0.804
Time Waited-Mean Others' Contributions$_{t-1}$	0.009	0.005	1.79	0.074*

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Reciprocity in PG Contributions, NoSort and Reverse Treatments

Finding 2: *Observed differences between rule-breakers and rule-followers cannot be explained by simple models of distributional preferences. Norm-dependent utility and norm reciprocity account for cooperative decay in un-sorted public goods experiments.*

4.2 Other Games

Having confirmed the predictions of the model in the public goods game and ruled out some competing hypotheses, we now test whether the model can predict prosocial behavior in three other well-studied games: trust games, dictator games, and ultimatum games. The findings below provide evidence that norm-dependence serves as a unifying explanation of prosocial behavior.

4.2.1 Trust Game Treatment

Recall that in the TG treatment, as in the PG treatment, subjects are randomly assigned into groups of eight and then sorted into two smaller groups of four according to the time they spent waiting in the RF task. The four who waited the longest are in the "rule-following group," and the four who waited the shortest amount of time are in the "rule-breaking group." As in the PG treatment, there is considerable noise in the measurement of rule-following since many people grouped with rule-breakers were themselves actually rule-followers who simply spent a second or two less time waiting than those who were sorted into rule-following groups. Nevertheless the predictions of the model are borne out again.

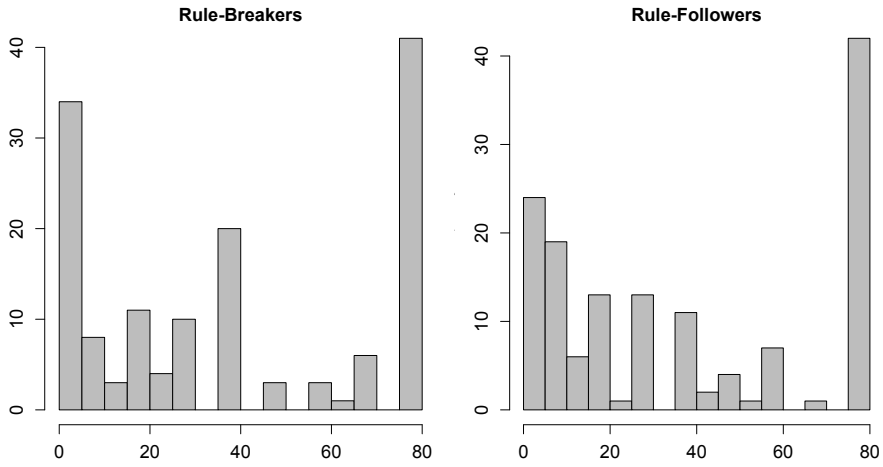

Figure 4: Histograms of Amount Sent in the TG treatment

Figure 4 displays histograms of the amount sent by first movers from rule-following and rule-breaking groups. The model presented in the Appendix A predicts that proposals in the TG depend on Proposer's norm sensitivity parameter ϕ as well as on her beliefs about the sensitivity parameter of the Responder. This implies that rule-followers and rule-breakers might propose similar amounts if they have similar beliefs. Note that consistent with the model there is little difference in the amount of trust between rule-following groups and rule-breaking groups. Unreported Wilcoxon tests cannot reject the null hypothesis of equal mean amount sent. However, the percent returned is higher in rule-following groups than in rule-breaking groups, as evidenced in figure 5, which plots the average amount returned by second movers to first movers

as a percent of the amount sent, for both group types in 3 bins.[24]

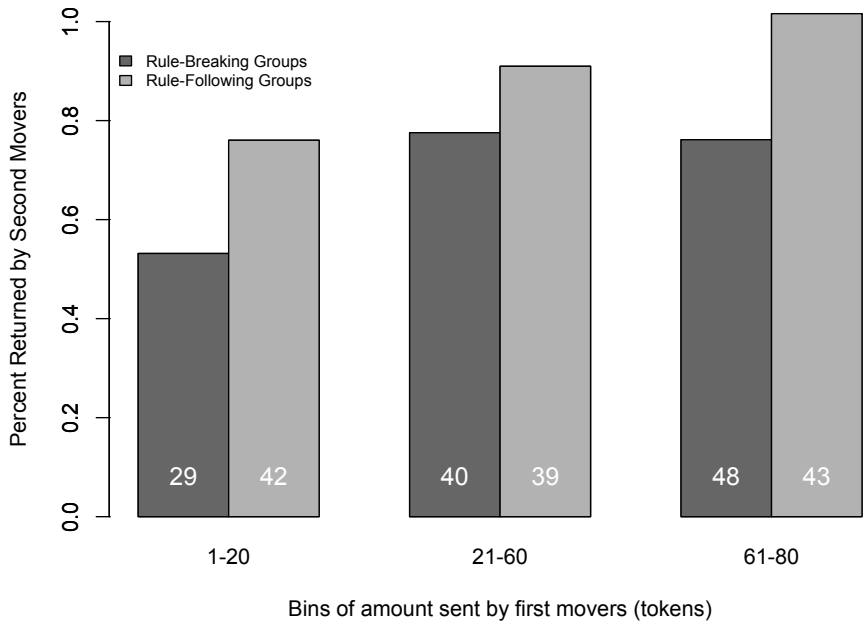

Figure 5: Barplots of $\frac{amount\_returned}{amount\_sent}$ by receivers in the TG treatment for 3 bins of amount sent. The white number in each bar displays the # of observations in the bin for that group type (i.e. for followers and breakers).

This is supported by Wilcoxon rank-sum tests of the null hypothesis of equal mean return on trust in rule-following and rule-breaking groups for each trial (1-3), where the first trial is defined as the first time a subject was in the role of first-mover, and observations are excluded where the first mover sent 0. In the first two trials, we reject the null hypothesis in favor of the alternative hypothesis that mean return on trust is higher in rule-following groups ($W_{43,42}$ = 752.5, p-value = 0.089 and $W_{42,37}$ = 633.5, p-value = 0.071, one-sided tests); however we cannot reject the null hypothesis for the third trial ($W_{39,38}$ = 698.5, p-value = 0.331, one-sided test).[25] Pooling the data and taking the mean return on trust for each subject over all three trials, another Wilcoxon test rejects the null hypothesis of equal mean returns ($W_{48,47}$ = 950.5, p-value = 0.092, one-sided test).

Finally, we note that our sample mean return is 25% of the amount received (75% of amount sent), and this is comparable to a previously observed mean of 20% for trust games in which subjects play both roles, reported in Table 1 of Johnson and Mislin (2011).

Finding 3: *In accordance with hypothesis 2, rule-followers exhibit greater positive reciprocity than rule-breakers in the trust game, and there are no differences in the amount sent.*

[24]In Appendix I we also report one other figure summarizing the full dataset from the TG treatment. Figure I2 shows, for each observation, the amount received by second-movers and the corresponding amounts returned and kept by group type.

[25]The number of observations changes because we only consider cases where first movers sent a positive amount.

4.2.2 Dictator Game Treatment

In the DG treatment, individuals were randomly assigned to their roles, so here we can exploit the entire distribution of RF task behavior to compare dictator giving for various percentile cutoffs defining rule-following and rule-breaking. As noted above, there is considerable noise in the measurement in the RF task, so it is reasonable to expect the effect size and significance to increase as we move further into the tails. With this fact in mind, panel (a) of Figure 6 displays mean amount sent by dictators of each type, where the type is defined by a percentile cutoff in the RF-task distribution. As we move right along the x-axis, we are moving further into the tails of the RF-task distribution; the final data points, comparing the top and bottom deciles of the distribution, contain 7 observations of each type.

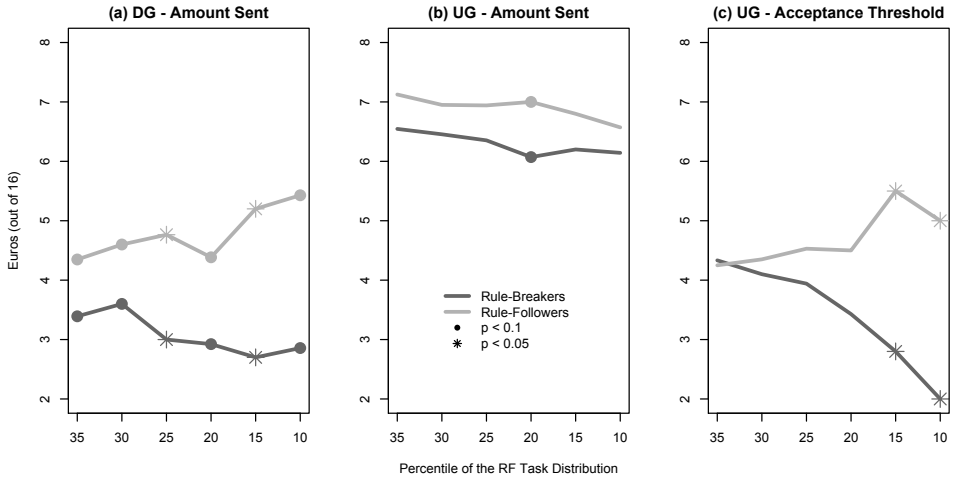

Figure 6: Each panel of the figure displays data relevant to one hypothesis, highlighting the comparison in behavior between rule-followers and rule-breakers. Panel (a) reports mean amounts sent by dictators in the DG treatment. Panel (b) reports mean amounts offered by proposers in the UG treatment. Panel (c) reports mean acceptance thresholds by responders in the UG treatment. The x-axis plots various percentile cutoffs defining rule-followers and rule-breakers; hence at the point labeled 25, we are comparing means for the upper and lower quartiles of the RF task distribution. At the point labeled 10, we are comparing deciles. Closed circles indicate that rule-follower means are significantly greater than rule-breaker means by a one-sided Wilcoxon test at the 0.1 level, and asterisks indicate significance at the 0.05 level.

At each reported cutoff, a Wilcoxon test rejects the null hypothesis of equal mean amount sent by rule-followers and rule-breakers in favor of our alternative hypothesis that rule-followers are more generous, and the magnitude of the difference grows as we move into the tails (as predicted by the model in Appendix A). When comparing the upper and lower decile of the distribution, rule-followers (so-defined) give nearly twice as much as rule-breakers! Moreover, if we simply classify as rule-followers all those who waited for at least 25 seconds in the RF task and as rule-breakers those who waited less than 25 seconds, our results remain essentially the same.

Finally, we note that our pooled sample of 67 dictators yields an average amount sent of 4.01 Euros or 25% of the total endowment, which is comparable to the mean of 28% reported in a

meta-study of dictator games (Engel, 2011).

Finding 4: *In accordance with hypothesis 3, rule-followers are more generous in dictator games than rule-breakers.*

4.2.3 Ultimatum Game Treatment

As in the DG treatment, we classify individuals as rule-followers and rule-breakers according to their behavior in stage 1, and we offer statistical comparisons for multiple percentile cutoffs. Recall that the model in Appendix A predicts no differences between high- and low-ϕ types in amount sent but predicts higher acceptance thresholds for high-ϕ types.[26] Again, due to the noise in our measure of rule-following (our proxy for ϕ), we would expect both the magnitude and significance of any such effect to increase as we move into the tails of the RF task distribution. Panel (b) of figure 6 displays the mean amount offered by rule-followers and rule-breakers in the UG treatment. Panel (c) displays mean acceptance thresholds.

As expected, there are no significant differences between rule-followers and breakers in amount sent (with the exception of a weakly significant and likely anomalous estimate at the quintile cutoff), though the mean is consistently 0.5 - 1 Euro higher among rule-followers. This provides strong evidence that giving in ultimatum games is largely a *strategic* decision. The fact that we observe the hypothesized difference in sender behavior between the DG and UG treatments is extremely important because it provides further evidence distinguishing our model from a model of explicit distributional preferences, which predicts a strong correlation between high offers in DG and UG.

Turning to second movers, as we move further into the tails of the RF-task distribution, the differences in acceptance threshold between rule-followers and rule-breakers become increasingly statistically and economically significant. For example, comparing the top and bottom decile of the waiting time distribution, rule-followers' mean threshold is 5 Euros and rule-breakers' is 2. This difference is statistically significant by a one-sided Wilcoxon Test ($W_{7,7}$ = 38.5, p-value = 0.032).

As in the DG treatment, our pooled sample mean offer from 69 proposers is comparable to figures reported in the UG literature. Proposers offer 43% of the endowment, compared to an average of 40% reported in a meta-study by Oosterbeek, Sloof, and Van De Kuilen (2004). Our pooled mean acceptance threshold set by responders is 27% of the endowment; unfortunately the meta-study does not report mean acceptance thresholds, though our observed threshold is comparable to the 33% reported in Table 2 of Schmitt, Shupp, Swope, and Mayer (2008).

[26] To be more precise, the model predicts little difference between high- and low-ϕ types in amount sent if the general belief in the population is that the rejection threshold is high. If everyone believes that the threshold is low, then high-ϕ individuals should send more than low-ϕ individuals (see Appendix A). We have weak evidence of this effect shown in panel (b) of Figure 6.

Finding 5: *In accordance with the first part of hypothesis 4, when we compare the most rule-following and the most rule-breaking individuals in the ultimatum game, we find that rule-followers set higher acceptance thresholds than rule-breakers. Moreover, as predicted, rule-followers and rule-breakers do not exhibit differences in amount sent.*

4.3 Individual Differences in the Rule-Following Task

A final feature of our design allows us to explore the determinants of RF-Task behavior at an individual level. Recall that we ran one NoRule-PG session with 24 subjects and a NoRule-Reverse-PG session with 24 subjects in which subjects in the first stage were not told that "the rule is to wait…" This allows us to distinguish the impact of the statement of the rule on waiting times from other factors that might influence waiting time (e.g. the context provided by our stop light task). Figure 7 displays histograms of waiting time by Rule/No-Rule treatment. Invoking a "rule" has a powerful impact on individual waiting times. Notably, when the rule is invoked, 62.5% of subjects spend at least 25 seconds waiting, indicating that they obey the rule without exception, though it costs them at least €2. Furthermore, average waiting time is 22.5 seconds, and many subjects who break the rule marginally while waiting at one or two of the five stop-lights nevertheless follow the rule in general. In the NoRule treatments, average waiting time is only 10.4 seconds, and only 12.5% of subjects wait at least 25 seconds. This suggests that the induced context of our RF task is responsible for some of the rule-following behavior we observe, but explicit statement of the rule plays a more important role.

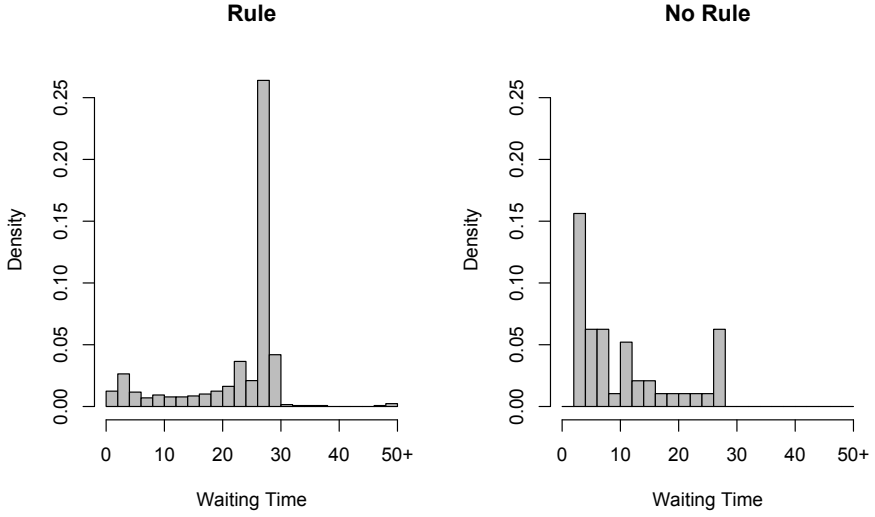

Figure 7: Histograms of Waiting Time in Seconds, Rule vs. No-Rule Treatments

At the end of each session all subjects answered the Moral Foundations Questionnaire designed to measure the strength of their respect for various moral values (Haidt and Joseph, 2004; Graham, Haidt, and Nosek, 2008). While the list is not necessarily exhaustive, the purpose is to

measure moral intuitions about the following five values: 1) aversion to doing *harm*; 2) concerns for justice or *fairness*; 3) love of country, family, and the *in-group*; 4) respect for *authority*; and 5) the desire for cleanliness and *purity*. Subjects answer 6 questions about each of these values using a Likert scale. We construct a score between 0 and 30 that represents the strength of their respect for each value. Table I1 in Appendix I summarizes the distribution of individual moral foundation scores pooled across treatments.

To explore other sources of individual differences in rule-following proclivity, we pool the data from all experimental sessions and report regression analysis explaining RF task behavior in terms of subjects' moral foundation scores with controls for demographic characteristics and the NoRule and Reverse treatments. The dependent variable is the total time the subject spent waiting at stop lights in the RF task, and the independent variables are subjects' scores for each of the five moral values, age, gender, dummy variables for the reverse-PG and NoRule treatments, an interaction dummy between NoRule and reverse-PG, field of study dummies, a dummy for non-European subjects, and a constant term. In the reverse-PG treatment, we also control for subjects' own mean contribution to the public good as well as the mean contribution of others in their group. Most of our subjects are business majors, so the field of study dummies indicate differences from the average business major. Note that we do not include a dummy for any other treatments since all other subjects were unaware of the details of the second stage when making their RF task decisions.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.808	9.232	3.012	0.003***
Reverse	0.066	3.487	0.019	0.985
NoRule	-10.291	1.412	-7.290	0.000***
Harm	-0.122	0.108	-1.136	0.257
Fairness	0.099	0.111	0.891	0.373
InGroup	-0.020	0.107	-0.191	0.849
Authority	0.014	0.104	0.131	0.896
Purity	0.090	0.093	0.963	0.336
Age	-0.003	0.004	-0.709	0.478
Female	2.554	0.789	3.238	0.001***
Non-European	-0.176	1.137	-0.155	0.877
Economics	-1.138	0.897	-1.270	0.205
Law	-0.862	1.647	-0.524	0.601
Psych	0.734	1.596	0.460	0.646
Other	-1.051	1.086	-0.967	0.334
Reverse_Contrib	-0.010	0.092	-0.103	0.918
Reverse_Others'_Contrib	-0.055	0.044	-1.255	0.210

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

N = 600; F-Statistic = 7.47, p-value < 0.01

Table 5: Determinants of Waiting Time, OLS Regression

Table 5 reports the estimation results. First, note again that subjects are substantially more likely to break the rule in the NoRule treatment than in the other treatments, which indicates that an explicitly stated verbal rule, with no strings attached, is sufficient to induce rule following. Second, we find that female subjects are less likely to break the rule than their male counterparts, and that age has no noticeable effect on rule breaking. Women are also less likely to cross at red lights in observational studies of pedestrian behavior in Amman, Jordan and Tel-Aviv, Israel (Hamed, 2001; Rosenbloom, 2009). In Amman, age is also negatively correlated with crossing, but because our sample consists of university students, our data may lack the variability necessary to identify an effect. Of course, age is likely to matter far less in a simulated environment because it no longer correlates with the ability to quickly cross the road. Our field-of-study dummies and the non-European dummy are insignificant. Surprisingly, none of the moral values scores are correlated with rule-following; we expected, at the least, that the instrument measuring respect for authority would correlate with RF-task behavior.

4.4 Discussion

In the last 30 years, there has been a proliferation of research on cooperative and altruistic behavior in the laboratory. In a variety of games, there is extensive evidence of prosocial behavior that is inconsistent with money-maximizing Nash play (see e.g. Cooper and Kagel, 2013, for a summary). To explain these observations, it has been proposed that these experiments reveal other-regarding preferences (e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). While these models have proved useful for categorizing behavior, there are a number of anomalous results for which they cannot account. In particular, the sensitivity of outcomes to choice context—a stylized fact reported in Hoffman, McCabe, and Smith (1996), Cherry, Frykblom, and Shogren (2002), List (2007), and Andreoni and Bernheim (2009), among others—is not consistent with any model in which preferences are defined over own and other payoffs. Varying the assignment of action rights, the visibility of actions, and the choice set itself all demonstrably impact measured pro-sociality, and in each of the aforementioned papers, the authors suggest that observed behaviors may be driven by social norms which have been activated by analogy between the lab setting and previous experiences and imported into the laboratory by subjects.

Our paper formalizes this idea and argues that norm-dependent preferences, in which individuals care not about particular payoff distributions but rather about whether their behavior corresponds to some social norm, provide a unifying explanation for sociality across laboratory games, contexts and cultures. We test this model in a variety of new experiments. In the model, based on the work of Kessler and Leider (2012), an individual's utility depends not only on the material payoff he receives, but also on the distance between the chosen action and the relevant social norm. Thus, since different contexts evoke different norms, behavior differs across con-

texts, and most importantly for us, differences in individual behavior *within* a context are driven by differences in the extent to which individuals care about social norms.

Thus, it is crucial to our objective to develop a means of measuring the extent to which individuals care about social norms *before* observing their play in the relevant games. Our extremely simple RF task, which provides us with a continuous measure of the extent to which individuals are willing to follow an arbitrary, costly rule, serves precisely this purpose. And our data from public goods, trust, dictator and ultimatum games indicate that norm-dependent utility and the accompanying concept of norm reciprocity accurately characterizes behavioral heterogeneity in both repeated and one-shot games. Importantly, behavior from our pooled samples is comparable to that observed in the literature (see Section 4).

Our main treatment focuses on one of the best studied of these games, the repeated public goods game, a well-known feature of which is a pattern of declining contributions over time. As with context effects in one-shot games, this observation is inconsistent with many models of other-regarding preferences, in which contributions should be insensitive to repetition. Numerous attempts have been made to explain these observations (see e.g. Andreoni, 1995a; Houser and Kurzban, 2002, for a lively discussion of whether contributions are driven by kindness or confusion), and one promising line of research has focused on the presence of *types* such as conditional and unconditional cooperators (Fischbacher, Gächter, and Fehr, 2001; Kurzban and Houser, 2005). Extending this line of argument, Fischbacher and Gächter (2010) propose that declining contributions in repeated public goods games are driven by individual *preferences* for conditional cooperation.

Our model provides an alternative way of thinking about conditionality: we argue that people are not specifically conditional *cooperators*, rather, they are conditional *norm-followers*, where the norm may or may not require cooperation. In particular, when we model play in repeated games, we assume that individuals are not blind adherents who will follow norms when others choose not to; when individuals observe that others refuse to adhere to norms, they do the same. The data from our PG treatments also confirm this aspect of the model. When we sort individuals without their knowledge into groups of rule-followers and rule-breakers, we observe declining contributions to the public good only among rule-breakers. Rule-following groups sustain high levels of contribution for the entire experiment!

In other experiments, the absence of 'cooperative decay' has generally been observed in outlying cases or with the introduction of communication and/or punishment (Isaac and Walker, 1988; Fehr and Gächter, 2000a; Bochet, Page, and Putterman, 2006; Kosfeld, Okada, and Riedl, 2009; Xiao and Houser, 2011).[27] Indeed, when we randomly assigned individuals to groups, we once again observed cooperative decay, and rule-followers' contributions were indistinguishable

[27]Exceptions to this rule exist; for example, subjects can achieve sustained cooperation when they make binding, incremental, publicly observable contributions in real-time (Kurzban, McCabe, Smith, and Wilson, 2001). Similarly, allowing individuals to form their own groups increases average contributions, but there is still a tendency for contributions to decline over time (Page, Putterman, and Unel, 2005).

from rule-breakers' overall, providing support for norm-reciprocity.

Existing models of reciprocity assume that players' prosocial behavior is conditional on others being "(un)fair" or "(un)kind."[28] These models start from assumptions regarding the distributional preferences of the players, which allow them to evaluate at each node of the game how "good" or "bad" the previous choices of others were with respect to distributional goals. As Charness and Rabin put it, "any reciprocal model must embed assumptions about distributional preferences" (p. 824). These evaluations then evoke reciprocal responses, changing the social preferences of the players so that they may be willing to sacrifice their own material payoff to either benefit or harm others. These models are capable of explaining some stylized observations in a variety of games, but there are substantial differences from our approach. First, our model need not assume that players have preferences over a specific distribution of payoffs; reciprocity results from players observing others' (non)-adherence to the norm, whatever it may be. Moreover, reliance on fixed assumptions about distributional preferences means that other models cannot explain why behavior differs across contexts, despite extensive evidence that this is the case. Most importantly, earlier models say nothing about why our rule-following task would be related to reciprocal (or for that matter, conditionally cooperative) proclivities: our model explains this connection directly.

A reader may object that the norm in our model is exogenous and thus any behavior can be rationalized by the model, while reciprocal behavior in the reciprocity models emerges "endogenously" in response to past actions. As a technical statement this is true (though it is also technically true that norm-reciprocity is an endogenous response in our model). More importantly, such arguments miss a broader philosophical point that, in fact, makes us wish to emphasize the explicit exogeneity of the norm here. Much of the motivation of the social preferences literature has derived from a desire to explain pro-social behavior, or "fairness." As Wilson (2012) has argued, the evaluation of an act as either pro-social or anti-social, fair or unfair, relies on the presumption of a shared conception of pro- (and anti-) sociality among agents; the word "social" gives this away – and this shared conception is necessarily *outside* of the model in most models of other-regarding preferences. In fact, it is at best implicit in the form of the utility function specified; for example, models based on inequity aversion assume norms of equity. To paraphrase Wilson, the following conversation fits neatly into the social preferences framework:

> *Economist 1:* Well, what did she do?
> *Economist 2:* She did what was *fair*.
> *Economist 1:* Why did she do that?
> *Economist 2:* To be fair!

The language of social preferences tends to circularly treat fairness as both a motivation *and*

[28] See e.g. Fehr and Gächter (2000b); Charness and Rabin (2002); Dufwenberg and Kirchsteiger (2004); Falk and Fischbacher (2006); Cox, Friedman, and Gjerstad (2007); Cox, Friedman, and Sadiraj (2008); Battigalli and Dufwenberg (2009).

an outcome, when in reality, "[the] rules that guide fair behavior are not located in an individual's private utility function but instead reside in the connections that the individual has to his cultural environs" (Wilson, 2012, p. 407). Thus, a model of norm-dependent preferences explicitly acknowledges that norms and conventions are *external* to the individual, while remaining formally agnostic on what sorts of actions may be normative in any particular environment.[29] Moreover, it highlights what we believe is the most important question going forward: how do we identify norms and account for their creation, spread, and evolution?

Recent work by Krupka and Weber (2013) and Gächter, Nosenzo, and Sefton (2013) has begun to answer the first part of this question. For example, relying on a theoretical model similar to ours in which behavior depends on norms and on unobserved heterogeneity in norm-following preferences, Krupka and Weber (2013) use simple coordination games to identify social norms,[30] and they find that differences in elicited norms predict differences in behavior across contexts in simple dictator games. Future work should combine their method of identifying norms with our method of measuring preference heterogeneity to test the more precise predictions of models of norm-following preference.

4.4.1 Intentionality and Mind-Reading

It is also interesting to reconsider the literature on "intentionality" and "theory of mind" in light of the model and experimental findings. From a variety of experiments, there is evidence that "intentions matter" when deciding whether to engage in positive or negative reciprocity (McCabe, Smith, and LePore, 2000; McCabe, Rigdon, and Smith, 2003; Falk, Fehr, and Fischbacher, 2008; Smith and Wilson, 2013). When second movers know that a cooperative first-mover has incurred an opportunity cost in order to take a cooperative action, they are more likely to positively reciprocate, and absence of generosity is more likely to be punished when the only alternative to non-generous behavior is kindness. If instead, the non-generous individual *also* had the opportunity to harm their counterpart, lack of generosity is met with indifference. All these findings can be accommodated by considering the norms induced by the exogenous manipulation of the choice set. Indeed one way of thinking about social cognition is that an important part of "theory of mind" is the ability to infer social norms from context.[31]

[29] As with fairness, judgments about what constitutes a 'good' or a 'bad' action, which one might (or ought to) reciprocate, are necessarily outside the model because they too rely on a shared sociality (see Wilson, 2008). Our model resolves this issue by embedding assumptions about preferences for adhering to socially defined norms, not for particular distributions. Agents in our model exhibit *norm-reciprocity*, a tendency to follow norms only when others do so as well. Given some norm, an action is judged on whether it was consistent with the norm. This allows us to model reciprocity in a variety of novel contexts while again remaining agnostic as to what behaviors merit a reciprocal response in any particular case.

[30] Participants are given a list of actions and asked to identify how they think *others* will evaluate each action's "social appropriateness". They are paid if their answers correspond to the modal answer in their anonymous group.

[31] Findings from neuroeconomics suggest that the cognitive mechanisms underlying cooperation and reciprocity rely on the parts of the brain associated with social exchange (McCabe, Houser, Ryan, Smith, and Trouard, 2001; Kosfeld, Heinrichs, Zak, Fischbacher, and Fehr, 2005; Knoch, Pascual-Leone, Meyer, Treyer, and Fehr, 2006). Thus,

4.4.2 Behavioral Types

Other recent research has demonstrated that experimental decisions can identify behavioral *types* (e.g. McCabe, Houser, Ryan, Smith, and Trouard, 2001; Houser, Keane, and McCabe, 2004; Kurzban and Houser, 2005; Wilson, Jaworski, Schurter, and Smyth, 2012) and that this information can be used to sustain cooperation among a subset of experimental subjects. For example, Gunnthorsdottir, Houser, and McCabe (2007) regroup subjects in public goods games according to their initial contributions and find that assortative matching supports cooperation over time. Similarly, Rigdon, McCabe, and Smith (2007) show that endogenous sorting of cooperators in a repeated trust game sustains cooperation among the positively sorted. In general, behavioral typing from experimental data relies on early-period decisions in the relevant experiment to classify types, which may confound interpretation of the results.[32] We also sort our subjects by type without their knowledge, but instead of identifying types based on early decisions in the repeated game, we use unrelated behavior from our RF task to perform our sorting. Subjects decide to what extent they will follow the rule in private, without knowledge of the behavior of others and without knowledge of the second stage of the experiment. Combining that with the fact that our task proved predictive across four of the major experiments used to study other-regarding behavior, we argue that this provides strong evidence that our RF task measures a quite fundamental characteristic of individuals.[33]

4.4.3 What Do We Know About Norm-Following?

An important issue for future research will be to explain *why* people care about norms. One candidate comes from Bernheim (1994); following this argument one might argue that actions in accordance with a norm provide a signal of status, and thus people follow norms to gain status. Indeed, evidence on "audience effects" indicates that behavior is affected by how an agent's actions are perceived by others (Kurzban, DeScioli, and O'Brien, 2007; Andreoni and Bernheim, 2009). However, while these models provide a proximate explanation for norm-following, they do not explain *why* norm-following signals status or why individuals seek status. Recent work in economics by Alger and Weibull (Forthcoming) offers an evolutionary foundation for preferences for following one particular kind of norm. Similarly, psychologists have recently argued for a culture-gene co-evolutionary origin of norm-following, in which a species that becomes suf-

people who are in upper tail of the autism spectrum may provide an interesting boundary case for the model, since they are often capable of following rationally articulated rules but not of easily identifying and following social norms. E.g. there is evidence that people with autism spectrum disorders are less sensitive to "audience effects" in charitable decision-making (Izuma, Matsumoto, Camerer, and Adolphs, 2011).

[32] One interesting exception is Rietz, Sheremeta, Shields, and Smith (forthcoming) who implement a surprise restart of the experiment after a one-shot game and use behavior in the first game to type subjects in a repeated version of the same game. See also Gunnthorsdottir, McCabe, and Smith (2002) who use a 'Machiavellianism' survey instrument to type individuals in a one-shot trust game.

[33] In subsequent research, we have also found evidence that the RF task predicts cooperation in a common pool resource game (Kimbrough and Vostroknutov, 2013).

ficiently reliant on culturally transmitted information will tend to evolve a "norm psychology" to reinforce transmission. This psychology then interacts with cultural group selection to influence the relative frequency of certain cultural phenotypes within and between groups (Chudek and Henrich, 2011). Whatever the source, the idea that people care about the social expectations embodied in norms has a long history among economists. The same intuition is personified in Adam Smith's notion of the "impartial spectator" through whose eyes individuals evaluate the propriety, not of outcomes, but of *actions*. In this light, he discussed the desire to conform to norms and conventions:

> [A person] desires, not only praise, but praise-worthiness; or to be that thing which, though it should be praised by nobody, is, however, the natural and proper object of praise. He dreads, not only blame, but blame-worthiness; or to be that thing which, though it should be blamed by nobody, is, however, the natural and proper object of blame. (Smith, 1759, §. 3.1)

Similarly, David Hume anticipated the fundamental *conditionality* of norm-following behavior (what he calls "justice"):

> Taking any single act, my justice may be pernicious in every respect; and *'tis only upon the supposition, that others are to imitate my example*, that I can be induc'd to embrace that virtue [...] (cf. Hume, 1740, §3.2.2, *italics added*)

These ideas also have precedent in the literature from social psychology on the "role-rule" model of human social behavior, which argues that many decisions can be explained by assuming that people are trying to "play by the rules," where the rules are determined by the individual's perceived role in the interaction (Harré and Secord, 1972). Moreover, the fact that individuals were willing to incur costs in obedience to an arbitrary rule comes as no surprise. In previous experiments, subjects have exhibited costly obedience to experimenters, even when it meant administering "painful" punishment to others, as in the famous Milgram experiment and numerous replications (Milgram, 1963; Zimbardo, 2007). Implicit in our argument is that any such "experimenter demand" effect is motivated by a desire to conform to perceived norms of laboratory behavior.

Moreover, the fact that many individuals who exhibited declining contributions in our "rule-breaking" and unsorted groups were not themselves gross violators of the rule could be viewed as an instance of the "broken windows" effect (in which individuals who observe violations of a social norm are more likely to violate the same norm, see Wilson and Kelling, 1982; Keizer, Lindenberg, and Steg, 2008). Our norm-reciprocity model captures precisely the intuition of the broken windows effect.[34]

[34]This is also consistent with evidence that individuals tend to conform to the (implicit or explicit) norms established by those whose actions they observe (Frey and Meier, 2004; Bardsley and Sausgruber, 2005; Alpizar, Carlsson, and Johansson-Stenman, 2008; Bicchieri and Xiao, 2008).

4.4.4 Some Practical Implications

Outside the lab, others have argued for the crucial role of norms in economics. An extensive literature in sociology, going back at least to Granovetter (1973), discusses the role of norms in sustaining cooperation within networks, and Granovetter (2005) discusses the implications of this perspective for economics more generally. Similarly, norms feature prominently in Greif's (2006) account of the emergence of modern institutions. In his telling, an important function of many institutions is to instill and reinforce mutually beneficial norms of behavior. Thus norms have instrumental value in supporting wealth creation.

However, our findings imply that the total social value of many rules and norms may be *underestimated* when we consider only their instrumental value, i.e. their value in resolving particular coordination problems, reducing transactions costs and so on. In fact it is possible that even seemingly costly and arbitrary practices may persist simply because they allow others to screen for cooperators. Applying the logic of Demsetz (1967), we would expect that even if rules and norms impose direct costs on their adherents, they can persist as long as the value of the screening function they provide exceeds this cost.[35]

There is a clear role for screening mechanisms to promote cooperation in human social and economic relations, as it is well known that many potential mutually beneficial transactions are plagued by incentive problems. Despite these incentive problems, many people are involved in organizations and groups that provide local public (club) goods, in-group risk sharing, and so on. If individuals have some means of identifying norm-following types *ex ante*, then many otherwise incentive incompatible cooperative endeavors become feasible. Our argument implies that imposing costly rules could provide one way in which a prospective cooperator might use observable behavior to identify similar others.

This idea too has precedent. In the literature on the economics of religion, it has been argued that religious strictures regarding the choice of food items and articles of clothing may act as screening mechanisms (similar to our costly RF task) that allow members of religious groups to distinguish sincere prospective members from free riders (Iannaccone, 1992).[36] By imposing a cost on entrants, these groups are able to maintain a high level of public (or club) good provision for their current members. In a recent experiment, Aimone, Iannaccone, Makowsky, and Rubin (forthcoming) test this hypothesis directly in a public goods game with endogenous group formation in which the cost of joining various groups differs. The authors find that individuals who join groups with higher entry costs also contribute more to the public good, which suggests that costly screening mechanisms can effectively promote cooperation, even in an environment where some individuals could incur the cost strategically to gain access to a valuable,

[35]One might ask us to perform a cost-benefit analysis on rule-following in our own experiment. However, such an analysis doesn't make sense for two reasons: 1) individuals are unable to make such calculations in the rule-following task because they have no foreknowledge of the second stage, and 2) the relative costs and benefits are arbitrarily chosen by the experimenter.

[36]See also Gintis, Smith, and Bowles (2001).

high-contribution group.

5 Conclusion

We explore a unifying framework for prosocial behavior in which individuals trade off own payoffs against a desire to adhere to social norms. To illustrate the argument, we derive some implications of a simple model of norm-dependent preferences and extend it to a dynamic setting to account for conditional cooperation and related phenomena through "norm-reciprocity", wherein individuals prefer to adhere to social norms only insofar as they observe others doing the same. We argue that our explanation both encompasses and supersedes earlier models based on explicit distributional preferences. The model accounts for many of the observations that these models were developed to explain without relying on particular assumptions about distributional preferences, and it accounts for a variety of observations that cannot be explained with these models, such as context effects and cooperative decay in repeated games. However, as we note above, these two types of model are not mutually exclusive because distributional preference models are implicit versions of a norm-dependent model where the social norm is assumed into the form of the utility function. The model we present can be rewritten to reflect any standard model of distributional preferences, but it is more general because the norm is allowed to vary across contexts.

We develop an extremely simple experiment that allows us to measure a proxy for the crucial parameter of the model which reflects the extent to which an individual cares about norms. Using this information, we show that norm-following proclivity predicts play in dictator, ultimatum, trust and public goods games, providing strong support for the idea that heterogeneous play in these games is driven by heterogeneous attitudes toward social norms.

One important avenue for future work will be to further explore the robustness of our norm-sensitivity measurement experiment, in which subjects are instructed to follow a costly rule. First, since our experiment indicates that individuals are extremely responsive to a simple statement that "the rule is…" it will be useful to develop context-minimal screening tasks to reduce noise in the proxy for ϕ. Second it will also be interesting to investigate whether there are *within-subjects* differences in the responsiveness to exogenous manipulation of context, consistent with the norm-dependent utility framework. We have begun work on this issue in a second paper which generalizes the model reported here (Kimbrough, Miller, and Vostroknutov, 2013).

The most important unanswered question, though, and the one that we hope this research will encourage others to ask, is "where do norms come from?" Norms are exogenous in our model, but we could also think of them as being specific to an individual's identity, group membership, or culture (this is one way to view Akerlof and Kranton, 2000). To the extent that this is true, the same experimental procedures may induce entirely different norms, depending on the cultural background of the subjects. Thus, since we know that high-ϕ individuals care about

norms, observing their behavior may reveal perceived norms, where the norm is uncertain or ambiguous.

Fundamentally, if the social norms associated with a particular context differ across cultures, then cross-cultural behavioral differences in laboratory experiments also come as no surprise – and we can explain these differences while maintaining that people facing different norms nevertheless have the same underlying motivations.[37]

As Wilson (2008) has argued, "in general, cooperative outcomes are the product of human agreement, tacit or otherwise, on the social context of the interaction" (p. 374). Since our experimental environment suppresses communication, any agreement on the norms of action is necessarily tacit, and the extent of tacit agreement is likely tied to the extent to which prospective cooperators share a cultural/experiential background. Although our subject pool contains individuals from a large number of nations, the preponderance of our subjects hail from Western Europe and were raised according to the rules and norms common to European culture(s). This likely encouraged cooperation among our high-ϕ types by increasing agreement about the norm. Note that this argument has interesting implications for the literature on diversity and social trust. For example, Putnam (2007) has shown that levels of trust are negatively correlated with diversity in a sample of 41 US regions/metro areas. Under our model, if two (or more) parties disagree about the norm, they may fail to cooperate *even though* both want to follow norms that are prosocial to some degree. This is an important issue for future research that we hope to pursue.

In a similar vein, Roth et al.'s (1991) findings that subject behavior differs across cultures in non-market contexts—but is essentially the same in market contexts—has interesting implications. One interpretation is that markets are norm-free, that is, behavior in markets is culturally invariant because markets work around or outside of normative concerns.[38] Another interpretation is that the norms associated with markets are common across cultures. This is likely true to some extent, though clearly there are cultural differences in the types of things that are viewed as commodities and the kinds of market transactions that are deemed acceptable (Roth, 2007). However, both hypotheses are difficult to reconcile with the evidence in Henrich et. al (2010) that greater exposure to markets and to large-scale institutions such as organized religion are both correlated with experimental measures of other-regarding and cooperative behavior. Instead, one might argue that certain norms are *embedded* in market institutions, and they are transmitted (epidemiologically?) through repeated interaction. We leave these questions for future research.

[37]See e.g. Roth, Prasnikar, Okuno-Fujiwara, and Zamir (1991); Henrich, Boyd, Bowles, Camerer, Fehr, Gintis, McElreath, Alvard, Barr, Ensminger, Smith Henrich, Hill, Gil-White, Gurven, Marlowe, Patton, and Tracer (2005); Herrmann, Thöni, and Gächter (2008); Henrich, Heine, Norenzayan, et al. (2010).

[38]This is related to the argument in Fehr and Schmidt (1999), where the effects of social preferences in their model dissipate when individuals are small relative to the market; competition limits the effectiveness of prosocial action.

References

ACEMOGLU, D., AND M. O. JACKSON (2012): "History, Expectations, and Leadership in the Evolution of Social Norms," MIT Working Paper.

AIMONE, J., L. R. IANNACCONE, M. MAKOWSKY, AND J. RUBIN (forthcoming): "Endogenous group formation via unproductive costs," *Review of Economic Studies*.

AKERLOF, G. A., AND R. E. KRANTON (2000): "Economics and Identity," *Quarterly Journal of Economics*, 115(3), 715–753.

ALGER, I., AND J. WEIBULL (Forthcoming): "*Homo Moralis* - Preference Evolution under Incomplete Information and Assortative Matching," *Econometrica*.

ALPIZAR, F., F. CARLSSON, AND O. JOHANSSON-STENMAN (2008): "Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica," *Journal of Public Economics*, 92(5), 1047–1060.

ANDERSEN, S., S. ERTAÇ, U. GNEEZY, M. HOFFMAN, AND J. A. LIST (2011): "Stakes Matter in Ultimatum Games," *American Economic Review*, 101(7), 3427–39.

ANDREONI, J. (1995a): "Cooperation in Public-Goods Experiments: Kindness or Confusion?," *American Economic Review*, 85(4), 891–904.

——— (1995b): "Warm-Glow Versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments," *Quarterly Journal of Economics*, 110(1), 1–21.

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.

ANDREONI, J., AND J. MILLER (2002): "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica*, 70.

AXELROD, R. (1986): "An Evolutionary Approach to Norms," *American Political Science Review*, pp. 1095–1111.

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

BARDSLEY, N., AND R. SAUSGRUBER (2005): "Conformity and reciprocity in public good provision," *Journal of Economic Psychology*, 26(5), 664–681.

BATTIGALLI, P., AND M. DUFWENBERG (2009): "Dynamic psychological games," *Journal of Economic Theory*, 144, 1–35.

BERG, J., J. DICKHAUT, AND K. MCCABE (1995): "Trust, reciprocity, and social history," *Games and Economic Behavior*, 10(1), 122–142.

BERNHEIM, B. D. (1994): "A Theory of Conformity," *Journal of Political Economy*, pp. 841–877.

BICCHIERI, C., AND E. XIAO (2008): "Do the right thing: but only if others do so," *Journal of Behavioral Decision Making*, 22(2), 191–208.

BOCHET, O., T. PAGE, AND L. PUTTERMAN (2006): "Communication and punishment in voluntary contribution experiments," *Journal of Economic Behavior & Organization*, 60(1), 11–26.

BOLTON, G. E., AND A. OCKENFELS (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90(1), 166–193.

BURKS, S., J. CARPENTER, AND E. VERHOOGEN (2003): "Playing both roles in the trust game," *Journal of Economic Behavior & Organization*, 51(2), 195–216.

BURNHAM, T., K. MCCABE, AND V. L. SMITH (2000): "Friend-or-foe intentionality priming in an extensive form trust game," *Journal of Economic Behavior & Organization*, 43, 57–73.

CHARNESS, G. B., AND M. RABIN (2002): "Understanding Social Preferences With Simple Tests," *Quarterly Journal of Economics*, 117(3), 817–869.

CHERRY, T. L., P. FRYKBLOM, AND J. F. SHOGREN (2002): "Hardnose the Dictator," *American Economic Review*, 92(4), 1218–1221.

CHUDEK, M., AND J. HENRICH (2011): "Culture–gene coevolution, norm-psychology and the emergence of human prosociality," *Trends in Cognitive Sciences*, 15(5), 218–226.

COOPER, D. J., AND J. KAGEL (2013): "Other Regarding Preferences: A Selective Survey of Experimental Results," forthcoming in *The Handbook of Experimental Economics, Vol. 2*, eds. Kagel, J. and Roth, A.

COX, J. C., D. FRIEDMAN, AND S. D. GJERSTAD (2007): "A Tractable Model of Reciprocity and Fairness," *Games and Economic Behavior*, 59(1), 17–45.

COX, J. C., D. FRIEDMAN, AND V. SADIRAJ (2008): "Revealed Altruism," *Econometrica*, 76(1), 31–69.

DEMSETZ, H. (1967): "Toward a theory of property rights," *American Economic Review*, 57(2), 347–359.

DREBER, A., T. ELLINGSEN, M. JOHANNESSON, AND D. G. RAND (2012): "Do people care about social context? Framing effects in dictator games," *Experimental Economics*, pp. 1–23.

DUFWENBERG, M., AND G. KIRCHSTEIGER (2004): "A theory of sequential reciprocity," *Games and Economic Behavior*, 47, 268–298.

ECKEL, C., AND P. GROSSMAN (1996): "Altruism in anonymous dictator games," *Games and Economic Behavior*, 16, 181.

ELLICKSON, R. C. (1989): "A Hypothesis of Wealth-Maximizing Norms: Evidence from the Whaling Industry," *Journal of Law, Economics & Organization*, 5(1), 83–97.

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

ENGEL, C. (2011): "Dictator games: A meta study," *Experimental Economics*, 14(4), 583–610.

ENGEL, C., AND M. KURSCHILGEN (2013): "The "Jurisdiction Of The Man Within": Intrinsic Norms in a Public Goods Experiment," mimeo, Max Planck Institute for Research on Collective Goods.

ENGELMANN, D., AND M. STROBEL (2004): "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments," *American Economic Review*, 94(4), 857–869.

FALK, A., E. FEHR, AND U. FISCHBACHER (2008): "Testing theories of fairness – Intentions matter," *Games and Economic Behavior*, 62, 287–303.

FALK, A., AND U. FISCHBACHER (2006): "A theory of reciprocity," *Games and Economic Behavior*, 54, 293–315.

FEHR, E., AND U. FISCHBACHER (2004): "Third-party punishment and social norms," *Evolution and human behavior*, 25(2), 63–87.

FEHR, E., AND S. GÄCHTER (2000a): "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 90(4), 980–994.

——— (2000b): "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives*, 14(3), 159–181.

FEHR, E., AND K. M. SCHMIDT (1999): "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114(3), 817–868.

FISCHBACHER, U. (2007): "z-Tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, 10(2), 171–178.

FISCHBACHER, U., AND S. GÄCHTER (2010): "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments," *American Economic Review*, 100(1), 541–556.

FISCHBACHER, U., S. GÄCHTER, AND E. FEHR (2001): "Are people conditionally cooperative? Evidence from a public goods experiment," *Economics Letters*, 71(3), 397–404.

FISMAN, R., S. KARIV, AND D. MARKOVITS (2007): "Individual Preferences for Giving," *American Economic Review*, 97(5), 1858–1876.

FORSYTHE, R., J. HOROWITZ, N. SAVIN, AND M. SEFTON (1994): "Fairness in simple bargaining experiments," *Games and Economic behavior*, 6(3), 347–369.

FREY, B., AND S. MEIER (2004): "Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment," *American Economic Review*, pp. 1717–1722.

GÄCHTER, S., D. NOSENZO, AND M. SEFTON (2013): "Peer Effects in Pros-Social Behavior: Social Norms of Social Preferences?," *Journal of the European Economic Association*, 11(3), 548–573.

GINTIS, H., E. SMITH, AND S. BOWLES (2001): "Costly signaling and cooperation," *Journal of Theoretical Biology*, 213(1), 103–119.

GOEREE, J., AND C. HOLT (2001): "Ten little treasures of game theory and ten intuitive contradictions," *American Economic Review*, pp. 1402–1422.

GRAHAM, J., J. HAIDT, AND B. NOSEK (2008): *The Moral Foundations Quiz*. www.yourmorals.org.

GRANOVETTER, M. S. (1973): "The Strength of Weak Ties," *American Journal of Sociology*, pp. 1360–1380.

——— (2005): "The impact of social structure on economic outcomes," *The Journal of Economic Perspectives*, 19(1), 33–50.

GREIF, A. (2006): *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*, (Political Economy of Institutions and Decisions). Cambridge University Press.

GUNNTHORSDOTTIR, A., D. HOUSER, AND K. MCCABE (2007): "Disposition, history and contributions in public goods experiments," *Journal of Economic Behavior & Organization*, 62(2), 304–315.

GUNNTHORSDOTTIR, A., K. MCCABE, AND V. SMITH (2002): "Using the Machiavellianism instrument to predict trustworthiness in a bargaining game," *Journal of Economic Psychology*, 23(1), 49–66.

GÜTH, W., R. SCHMITTBERGER, AND B. SCHWARZE (1982): "An experimental analysis of ultimatum bargaining," *Journal of Economic Behavior & Organization*, 3(4), 367–388.

HAIDT, J., AND C. JOSEPH (2004): "Intuitive ethics: How innately prepared intuitions generate culturally variable virtues," *Daedalus*, 133(4), 55–66.

HALEVY, Y., AND M. S. PETERS (2009): "Other-Regarding Preferences: Outcomes, Intentions, or Interdependence," mimeo.

HAMED, M. (2001): "Analysis of pedestrians' behavior at pedestrian crossings," *Safety Science*, 38(1), 63–82.

HARRÉ, R., AND P. SECORD (1972): *The Explanation of Social Behaviour*. New York, Rowman & Littlefield.

HAYEK, F. A. (1973): *Law, Legislation and Liberty, vol. 1: Rules and Order*. The University of Chicago Press.

HENRICH, J., R. BOYD, S. BOWLES, C. CAMERER, E. FEHR, H. GINTIS, R. MCELREATH, M. ALVARD, A. BARR, J. ENSMINGER, N. SMITH HENRICH, K. HILL, F. GIL-WHITE, M. GURVEN, F. W. MARLOWE, J. Q. PATTON, AND D. TRACER (2005): ""Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies," *Behavioral and Brain Sciences*, 28, 795–855.

HENRICH, J., J. ENSMINGER, R. MCELREATH, A. BARR, C. BARRETT, A. BOLYANATZ, J. CARDENAS, M. GURVEN, E. GWAKO, N. HENRICH, ET AL. (2010): "Markets, religion, community size, and the evolution of fairness and punishment," *Science*, 327(5972), 1480–1484.

HENRICH, J., S. J. HEINE, A. NORENZAYAN, ET AL. (2010): "The weirdest people in the world," *Behavioral and Brain Sciences*, 33(2-3), 61–83.

HERRMANN, B., C. THÖNI, AND S. GÄCHTER (2008): "Antisocial Punishment Across Societies," *Science*, 319, 1362–1367.

HOFFMAN, E., K. MCCABE, K. SHACHAT, AND V. SMITH (1994): "Preferences, Property Rights, and Anonymity in Bargaining Games," *Games and Economic Behavior*, 7(3), 346–380.

HOFFMAN, E., K. MCCABE, AND V. L. SMITH (1996): "Social Distance and Other-Regarding Behavior in Dictator Games," *American Economic Review*, 86(3), 653–60.

HOFFMAN, E., AND M. L. SPITZER (1985): "Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice," *Journal of Legal Studies*, 14, 259–298.

HOUSER, D., M. KEANE, AND K. MCCABE (2004): "Behavior in a dynamic decision problem: An analysis of experimental evidence using a Bayesian type classification algorithm," *Econometrica*, 72(3), 781–822.

HOUSER, D., AND R. KURZBAN (2002): "Revisiting kindness and confusion in public goods experiments," *The American Economic Review*, 92(4), 1062–1069.

HUME, D. (1740): *A Treatise of Human Nature*. Oxford: Oxford University Press, (2003) edn.

IANNACCONE, L. (1992): "Sacrifice and stigma: reducing free-riding in cults, communes, and other collectives," *Journal of Political Economy*, pp. 271–291.

ISAAC, R., AND J. WALKER (1988): "Group size effects in public goods provision: The voluntary contributions mechanism," *Quarterly Journal of Economics*, 103(1), 179–199.

IZUMA, K., K. MATSUMOTO, C. F. CAMERER, AND R. ADOLPHS (2011): "Insensitivity to social reputation in autism," *Proceedings of the National Academy of Sciences*, 108(42), 17302–17307.

JOHNSON, N. D., AND A. A. MISLIN (2011): "Trust games: A meta-analysis," *Journal of Economic Psychology*, 32(5), 865–889.

KEIZER, K., S. LINDENBERG, AND L. STEG (2008): "The spreading of disorder," *Science*, 322(5908), 1681–1685.

KESSLER, J. B., AND S. LEIDER (2012): "Norms and Contracting," *Management Science*, 58(1), 62–77.

KIMBROUGH, E. O., J. B. MILLER, AND A. VOSTROKNUTOV (2013): "A Model of Frames and Norms in Games," mimeo.

KIMBROUGH, E. O., V. L. SMITH, AND B. J. WILSON (2008): "Historical Property Rights, Sociality, and the Emergence of Impersonal Exchange in Long-Distance Trade," *American Economic Review*, 98(3), 1009–39.

KIMBROUGH, E. O., AND A. VOSTROKNUTOV (2013): "The Social and Ecological Determinants of Common Pool Resource Sustainability," mimeo.

KNOCH, D., A. PASCUAL-LEONE, K. MEYER, V. TREYER, AND E. FEHR (2006): "Diminishing Reciprocal Fairness by Disrupting the Right Prefrontal Cortex," *Science*, 314, 829–832.

KOSFELD, M., M. HEINRICHS, P. ZAK, U. FISCHBACHER, AND E. FEHR (2005): "Oxytocin increases trust in humans," *Nature*, 435(7042), 673–676.

KOSFELD, M., A. OKADA, AND A. RIEDL (2009): "Institution Formation in Public Goods Games," *American Economic Review*, 99(4), 1335–1355.

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

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.

KURZBAN, R., P. DESCIOLI, AND E. O'BRIEN (2007): "Audience Effects on Moralistic Punishment," *Evolution and Human Behavior*, 28(2), 75–84.

KURZBAN, R., AND D. HOUSER (2005): "Experiments Investigating Cooperative Types in Humans: A Complement to Evolutionary Theory and Simulations," *Proceedings of the National Academy of Sciences*, 102(5), 1803–1807.

KURZBAN, R., K. MCCABE, V. L. SMITH, AND B. J. WILSON (2001): "Incremental commitment and reciprocity in a real-time public goods game," *Personality and Social Psychology Bulletin*, 27(12), 1662–1673.

LEVINE, D. K. (1998): "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, 1(3), 593–622.

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

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.

MCCABE, K., D. HOUSER, L. RYAN, V. SMITH, AND T. TROUARD (2001): "A functional imaging study of cooperation in two-person reciprocal exchange," *Proceedings of the National Academy of Sciences*, 98(20), 11832.

MCCABE, K. A., M. L. RIGDON, AND V. L. SMITH (2003): "Positive reciprocity and intentions in trust games," *Journal of Economic Behavior and Organization*, 52, 267–275.

MCCABE, K. A., V. L. SMITH, AND M. LEPORE (2000): "Intentionality detection and "mindreading": Why does game form matter?," *Proceedings of the National Academy of Sciences*, 97(8), 4404–4409.

MERTON, R. K. (1957): *Social Theory and Social Structure*. Free Press.

MILGRAM, S. (1963): "Behavioral study of obedience," *Journal of Abnormal Social Psychology*, 67, 371–378.

OOSTERBEEK, H., R. SLOOF, AND G. VAN DE KUILEN (2004): "Cultural differences in ultimatum game experiments: Evidence from a meta-analysis," *Experimental Economics*, 7(2), 171–188.

OSBORNE, M. J., AND A. RUBINSTEIN (1994): *A Course in Game Theory*. Cambridge, Mass.: MIT Press.

PAGE, T., L. PUTTERMAN, AND B. UNEL (2005): "Voluntary association in public goods experiments: Reciprocity, mimicry and efficiency," *The Economic Journal*, 115(506), 1032–1053.

PUTNAM, R. D. (2007): "E pluribus unum: Diversity and community in the twenty-first century, the 2006 Johan Skytte Prize Lecture," *Scandinavian Political Studies*, 30(2), 137–174.

R DEVELOPMENT CORE TEAM (2013): *R: A Language and Environment for Statistical Computing*R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

RABIN, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83(5), 1281–1302.

RIETZ, T. A., R. M. SHEREMETA, T. W. SHIELDS, AND V. L. SMITH (forthcoming): "Transparency, Efficiency and the Distribution of Economic Welfare in Pass-Through Investment Trust games," *Journal of Economic Behavior & Organization*.

RIGDON, M. L., K. A. MCCABE, AND V. L. SMITH (2007): "Sustaining Cooperation in Trust Games," *The Economic Journal*, 117(522), 991–1007.

ROSENBLOOM, T. (2009): "Crossing at a red light: Behaviour of individuals and groups," *Transportation Research Part F: Traffic Psychology and Behaviour*, 12(5), 389–394.

ROTH, A. E. (2007): "Repugnance as a Constraint on Markets," *The Journal of Economic Perspectives*, 21(3), 37–58.

ROTH, A. E., V. PRASNIKAR, M. OKUNO-FUJIWARA, AND S. ZAMIR (1991): "Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study," *American Economic Review*, pp. 1068–1095.

SCHMITT, P., R. SHUPP, K. SWOPE, AND J. MAYER (2008): "Pre-commitment and personality: Behavioral explanations in ultimatum games," *Journal of Economic Behavior & Organization*, 66(3), 597–605.

SHERIF, M. (1936): *The Psychology of Social Norms*. Harper.

SKYRMS, B. (2004): *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press, Cambridge.

SMITH, A. (1759): *The Theory of Moral Sentiments*. Liberty Fund: Indianapolic (1982).

SMITH, V. L. (2008): *Rationality in Economics: Constructivist and Ecological Forms*. New York, Cambridge University Press.

SMITH, V. L., AND B. J. WILSON (2013): "*Sentiments*, Conduct and Trust in the Laboratory," Economic Science Institute Working Paper.

WILSON, B. J. (2008): "Language games of reciprocity," *Journal of Economic Behavior & Organization*, 68(2), 365–377.

——— (2012): "Contra Private Fairness," *American Journal of Economics and Sociology*, 71(2), 407–435.

WILSON, B. J., T. JAWORSKI, K. SCHURTER, AND A. SMYTH (2012): "The Ecological and Civil Mainsprings of Property: An Experimental Economic History of Whalers' Rules of Capture," *Journal of Law, Economics and Organization*, 28(4), 617–656.

WILSON, J. Q., AND G. KELLING (1982): "The police and neighborhood safety: Broken windows," *The Atlantic Monthly*, 127, 29–38.

XIAO, E., AND D. HOUSER (2011): "Punish in Public," *Journal of Public Economics*, 95(7), 1006–1017.

YOUNG, H. (1993): "The Evolution of Conventions," *Econometrica*, 61(1), 57–84.

YOUNG, H. P. (1998): "Social Norms and Economic Welfare," *European Economic Review*, 42(3), 821–830.

ZIMBARDO, P. (2007): *The Lucifer Effect: Understanding How Good People Turn Evil*. New York, ME Sharpe.

Appendix (for online publication)

A Norm-Dependence in Some One-Shot Games

In all three models of games introduced below we maintain the same definitions: $g : [0,1] \to [0,1]$ is a strictly convex increasing differentiable function with $g(0) = 0$ and $g(1) = 1$, which represents the disutility of deviation from the norm; $\phi_p, \phi_r \geq 0$ is the norm sensitivity parameter of Proposer and Responder (in Trust and Ultimatum games).

Trust Game. In the trust game (Berg, Dickhaut, and McCabe, 1995) the Proposer decides to keep $x \in [0,1]$ and send $1-x$ to the Responder. The Responder receives $3(1-x)$ and then chooses $y \in [0, 3(1-x)]$, the amount she wants to return to the Proposer. The payoff of the Proposer is $x+y$ and the payoff of the Responder is $3(1-x)-y$. Suppose that the norm is for the Proposer to send everything to the Responder ($x=0$) and for the Responder to send back some amount $y = r_x(1-x)$, where the fraction $r_x \in [0, 1.5]$ is weakly decreasing in x. Here we assume that the Responder reciprocates by returning a (weakly) higher fraction of the offer $(1-x)$, the higher the offer is.

The norm-dependent utilities of the Proposer and the Responder are

$$U_p(x,y) = x + y - \phi_p g(x) \qquad U_r(x,y) = 3(1-x) - y - \phi_r g(\|y - r_x(1-x)\|).$$

Here $\|y - r_x(1-x)\| = |(y - r_x(1-x))/((3-r_x)(1-x))|$ is the normalization that is necessary to keep the deviations from the norm in the $[0,1]$ interval, ensuring that the highest possible disutility from deviation from the norm is equal to ϕ_r in all of the Responder's subgames.

In the SPNE the Responder chooses

$$y^*(x, \phi_r) = \underset{y \leq 3(1-x)}{\operatorname{argmax}} \; 3(1-x) - y - \phi_r g(\|y - r_x(1-x)\|)$$

which weakly increases in ϕ_r with $y^*(x, \phi_r) \to 0$ as $\phi_r \to 0$ and $y^*(x, \phi_r) \to r_x(1-x)$ as $\phi_r \to \infty$. The Proposer takes the best responses of the Responder into account and chooses

$$x^*(\phi_p, \phi_r) = \underset{x \in [0,1]}{\operatorname{argmax}} \; x - \phi_p g(x) + y^*(x, \phi_r).$$

Higher values of both ϕ_p and ϕ_r push the optimal proposal towards $x = 0$ (i.e. send everything).

Introducing incomplete information into the trust game is not particularly difficult. Here the Responder's optimal strategy is unchanged when we introduce uncertainty about ϕ_p. But the Proposer now chooses an optimal offer x that solves

$$\max_{x \in [0,1]} \; x - \phi_p g(x) + \int_0^\infty y^*(x, \phi_r) dF(\phi_r)$$

where F is the Proposer's belief regarding the Responder's norm sensitivity parameter ϕ_r.

Notice that the payoff of the Proposer $x - \phi_p g(x) + \int_0^\infty y^*(x, \phi_r) dF(\phi_r)$ might be easily decreasing in x if he believes that ϕ_r is high enough. Therefore, there is no reason to expect strong behavioral differences between high-ϕ and low-ϕ Proposers here, since the the optimal amount sent depends on beliefs about ϕ_r. On the other hand, high-ϕ Responders will always behave more reciprocally than low-ϕ Responders.

Dictator Game. Let the choice of the Proposer in the dictator game be $x \in [0,2]$ and assume that the norm prescribes equal division of the pie. Then the Proposer's norm-dependent utility can be defined as

$$U_p(x) = x - \phi_p g(|x-1|).$$

Thus, following the norm gives no disutility to the Proposer, while choosing to keep everything ($x = 2$) gives the maximal disutility of ϕ_p.

Given the assumptions on g, the optimal choice $x^*(\phi_p) \in [1, 2]$ of the Proposer is weakly decreasing in ϕ_p. For $\phi_p \to 0$, Proposer chooses $x^* \to 2$; for $\phi_p \to \infty$ she chooses $x^* \to 1$. Since g is assumed strictly convex, all intermediate values $x^*(\phi_p) \in (1, 2)$ are possible for some ϕ_p.[39] Note also that the evidence that people are sensitive to the "price" of giving (reported in Andreoni and Miller, 2002; Fisman, Kariv, and Markovits, 2007) is consistent with the tradeoffs implied by our model.

Ultimatum Game. In the ultimatum game (Güth, Schmittberger, and Schwarze, 1982) the Proposer chooses a division of the pie $x \in [0, 2]$, where x is the amount she decides to keep for herself. Then, the Responder decides to accept the division $(x, 2 - x)$ or to reject it, in which case both players get 0. Suppose that the norm prescribes the Proposer to divide the pie equally and the Responder to accept any offer $2 - x \geq 1$, which gives him at least half, and reject all offers $2 - x < 1$, which give the Responder less than half. Then, the norm-dependent utility of the Proposer is defined by

$$U_p(x, A) = x - \phi_p g(|x - 1|) \qquad U_p(x, R) = -\phi_p g(|x - 1|).$$

Here (x, A) stands for offer x followed by acceptance and (x, R) stands for offer x followed by rejection. The norm-dependent utility of the Responder is given in Table 6.

	Accept	Reject
$x > 1$	$U_r(x, A) = 2 - x - \phi_r$	$U_r(x, R) = 0$
$x \leq 1$	$U_r(x, A) = 2 - x$	$U_r(x, R) = -\phi_r$

Table 6: Responder's utility in the Ultimatum Game.

As in the dictator game the norm-dependent utility is the material payoff minus the disutility from the deviation from the norm. For the Proposer we maintain the same assumptions on g and ϕ_p as in the dictator case. For the discrete choice of the Responder we assume that he loses utility $\phi_r \geq 0$ if his action is not in accordance with the norm (this is without loss of generality; see Kimbrough, Miller, and Vostroknutov, 2013).

Now we are ready to characterize the Subgame Perfect Nash Equilibrium (SPNE). When $x \leq 1$, the Responder always chooses to accept ($2 - x \geq -\phi_r$). When $x > 1$ and $\phi_r \leq 1$, the Responder will accept all offers $2 - x \geq \phi_r$ and reject all smaller offers. For $\phi_r > 1$, the Responder will reject all offers $(2 - x)$ below 1 and accept all offers at or above 1. In other words, the Responder accepts an offer $(2 - x)$ if $x \leq x_r^*(\phi_r) = \max\{2 - \phi_r, 1\}$.

The Proposer takes into account this best response of the Responder. Let

$$x_p^*(\phi_p) = \operatorname{argmax}_x x - \phi_p g(|x - 1|).$$

$x_p^*(\phi_p)$ is in the interval $[1, 2]$. There are two possibilities:

if $x_p^*(\phi_p) \leq x_r^*(\phi_r)$ the Proposer chooses $x_p^*(\phi_p)$, i.e. the allocation that maximizes his utility;

[39] We could also consider another norm governing choice in the dictator game. Suppose that the dictator first earned the right to allocate the money through some pre-play task and then faces the decision in the dictator game. By introducing competition to assign "property rights," the normative action changes. Players who have earned the right to be dictator also believe they have earned a right to a larger share of the pie. In this case we can assume that $\eta = k > 1$. The utility of the Proposer becomes $U_p(x) = x - \phi_p g(|x - k|/k)$, so that $x^*(\phi_p) \in [k, 2]$ and the Proposer will choose to keep a larger share of the money. Indeed, in a variety of laboratory experiments with earned rights, this is exactly what is observed (e.g Hoffman, McCabe, and Smith, 1996). Moreover, if $\eta = 2$, we get $x^*(\phi_p) = 2$ for all ϕ_p, or, in other words, the proposer will keep the entire pie as was observed in Cherry, Frykblom, and Shogren (2002) who combined earned rights and double-blind protocols, sharply shifting social expectations towards selfishness.

if $x_p^*(\phi_p) > x_r^*(\phi_r)$ the Proposer chooses $x_r^*(\phi_r)$, or the smallest allocation that will be accepted by the Responder.[40]

The SPNE just described has an important implicit assumption: all the parameters of the norm-dependent utilities should be common knowledge. In particular, the Proposer should exactly know the norm sensitivity ϕ_r of the Responder. This sounds rather unrealistic. Let us instead assume that the Proposer holds a belief that the Responder's ϕ_r is distributed according to some cdf F. Then it can be shown that in a Perfect Bayesian Nash Equilibrium the Proposer's optimal behavior changes to

$$x^*(\phi_p) = \underset{x\in[0,2]}{\operatorname{argmax}} F(2-x)x - \phi_p g(|x-1|).$$

Here, the best responses of the Responder stay the same and do not depend on any incompleteness of information regarding ϕ_r or ϕ_p.

Crucial for our experiment is the implication that the Proposers' behavior depends on his beliefs about the Responder's type. In some cases, low-ϕ types may want to send large amounts if they believe their counterpart is a high-ϕ type. On the other hand, high-ϕ Responders will always be more likely to reject low offers.

B Public Goods Game Proofs

Proof of Lemma 1. According to s^0, since $m_{t-1} \neq \eta$, in period t all players should contribute 0 and continue doing so until the end of the game on the s^0-induced path. If some player i deviates in any period, then she can contribute a maximum of 1. Given fixed contributions of others, this will make the average contribution next period equal to $\frac{1}{n}$, which, by assumption, is less than η. Therefore, the deviation will fail to induce norm-following in the next period, and all players will have the same standard stage utility as in the game without norm-dependence. This implies that no single player deviation in any period (or multiple periods for that matter) can be profitable since the stage utility without norm-dependence decreases in contribution. ■

Proof of Proposition 1. Notice that from the assumption in 1.1 it follows that $\frac{1}{n}\Sigma < \eta$ (since $x_i^* \leq 1$). The same holds for 1.2: summing up n conditions $\Sigma_{-i} \leq (1-\alpha)n\eta$ for all i gives

$$\frac{1}{n}\Sigma \leq \frac{n-\alpha n}{n-1}\eta < \eta$$

where the last inequality follows from the game specific assumption $n\alpha > 1$. This means that in the first period, in accordance with s^0, each player j chooses x_j^* which results in $m_1 = \frac{1}{n}\Sigma < \eta$. Therefore, in period 2 all players reciprocate against the norm-violators and stop following the norm which results in 0 contributions by all players in period 2 and thereafter (on the s^0-induced path).

Case 1.1. Deviation in period 1. By assumption $\Sigma_{-i} < n\eta - 1$ for all $i \in N$. This implies that, if all players $j \neq i$ choose x_j^*, in accordance with s_j^0, then player i cannot choose an x_{i1} sufficiently large that the average contribution reaches η (since $x_{i1} \leq 1$). This means that player i cannot induce norm-following in period 2 and, given any contribution by i, all players will contribute 0 in the next period. Therefore, whatever

[40]Note that our model predicts that selfish behavior by the proposer is increasing in the stakes, because responders will be willing to accept a smaller share of a larger total pie, consistent with the evidence reported in Andersen, Ertaç, Gneezy, Hoffman, and List (2011).

player i does in period 1, the continuation of the game is the same. Thus, he should follow s_i^0.

Case 1.2. Deviation in period 1. Suppose that there is player i with $\Sigma_{-i} \geq n\eta - 1$.[41] Here player i can contribute $x > x_i^*$ in order to make the average of contributions $\frac{1}{n}\Sigma$ after period 1 to be equal to η ($x = n\eta - \Sigma_{-i}$).[42] This will encourage norm-following in the next period so that all other players j contribute x_j^* in period 2. Suppose that player i deviates in this way in the first k periods, thus inducing norm-following by the others and then switches back to s_i^0, which maximizes his utility given the choices of others. The payoff on the s^0 path is

$$1 - x_i^* + \alpha\Sigma - \phi_i g(\|\eta - x_i^*\|) + T - 1.$$

The payoff with the described deviation is

$$k[1 - x + \alpha(\Sigma_{-i} + x) - \phi_i g(\|\eta - x\|)] + 1 - x_i^* + \alpha\Sigma - \phi_i g(\|\eta - x_i^*\|) + T - k - 1.$$

No deviation occurs when

$$-x + \alpha(\Sigma_{-i} + x) - \phi_i g(\|\eta - x\|) \leq 0.$$

Substituting x and rearranging we obtain

$$\Sigma_{-i} \leq (1 - \alpha)n\eta + \phi_i g(\|\eta - n\eta + \Sigma_{-i}\|).$$

Now, for this to be an equilibrium for all ϕ, including $\phi = 0$, we must have

$$\Sigma_{-i} \leq (1 - \alpha)n\eta$$

which is true by the assumption of the Proposition and holds for all $k \leq T$. Since there are only two types of behavior exhibited by other players (contribute x_j^* for all $j \neq i$ or contribute nothing) the described deviation is the best possible for player i. Since this deviation is not profitable, under the assumption, we can conclude that there are no profitable deviations in the first period.

For the cases 1.1 and 1.2 it is left to show that no player i would like to deviate in periods after the first. But this is guaranteed by Lemma 1.

In case 1.3 there is player i with $\Sigma_{-i} \geq n\eta - 1$ who can, as in case 1.2, deviate to $n\eta - \Sigma_{-i}$ in period 1. This deviation will be now profitable since in case 1.3 $n\eta - 1 > (1 - \alpha)n\eta$ and the no deviation condition is $\Sigma_{-i} \leq (1 - \alpha)n\eta$. This completes the proof. ■

Proof of Proposition 2. On the path induced by s^1 all players contribute η in periods 1 to $T - 1$ and x_j^* for all $j \in N$ in period T. There can be no deviation in the last period. Therefore, we need to check that no player wants to deviate in periods 1 through $T - 1$. The payoff of player i on the s^1-induced path is

$$(T - 1)(1 - \eta + \alpha n\eta) + 1 - x_i^* + \alpha\Sigma - \phi_i g(\|\eta - x_i^*\|). \tag{1}$$

Now we need to identify the best possible deviation of player i in period k. Any deviation in period k to some contribution less than η will terminate norm-following, and in the next period all other players will contribute 0, according to s_j^0 for all $j \neq i$.[43] Therefore, player i's best choice is to deviate to x_i^*. In all later periods, all players will contribute 0, and player i's best response is to also contribute 0 in all periods after

[41] Otherwise we are back to case 1.1.

[42] Deviating to any larger x will only decrease i's payoff without changing anything else. Deviating to less than x will not activate norm-following by others and will leave i with less payoff than he could acquire under s_i^0. Thus, x is the best deviation that activates norm-following.

[43] Deviation to some contribution greater than η is not profitable, as it induces the same behavior in others and

k (by Lemma 1). Thus, the total payoff received by i from her best deviation to x_i^* in period $k \in 1 \dots T-1$ and later to 0 generates

$$(k-1)(1-\eta+\alpha n\eta)+1-x_i^*+\alpha[\eta(n-1)+x_i^*]-\phi_i g(\|\eta-x_i^*\|)+T-k. \tag{2}$$

There is no deviation if

$$\Sigma_{-i} \geq \frac{(T-k)\eta-(T-k-1)\alpha\eta n-\alpha\eta}{\alpha}.$$

Since, by assumption, $\alpha n > 1$ the RHS is increasing in k. So for this inequality to hold for all k we need it to hold for $k = T-1$ which is:

$$\Sigma_{-i} \geq \frac{\eta-\alpha\eta}{\alpha} = \eta\left[\frac{1}{\alpha}-1\right].$$

The best deviation from s_i^1 is unprofitable if this condition is satisfied. This completes the proof. ■

Proof of Proposition 3. On the $s^{2\ell}$-induced path all players contribute η in the first ℓ periods, then contribute x_j^* for all $j \in N$ in period $\ell+1$ and contribute 0 thereafter. The proof is analogous to the one of Proposition 2. The payoff on the $s^{2\ell}$-induced path is

$$\ell(1-\eta+\alpha n\eta)+1-x_i^*+\alpha\Sigma-\phi_i g(\|\eta-x_i^*\|)+T-\ell-1.$$

The best deviation during the cooperative stage in period $k \in 1..\ell$ is

$$(k-1)(1-\eta+\alpha n\eta)+1-x_i^*+\alpha[\eta(n-1)+x_i^*]-\phi_i g(\|\eta-x_i^*\|)+T-k.$$

The reasons why this deviation is the best are the same as in Proposition 2. The deviation is not profitable if

$$\Sigma_{-i} \geq \frac{-\eta(\ell-k)(\alpha n-1)+\eta-\alpha\eta}{\alpha}.$$

The RHS is increasing in k and decreasing in ℓ. Thus the RHS is highest when $k = \ell$. So we obtain the same condition as in Proposition 2:

$$\Sigma_{-i} \geq \eta\left[\frac{1}{\alpha}-1\right].$$

Now we need to check that there are no profitable deviations in periods after ℓ where the players contribute x_j^* and then 0. By Proposition 1 we know that such profitable deviations do not exist. This completes the proof. ■

Proof of Proposition 4. We can classify all non-terminal histories h of Γ by the average contribution $m(h)$. There are two cases: $m(h) = \eta$ and $m(h) \neq \eta$. Strategy profile s^0 is a NE of Γ by Proposition 1. Strategy profile $s^0(h)$ restricted to subgame $\Gamma(h)$ with $m(h) = \eta$ is the same as s^0 only with less periods. Therefore, by Proposition 1, $s^0(h)$ is a NE of $\Gamma(h)$. Strategy profile $s^0(h)$ restricted to subgame $\Gamma(h)$ with $m(h) \neq \eta$ is a NE of $\Gamma(h)$ by Lemma 1. Therefore, s^0 restricted to any subgame is a NE of that subgame, i.e. s^0 is a SPNE of Γ.

Proof of Proposition 5. We can classify all non-terminal histories h of Γ by the average contribution $m(h)$. There are two cases: $m(h) = \eta$ and $m(h) \neq \eta$. Strategy profile s^1 is a NE of Γ by Proposition 2. Strategy profile $s^1(h)$ restricted to subgame $\Gamma(h)$ with $m(h) = \eta$ is the same as s^1 only with fewer periods.

decreases i's stage payoff instead of increasing it.

Therefore, by Proposition 2, $s^1(h)$ is a NE of $\Gamma(h)$. Strategy profile $s^1(h)$ restricted to subgame $\Gamma(h)$ with $m(h) \neq \eta$ is a NE of $\Gamma(h)$ by Lemma 1. Therefore, s^1 restricted to any subgame is NE or, i.e. s^1 is a SPNE of Γ. ■

B.1 Repeated Public Goods Game with Incomplete Information

In this section we introduce incomplete information into the game Γ. The reason for doing this is that the norm sensitivity parameters ϕ, are not only unobservable but also may be very hard to deduce from observed behavior. For example, in the setup of Proposition 2 players with a wide range of ϕ's chose to contribute η in equilibrium, which would not allow observers to estimate their ϕ.

We represent incomplete information as uncertainty about x_i^*, the optimal choice of player i in one-shot game.[44] Assume that, for each player i, x_i^* is distributed according to some F_i. All F_i are common knowledge. In what follows we will construct two Perfect Bayesian Nash Equilibria (PBNE) of Γ with incomplete information structure given by $(F_i)_{i\in N}$ which correspond to the SPNE depicted in Propositions 4 and 5.

It is easier to start with the cooperative equilibrium (Proposition 5). The strategy profile s^1 will also constitute a PBNE. We only need to introduce common beliefs about others' types to each history. For histories h with $m(h) = \eta$ let $\mu_i(h) = F_i$ be the common belief of all players but i about her type after h. For each history with $m(h) \neq \eta$ let $\mu_i(h) = \delta_{x_i(h)}$ be the point mass at $x_i(h)$, the choice of i in the last period of h. Thus, we assume that whenever a player observes someone choosing some contribution $x < \eta$ she believes that this player has chosen according to her one-shot game best reply. In principle, the beliefs can be defined in many other ways, but for our particular game this is irrelevant. The only thing that matters is that the prior common belief is $(F_i)_{i\in N}$ and that this belief does not get updated if all players choose η (see Osborne and Rubinstein (1994) section 12.3 for the relevant definitions).

Before we proceed let us give some definitions. Let $E := \sum_{j\in N} E_{F_j}[x_j^*]$ be the sum of expected values of x_j^* given F_j and let E_{-i} be the same sum without player i.

Proposition 6. Suppose $\frac{1}{n} < \eta$. If for all i it is true that $E_{-i} \geq \eta\,[1/\alpha - 1]$ then the strategy profile s^1 together with the beliefs $(\mu_i(h))_{i\in N, h\in H\setminus Z}$ constitute a PBNE.

Proof. The proof that no type of any player wants to deviate from the strategy s^1 in Γ is exactly the same as that of Proposition 2. Only Σ_{-i} is replaced with E_{-i} and Σ is replaced with $E_{-i} + x_i^*$. Similarly, no player wants to deviate from $s^1(h)$ in subgames h with $m(h) \neq \eta$ by Lemma 1: uncertainty plays no role here. In all subgames $s^1(h)$ with $m(h) = \eta$ the proof of Proposition 2 is used again since by assumption the beliefs after any such history are $(F_i)_{i\in N}$ as in the supergame. ■

Proposition 7. Suppose that $\frac{1}{n} < \eta$ and the assumptions of Proposition 1.1 or 1.2 hold with Σ_{-i} replaced with E_{-i}. Then the strategy profile s^0 together with the beliefs $(\mu_i(h))_{i\in N, h\in H\setminus Z}$ constitute a PBNE.

Proof. The proof that no type of any player wants to deviate from the strategy s^0 in Γ is exactly the same as that of Proposition 1. Only Σ_{-i} is replaced with E_{-i} as above and Σ is replaced with $E_{-i} + x_i^*$. Similarly, no player wants to deviate from $s^0(h)$ in subgames h with $m(h) \neq \eta$ by Lemma 1: uncertainty plays no role here. In all subgames $s^0(h)$ with $m(h) = \eta$ the proof of Proposition 2 is used again since by assumption the beliefs after any such history are $(F_i)_{i\in N}$ as in the supergame. ■

[44] We could have defined uncertainty directly over ϕ_i, but we find it more convenient to work with uncertainty over x_i^*. The two formulations are equivalent: ϕ_i's that correspond to some x_i^* can be found through the inverse of $x_i^*(\phi_i)$.

C Instructions for the Rule Following Stage

General information
You are now participating in a decision making experiment. If you follow the instructions carefully, you can earn a considerable amount of money depending on your decisions and the decisions of the other participants. Your earnings will be paid to you in CASH at the end of the experiment

This set of instructions is for your private use only. **During the experiment you are not allowed to communicate with anybody**. In case of questions, please raise your hand. Then we will come to your seat and answer your questions. Any violation of this rule excludes you immediately from the experiment and all payments. The research organization METEOR has provided funds for conducting this experiment.

Part I
In Part I of this experiment, you control a stick figure that will walk across the screen.

Once the experiment begins, you can start walking by clicking the **"Start"** button on the left of the screen. Your stick figure will approach a series of stop lights and will stop to wait at each light. To make your stick figure walk again, click the **"Walk"** button in the middle of the screen.

The rule is to wait at each stop light until it turns green.

Your earnings in Part I are determined by the amount of time it takes your stick figure to walk across the screen. **Specifically, you begin with an initial endowment of 8 Euro**. Each second, this endowment will decrease by **0.08 Euro**.

This is the end of the instructions for Part I. If you have any questions, please raise your hand and an experimenter will answer them privately. Otherwise, please wait quietly for the experiment to begin.

D Instructions for the Dictator Game

Part II

In this part, there will be two types of people, Red and Blue. Throughout Part II, you will be either a Red person or a Blue person depending on a random choice by the computer.

You will be paired with a person of the other type.

You will interact with only 1 other person in the room. In this experiment a Blue person makes a choice and a Red person does not choose. The amount of money you receive depends on the decision made by the Blue person in your pair.

Instructions for Blue People

Each Blue person begins with 16 Euro. A Blue person chooses how to allocate this money between him/herself and a Red person he/she is paired with. To specify an allocation, the Blue person types the amount he/she wants to allocate to him/herself and the amount he/she wants to allocate to the Red person and then clicks OK. The two amounts must sum up to 16 Euro.

Instructions for Red People

After the Blue person chooses an allocation, the Red person will see how much money the Blue person allocated to the Red person. The Red person does not make any decisions.

This is the end of the instructions for Part II. If you have any questions please raise your hand and an experimenter will come by to answer them.

E Instructions for the Ultimatum Game

Part II

In this part, there will be two types of people, Red and Blue. Throughout Part II, you will be either a Red person or a Blue person depending on a random choice by the computer.

You will be paired with a person of the other type. You will interact with only 1 other person in the room. The amount of money you receive depends on the decision you make and on the decision of the person you are paired with.

Each person makes his/her decision without knowing the decision of the other person.

Instructions for Blue People

Each Blue person begins with 16 Euro. A Blue person chooses how to allocate this money between him/herself and a Red person he/she is paired with. To specify an allocation, the Blue person types the amount he/she wants to allocate to him/herself and the amount he/she wants to allocate to the Red person and then clicks OK. The two amounts must sum up to 16 Euro.

Instructions for Red People

The Red person makes a decision that determines whether he/she will Accept or Reject the allocation chosen by the Blue person. If the Red person accepts an allocation, then he/she gets the amount of money specified by this allocation and the Blue person gets his/her part. If the Red person rejects an allocation, then both the Red and the Blue person get nothing.

In practice, the Red person chooses a single number (call it X). This number represents the minimum amount that the Red person is willing to accept. Once the number has been entered, the Red person will click "OK." Then, if the Blue person chose any amount (call it Y) that is greater than or equal to the Red person's chosen minimum (in other words, if Y X), the Red person accepts the allocation and receives Y Euro while the Blue person receives 16 - Y Euro. On the other hand, if Y is less than the Red person's chosen minimum (that is, if Y ¡ X), then the allocation is rejected and both persons receive nothing.

After both the Blue and the Red persons have decided, the outcome will be revealed and both people will have their earnings added to their total for the experiment.

This is the end of the instructions for Part II. If you have any questions please raise your hand and an experimenter will come by to answer them.

F Instructions for the Trust Game

Part II

This part of the experiment will consist of several periods.

In this part, there will be two types of people, Red and Blue. You will be both a Red person and a Blue person depending on the period. Each period you will be randomly paired with a person of the other type. In this experiment you will interact with 3 other people in the room.

Instructions for Blue People

Each Blue person begins each period with 80 tokens. A Blue person may choose to send some, all, or none of these tokens to a Red person he/she is paired with by typing the amount into a box in the center of the screen and then clicking "OK".

Any tokens that a Blue person sends to a Red person will be subtracted from the Blue persons account, multiplied by 3 and transferred to the Red person. Any tokens that a Blue person chooses not to send to the Red person remain the Blue persons earnings. (Only Blue people will be able to send tokens and have them multiplied.)

Instructions for Red People

Each Red person enters a period with 80 tokens. After the Blue person makes a decision, the Red person will see how many tokens were sent from the Blue person.

The amount sent by the Blue person will be multiplied by 3 and added to the Red persons account. Then the Red person decides to send some, all or none of these tokens to the Blue person by typing the amount into a box in the center of the screen and then clicking "OK". (Only Red people will make this decision.)

In each period, each Red person is paired with one Blue person for the entire period. (One "period" consists of one Blue person deciding how many tokens to send to one Red person and that Red person deciding how many of the multiplied tokens to send to the paired Blue person.)

Summary

A Blue persons earnings for a period are:
Earnings =
Starting tokens
minus Amount Sent to Red
plus Amount Received from Red

A Red persons earnings for a period are:
Earnings =
Starting tokens
plus Amount Received from Blue x 3
minus Amount Sent to Blue

At the end of the experiment the sum of your tokens from all periods will be converted to Euros at a rate of 100 tokens = 1 Euro and paid to you privately in cash, along with your earnings from Part 1 of the experiment. This is the end of the instructions. If you have any questions please raise your hand and an experimenter will come by to answer them.

G Instructions for the Public Goods Game

Part II

This part of the experiment will consist of several decision making periods. In each period, you are given an endowment of 50 tokens. Your task is to decide how to divide these tokens into either or both of two accounts: a private account and a group account. Each period you receive the sum of your earnings from your private account plus your earnings from the group account.

There are 4 people, including yourself, participating in your group. You will be matched with the same people for all of Part II.

Each token you place in the private account generates a cash return to you (and to you alone) of one cent (0.01 Euro). Tokens placed in the group account yield a different return.

Every member of the group receives the same return for each token you place in the group account. Similarly, you receive a return for every token that the other members of the group place in the group account. Thus, your earnings in each decision period are the number of tokens you place in your private account, plus the return from all tokens you and the other members of the group place in the group account.

Specifically, the total amount of tokens in the group account, that is, your group account tokens and the tokens placed in the group account by other members of the group, is doubled and then equally divided among 4 members of the group.

Here are two examples to make this clear:

(1) Suppose you place 0 tokens in the group account and the other members of your group place a total of 150 tokens in the group account. Your earnings from the group account would be (2 * 150) / 4 = 75 cents. Other members of the group would also receive 75 cents from the group account.

(2) Suppose you place 45 tokens in the group account and the other members of your group place a total of 15 tokens in the group account. The total group contribution is 60. Your earnings from the group account would be (2 * 60) / 4 = 30 cents. Other members of the group would also receive 30 cents from the group account.

Each period proceeds as follows:

First, decide on the number of tokens to place in the private and in the group accounts by entering numbers into the boxes labeled private and group. Your entries must sum to your token endowment which is always 50. While you make your decision, the 3 other members in your group will also divide their token endowments between the private and group accounts.

Second, after everyone has made a decision, your earnings for that decision period are the sum of your earnings from the private and group accounts.

As an example, suppose the total contribution to the group account at the end of the period was 120. Your contribution to the group account was 30, which means your contribution to the private account was 20. You would earn 80 cents this period, 20 from private account and (2 * 120) / 4 = 60 from the group account.

While you are deciding how to allocate your tokens, everyone else in your group will be doing so as well. When the period is over the computer will display your earnings for that period and your total earnings up to and including that period.

This is the end of the instructions. If you have any questions please raise your hand and an experimenter will come by to answer them.

H Moral Foundations Questionnaire

Part 1. When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please rate each statement using this scale:

0	1	2	3	4	5
not at all relevant	not very relevant	slightly relevant	somewhat relevant	very relevant	extremely relevant

1. Whether or not someone suffered emotionally
2. Whether or not some people were treated differently than others
3. Whether or not someone's action showed love for his or her country
4. Whether or not someone showed a lack of respect for authority
5. Whether or not someone violated standards of purity and decency
6. Whether or not someone was good at math
7. Whether or not someone cared for someone weak or vulnerable
8. Whether or not someone acted unfairly
9. Whether or not someone did something to betray his or her group
10. Whether or not someone conformed to the traditions of society
11. Whether or not someone did something disgusting
12. Whether or not someone was cruel
13. Whether or not someone was denied his or her rights
14. Whether or not someone showed a lack of loyalty
15. Whether or not an action caused chaos or disorder
16. Whether or not someone acted in a way that God would approve of

Part 2. Please read the following sentences and indicate your agreement or disagreement:

0	1	2	3	4	5
Strongly Disagree	Moderately Disagree	Slightly Disagree	Slightly Agree	Moderately Agree	Strongly Agree

1. Compassion for those who are suffering is the most crucial virtue.
2. When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.
3. I am proud of my countrys history.
4. Respect for authority is something all children need to learn.
5. People should not do things that are disgusting, even if no one is harmed.
6. It is better to do good than to do bad.
7. One of the worst things a person could do is hurt a defenseless animal.
8. Justice is the most important requirement for a society.
9. People should be loyal to their family members, even when they have done something wrong.
10. Men and women each have different roles to play in society.
11. I would call some acts wrong on the grounds that they are unnatural.
12. It can never be right to kill a human being.
13. I think its morally wrong that rich children inherit a lot of money while poor children inherit nothing.
14. It is more important to be a team player than to express oneself.
15. If I were a soldier and disagreed with my commanding officers orders, I would obey anyway because that is my duty.
16. Chastity is an important and valuable virtue.

*The Moral Foundations Questionnaire (full version, July 2008) by Jesse Graham, Jonathan Haidt, and Brian Nosek. For more information about Moral Foundations Theory and scoring this form, see: www.MoralFoundations.org

Moral Foundations Questionnaire: 30-Item Full Version Item Key, July 2008

–Below are the items that compose the MFQ30. Variable names are IN CAPS
–Besides the 30 test items there are 2 catch items, MATH and GOOD
–For more information about the theory, or to print out a version of this scale formatted for participants, or to learn about scoring this scale, please see: www.moralfoundations.org

---

PART 1 ITEMS (responded to using the following response options: not at all relevant, not very relevant, slightly relevant, somewhat relevant, very relevant, extremely relevant)

MATH - Whether or not someone was good at math [This item is not scored; it is included both to force people to use the bottom end of the scale, and to catch and cut participants who respond with last 3 response options]

Harm:
EMOTIONALLY - Whether or not someone suffered emotionally
WEAK - Whether or not someone cared for someone weak or vulnerable
CRUEL - Whether or not someone was cruel
Fairness:
TREATED - Whether or not some people were treated differently than others
UNFAIRLY - Whether or not someone acted unfairly
RIGHTS - Whether or not someone was denied his or her rights
Ingroup:
LOVECOUNTRY - Whether someones action showed love for his or her country
BETRAY - Whether or not someone did something to betray his or her group
LOYALTY - Whether or not someone showed a lack of loyalty
Authority:
RESPECT - Whether or not someone showed a lack of respect for authority
TRADITIONS - Whether or not someone conformed to the traditions of society
CHAOS - Whether or not an action caused chaos or disorder
Purity:
DECENCY - Whether or not someone violated standards of purity and decency
DISGUSTING - Whether or not someone did something disgusting
GOD - Whether or not someone acted in a way that God would approve of

PART 2 ITEMS (responded to using the following response options: strongly disagree, moderately disagree, slightly disagree, slightly agree, moderately agree, strongly agree)

GOOD It is better to do good than to do bad. [Not scored, included to force use of top of the scale, and to catch and cut people who respond with first 3 response options]

Harm:
COMPASSION - Compassion for those who are suffering is the most crucial virtue.
ANIMAL - One of the worst things a person could do is hurt a defenseless animal.
KILL - It can never be right to kill a human being.
Fairness:
FAIRLY - When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.
JUSTICE Justice is the most important requirement for a society.
RICH - I think its morally wrong that rich children inherit a lot of money while poor children inherit nothing.
Ingroup:
HISTORY - I am proud of my countrys history.
FAMILY - People should be loyal to their family members, even when they have done something wrong.
TEAM - It is more important to be a team player than to express oneself.
Authority:
KIDRESPECT - Respect for authority is something all children need to learn.
SEXROLES - Men and women each have different roles to play in society.
SOLDIER - If I were a soldier and disagreed with my commanding officers orders, I would obey anyway because that is my duty.
Purity:
HARMLESSDG - People should not do things that are disgusting, even if no one is harmed.
UNNATURAL - I would call some acts wrong on the grounds that they are unnatural.
CHASTITY - Chastity is an important and valuable virtue.

I Additional Tables and Figures

	Moral Foundation				
	Authority	Fairness	Harm	Ingroup	Purity
Mean	16.08	21.40	20.73	16.91	13.76
Std. Deviation	(4.67)	(4.11)	(4.55)	(4.33)	(4.94)

Table I1: Average Moral Foundation Scores (out of 30)

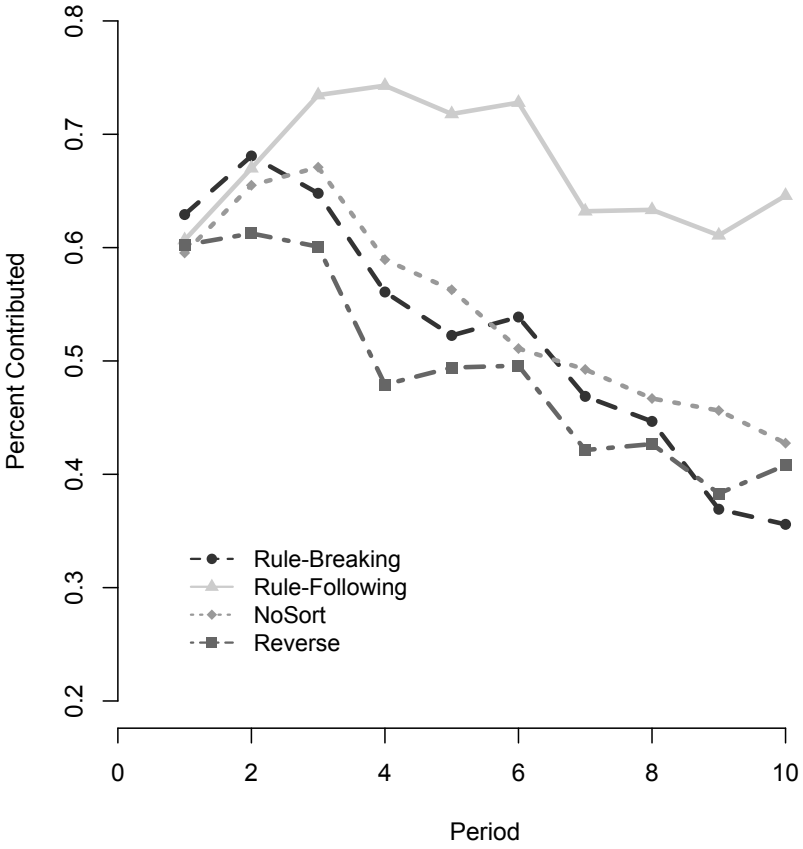

Figure I1: Time Series of Mean Group Public Good Contributions by Treatment, Rule-Following and Rule-Breaking. This figure includes sorted NoRule sessions in the computation of means for the rule-following and rule-breaking time series. While the rule-breakers mean is slightly higher than when the NoRule sessions are excluded (compare figure 3), this is no surprise since the NoRule treatment classifies as rule-breakers many of those who would have followed the rule had we stated it expressly. Moreover, clear differences between types remain.

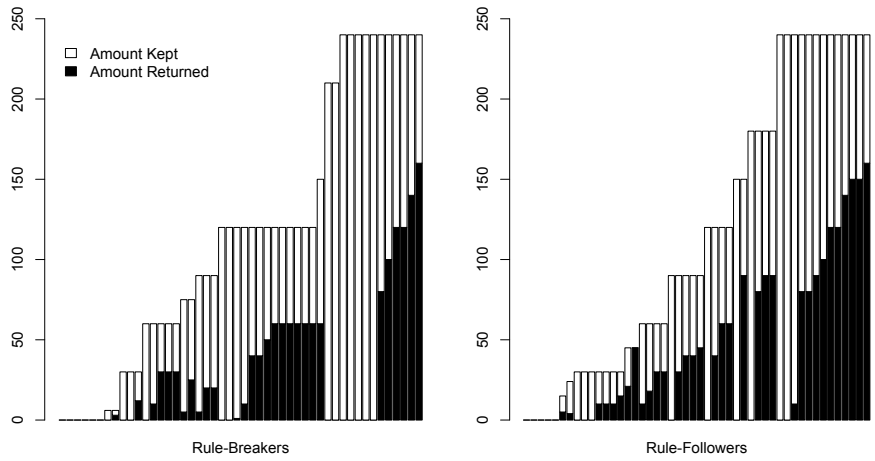

Figure I2: Amount Received, Kept and Returned in the TG Treatment, by Group Type