SOCIAL DECISION MAKING AS A COGNITIVE BEHAVIORIST VIEWS IT:

INVESTIGATIONS INTO A NORM-BASED UTILITY FUNCTION

By

JEFFREY RONALD DEWITT

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Psychology

Written under the direction of

Gretchen B. Chapman

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May 2017

ABSTRACT OF THE DISSERTATION

Social Decision Making as a Cognitive Behaviorist Views It:

Investigations into a Norm-Based Utility Function

By JEFFREY RONALD DEWITT

Dissertation Director:

Gretchen B. Chapman

Models of *social preferences* have been at the fore of research in decision science seeking to explain strategic behavior in experimental games. However, the source of the behavioral consistency these approaches seek to model could originate from a motivation to conform to and enforce social norms. Across four studies, the rationality of strategic choices in monetary games is measured as well as whether beliefs about the behavior and expectations of others are associated with pro-social decision making. Overall, the norms of everyday life may be difficult to manipulate reliably in traditional lab tasks, but measuring beliefs about salient norms consistently tracks the behavior of cooperators and defectors in social dilemmas.

Table of Contents

Part 1: A Rational Species

Scientific approaches to understanding human cooperation are motivated by different theoretical perspectives that help specify the research questions of interest and provide a framework for interpreting results. The current work examines a psychological account of cooperation with roots in decision theory, but its motivation is better understood by placing it within historical context and explanatory space.

At their broadest, theories of human cooperation can operate at the level of evolutionary function, ultimate explanations, or the level of social and psychological instantiation, proximate explanations (Scott-Phillips, Dickins & West, 2011). Ultimate explanations seek to answer the question of why cooperative behavior is adaptive or stable in equilibrium, whereas proximate explanations answer the question of what underlying mechanisms generate and sustain cooperation. As a result, theories of ultimate causation are rooted in the direct (personal) and indirect (kin) reproductive benefits of cooperating[1], whereas theories of proximate causation are embedded in either the social institutions or psychological tendencies of cooperative agents (Scott-Phillips, Dickins & West, 2011). The purpose of the present work is to test a proximate explanation for cooperation that focuses on its psychological underpinnings from the perspective of individual decision theory, though a more sociological view will also be commented on in the general discussion. In particular, Study 1 examines the extent to which behavior in cooperative lab experiments can be explained as the result of a

[1] This is true even for "multi-level" or "group selection" accounts (see, e.g., Lehmann, Keller, West & Roze, 2007).

consistent, goal-directed, process, while Study 2 investigates whether the underlying nature of this process is a concern for the welfare of others or primarily the result of social pressure.

Sociological and psychological explanations often describe patterns of behavior at different levels of analysis (e.g., macro vs. micro), but psychological theories themselves can be situated within different levels. For instance, in the information-processing approach to cognitive psychology, a useful division has been to explain behavior at the level of the cognitive computation being performed, the level of the mental representations and algorithms used in performing that computation, and at the biological level in which the algorithm is physically realized (Marr, 1982; Pylyshyn, 1999). The value of these divisions is that some behavioral patterns are best explained at the level of biology (e.g. poor performance on an Ishihara color vision task due to a missing or shifted photoreceptor), whereas others are better accounted for at higher levels of processing (e.g. when that same individual nevertheless chooses clothes which actually "match" due to transforming the inputs from her retina with semantic knowledge of how those perceptions "should" be paired). In regards to cooperative behavior, examples of biological explanations are those highlighting the influence of hormones such as oxytocin and testosterone (see, e.g., Rillings et al. 2012; van Honk, Montoya, Bos, van Vugt & Terburg, 2010), whereas computational accounts often focus on the integration of an individual's higher-level beliefs and preferences[2]. The current work

[2] These include not only the rational actor model, but also models of heuristic decision making and those in which beliefs and/or preferences emerge from lower-level attentional and memory processes (see Oppenheimer & Kelso, 2015 for a representative review of the latter type).

investigates the descriptive accuracy of adopting a specific computational approach - the rational actor model from decision theory.

The rational actor model was formalized in parallel with the zeitgeist of behaviorism in psychology and likewise embraced the analysis of overt behavior without relying on an analysis of internal mental processes. Assumptions were no longer made about the nature of "preferences" (e.g. that they be hedonic), but rather preferences became synonymous with the observable choices that "revealed" them, as were subjective beliefs through choices over lotteries (Samuelson, 1937; von Neumann & Morgenstern, 1944; Savage, 1954). As long as decision making conformed to several assumptions regarding consistency (e.g. that choices be transitive), behavior could be represented by a functional form and predicted "as if" it were maximizing a goal (utility), without an account of the underlying motivation for that goal or the process leading to goal pursuit. At the time, this abstraction was viewed as a positive; it meant the model could be applied to behavior resulting from both conscious and unconscious processes, as well as expanded to account for the actions of groups, firms, machines, and non-human animals alike (Binmore, 2010; Gintis, 2009; Bowles & Gintis, 2011). But with the dawn of the computational view of the mind during the cognitive revolution, psychologists turned to rational choice not only for an accurate predictive account of behavior, but also for a theory of the actual cognitive processes resulting in that behavior. This led to new empirical programs in psychology from the computational level (e.g. Kahneman & Tversky, 1979) down to the biological (e.g. Kable & Glimcher, 2007), which continue today. And in the area of strategic decision making or

interdependent choice, this led specifically to the field of behavioral game theory (Camerer, 2003).

Analytical game theory applies rational choice analysis to strategic interactions where no one person can unilaterally determine the outcome for all parties involved. The subjective value of possible strategy profiles (complete plans of action for all agents involved) is represented as a "payoff" and agents are often predicted to choose the strategy that maximizes this goal given their beliefs about the knowledge and payoffs of other players (i.e. games are often analyzed using the Nash equilibrium solution concept). The structure of these payoffs defines the type of game being played, e.g., social dilemmas are defined as games in which one's personal payoff is always maximized by choosing a particular strategy (the defection strategy), but the total payoff to all players is maximized when everyone chooses a different strategy (the cooperative strategy)[3]. At its inception, empirical tests of game theoretic predictions entailed assigning monetary payments to strategy profiles, defining the game type based on the structure of personal payments, and then comparing real-world behavior against the Nash equilibria predicted for self-regarding money-maximizing agents. While this simplifying assumption is a natural starting point that has proven pragmatic when behavior is filtered through certain social institutions, e.g., competitive markets (Smith, 1962), it need not be the case. In particular, the theory of rational choice is no

[3] Let C = the cooperative strategy, D = the defection strategy, and CD correspond to the situation in which the first player cooperates and the second defects. Formally, a social dilemma is defined by the following payoffs to the first player: CD < DD < CC < DC with an additional assumption that CD + DC < 2(CC) so that defection is a dominant strategy for both players and there is no gain from taking turns playing C and D (Gintis, 2009).

less threatened by socially motivated or social-context dependent preferences than it is by state- or time-dependent ones, provided that decision making is consistent within the state, time, or social-context the agent finds herself[4]. With this in mind, Andreoni & Miller (2002) conducted influential work seeking to describe nature of "social preferences" in a non-strategic task referred to as the dictator game (DG). And in the current work, Study 1 was designed to extend these earlier findings both within a more strategic setting (an experimental social dilemma) and across strategic settings of the same type (i.e., another social dilemma framed differently).

In the DG, participants are organized into pairs and one person, the dictator, is given an endowment of money, typically $10, which she may divide between herself and the other person, the recipient. The recipient makes no decisions in this task and has no means of protesting the allocation[5]. Because participants are paired anonymously and payments are made in private, any deviation by the dictator from keeping all of the endowment is often viewed as a measure of her *social preference*. Andreoni & Miller (2002) introduced "tokens" into this task and had participants make a series of DG choices that differed only in the exchange rate of tokens to "points" for each player, which were later converted into money. By utilizing different exchange rates, the authors created different budget constraints that would not affect the choices of money-maximizing agents, but could affect those with social preferences. Andreoni

[4] Gintis & Helbing (2015) exemplify this view when they argue that, "These state-dependent aspects of preferences render the empirical estimation of preferences somewhat delicate, but they present no theoretical or conceptual problems." p. 37

[5] This makes the DG not technically a "game" as there is no strategic interaction.

& Miller (2002) found that the vast majority of participants made choices across games consistent with the axioms of rational choice. This meant that the behavior of a majority of participants was rationalizable - their social decisions were internally consistent with one another and could therefore be modeled by a functional form that the individual acted "as if" she was trying to maximize. In particular, 43% could be fit exactly with a standard utility function while the remainder could be categorized as closely approximating one. However, only 47.2% of participants could be fit with a money-maximizing utility function (perfect selfishness), whereas 30.4% consistently preferred equal payments (Rawlisans/Leontief), and 22.4% acted in line with utilitarian values – consistently preferring the agent with the higher exchange rate (perfect substitutes). The finding of relatively coherent, yet heterogeneous, social preferences in non-strategic settings was subsequently supported in a richer DG design (Fisman, Kariv & Markovits, 2007) and later expanded to predict subsequent behavior in a strategic setting (Yang, Onderstal & Schram, 2016). In addition, work on the general consistency of social behavior across different game types (Blanco, Engelmann & Normann, 2011), within the same type of game over time (Volk, Thöni & Ruigrok, 2012; Mao, Dworkin, Suri & Watts, 2016), or both (Yamagishi et al., 2013) has broadly found support for consistency at the aggregate level (in terms of the overall distribution of choices) that decreases when analyzed for individual subjects. In the current work, Study 1 adds to this literature by examining individual level consistency both within and between two-person social dilemma tasks at a single time point. By focusing on a single game type in

a single experimental session, Study 1 will help establish a lower limit on the proportion of rationalizable decision makers in anonymous one-shot experimental social dilemmas.

While finding that social decisions are amenable to rational models supports game theoretic approaches to understanding behavior, the underlying nature of these preferences, such as those reported by Andreoni & Miller (2002), remains an active line of research. In particular, while the anonymous DG controls for strategic concerns such as the fear of reprisal or reputation building, the motivation to not violate social expectations could still be the true cause of pro-social behavior as opposed to a genuine concern for the welfare of others. This idea is supported by findings from Handgraff, Van Dijk, Vermunt, Wilke, and De Dreu (2008) who employed an innovation on the ultimatum game (UG) designed by Suleiman (1996). The standard UG allows the recipient of a dictator (now "proposer") allocation to respond by either accepting the offer, in which case it is carried out, or rejecting the offer, in which case both players end up with nothing. In Suleiman's (1996) version, choosing to reject an offer resulted in the offered split being reduced/discounted by a known factor, δ. If $\delta = 0$, "rejecting" has the same effect as the standard UG (both players received nothing), but if $\delta = 1$ then "rejecting" has no effect on the outcome and the players are effectively in a DG. Suleiman (1996) and Handgraaf et al. (2008) varied δ and both found that proposers made significantly more generous "offers" when $\delta = 1$ (they were playing the DG) than when δ was high but not exactly 1 (0.8 or 0.9). Handgraaf et al. (2008) argue that, essentially, competitive norms/expectations are activated when the recipient has any power to retaliate ($\delta = 0.8$ or 0.9), but that norms of social responsibility are activated

when the recipient is powerless ($\delta = 1$). This is to say that the source of consistent pro-social preferences may be an underlying preference to conform to the prevailing social norm as opposed to reflecting an actual "taste for fairness" (a point developed more fully in DeWitt, 2017b). In the current work, Study 2 contributes to this literature by employing a DG that varies both social expectations and social relationships to more clearly map the motivational dynamics of pro-social behavior.

Study 2 conceptually replicates the work of Dana, Cain & Dawes (2006) who provided a more direct test of the social expectations hypothesis by conducting a standard $10 DG and then presenting dictators with an unexpected opportunity to "exit" the game by accepting $9 instead of having their DG choice carried out. If a participant accepted the $9 exit payment, the recipient of their DG choice would never be informed about the DG opportunity. So while accepting $9 was dominated by the $10 DG, if some participants act generously to avoid violating the recipient's expectations they may prefer the quiet exit. Dana et al. (2006) found that a significant proportion of dictators chose to exit (28% and 43% in Studies 1 and 2 respectively), and this finding was verified in a more rigorous design by Broberg, Ellingsen & Johannesson (2007), who found that only 36% of participants had exit reservation prices consistent with selfish or social preferences (i.e. only 36% of participants demanded at least the dictator endowment to exit using a Becker-Degroot-Marschak procedure). In addition, when Dana et al. (2006) introduced an alternative design whereby recipients would never find out why they were receiving the amount given to them in the DG, exiting nearly vanished altogether - further supporting the view that the expectations of (even

anonymous) others can influence decision making. These results were bolstered by the findings of Dana, Weber & Kuang (2007) that dictators exploit "moral wiggle room" to behave in a money-maximizing fashion when, for example, they have plausible deniability due to common knowledge that a random timer may cut them off before deciding and instead enact a random allocation.

Although the research above calls into question the robustness of assuming socially motivated preferences, recent results complicate this interpretation. In particular, van der Weele, Kulisa, Kosfeld & Friebel (2014) replicated the plausible deniability treatment of Dana et al. (2007) with 2nd movers in a two-person trust game (TG) and a moonlighting game[6] and found no effect of introducing "wiggle room" in either task. Structurally, 2nd movers in these tasks are effectively playing a DG with an endowment determined by the other participant (1st mover) instead of by the experimenter. For example, in the TG, both participants receive the same initial endowment and the 1st mover, the investor, makes a choice of whether to invest any/all of their money by sending it to the other participant, the trustee. Any amount sent by the investor will be increased by a known multiplier and then the trustee faces a decision of whether or not to return any/all of the investment back to the investor (who cannot retaliate, thus placing the trustee in the role of dictator). Finding that trustees do not exploit plausible deniability to keep more money for themselves in the TG (or punish less in the moonlighting game), while they did keep more in Dana et al.'s (2007)

[6] The moonlighting game was a punishment version of the TG. Both players started with a large endowment, the 1st mover could take from the 2nd mover, and then the 2nd mover could punish the 1st mover at a cost.

DG, van der Weele et al. (2014) hypothesize that social preferences are more robust than social image concerns when agents have morally relevant information about their interaction partner (e.g. based on their decision to trust the 2nd mover). In the current work, Study 2 tests this explanation in an exit-version of the DG using participants that dictators presumably already have morally relevant information about – their own friends. However, this approach presupposes that decision making is internally consistent/rationalizable in these settings, which is the focus of Study 1.

Study 1

Study 1 seeks to examine behavioral consistency both within and between social dilemma tasks. To this end, a two-person sequential-move monetary Prisoner's Dilemma (PD) was introduced because 2nd mover choices control for strategic concerns and are thought to be an indicator of stable social preferences (see Figure 1). To measure between-task consistency, participants were instructed that they would be completing two separate tasks that were both, in fact, structurally the same (i.e., PDs). This idea was reinforced by describing the tasks differently (see next section), presenting separate quizzes on each task, and compensating participants from their own and others' decisions in one randomly selected round from within each task. To measure within-task consistency, participants made four decisions in each PD framing that differed in their temptation to defect by lowering the benefit from mutual cooperation and increasing with the benefit from mutual defection.

	Cooperate	Defect
Cooperate	Other person: \$7 **You: \$7**	Other person: \$10 **You: \$0**
Defect	Other person: \$0 **You: \$10**	Other person: \$3 **You: \$3**

	Cooperate	Defect
Cooperate	Other person: \$14 **You: \$14**	Other person: \$17 **You: \$7**
Defect	Other person: \$7 **You: \$17**	Other person: \$10 **You: \$10**

Figure 1. An example of the two different monetary Prisoner's Dilemmas used in Study 1. In both games, a money-maximizing agent's dominate strategy is to defect.

A consistent decision maker was defined as a person whose pattern of choices could be represented by a functional form and analyzed using analytical or psychological game theory. A well-known example of a goal consistent with the analytical approach is inequity aversion[7] (Fehr & Schmidt, 1999). In this model, a person evaluates the distribution of goods between herself and others and derives disutility from feelings of envy if she has received less than average (disadvantageous inequity) as well as something akin to guilt if she receives more (advantageous inequity). Formally, this is represented by:

[7] Moralists who value the outcome of all parties involved (e.g., utilitarians) or the outcome of the worst-off in particular (Rawlsians) would also be categorized as consistent by our definition.

$$U_i = x_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max(x_j - x_i, 0) - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max(x_i - x_j, 0)$$

where x_i represents the utility *i* derives from the receipt of good *x*, *n* is the number of people involved in the distribution, α_i is the envy parameter, and β_i is the guilt parameter. Applied to the first (top) PD in Figure 1, an inequity averse agent will cooperate with a cooperator if 10 – β(10-0) < 7, or, if β > 3/10. However, if β > 3/10, the function predicts that the agent will also cooperate if the benefits from cooperating increase which is our test of within task consistency. Likewise, if if β > 3/10, the model also predicts that the agent will cooperate in the second (bottom) PD in Figure 1 because 17 - β(10) < 14 for β > 3/10. Therefore, if a person cooperates at a given level of temptation in one game, between-task consistency requires that they also cooperate at that level of temptation in the other game.

If, instead, agents are motivated by their beliefs about other players' beliefs/actions (e.g., others' intentions (Rabin, 1993)) or normative expectations (Bicchieri, 2006), this too can be modeled, with some restrictions, using the theory of psychological games (Geanakoplos, Pearce, & Stacchetti, 1989; Battigalli & Dufwenberg, 2009). For example, in Bicchieri's (2006) norm-based utility function[8], agents are motivated by their own outcome as well as by avoiding norm violations. Assuming that a norm to cooperate exists in the PD, in this model a person will cooperate/conform in the first game of Figure 1 if 10 – *k*(7-0) < 7, or, if *k* > 3/7 where *k* represents the person's

[8] See the Supplemental Materials for a more detailed description of this model.

sensitivity to the cooperative norm in this social group. However, if $k > 3/7$, this model also predicts within-task consistency for situations involving larger norm-violations (i.e., those in which the benefits from cooperating increase) as well as between-task consistency because $17 - k(14-7) < 14$ for $k > 3/7$.

Method and Procedures

Participants (N=96, 48% female) were recruited from the Economics subject pool at a large U.S. university. They were offered $5 and the opportunity to earn additional compensation based on their own and others' choices. Participants entered the lab in groups of 10 – 18 and were randomly seated at computer stations separated by dividers. Instructions[9] were then provided on screen, as well as read aloud, stating that there were two separate tasks each involving four rounds of decision making and that final compensation would be determined by one randomly chosen round within each task. The tasks themselves were sequential-move monetary Prisoner's Dilemmas (PDs) framed either in the abstract or as a $10 public/private investment decision (referred to hereafter as a mini-Public Goods Game, mPGG (as in Sigmund, Hauert & Nowak, 2001)). After introducing a task, order counterbalanced, participants completed a brief quiz on the rules and questions were answered in private. Participants were then informed that, to maximize their choices, they would make decisions as both 1st and 2nd mover (only one choice would be randomly selected to be enacted for real compensation) and that 2nd movers would choose via the strategy method – providing a choice in response

[9] Full instructions available in the Supplemental Materials

to each decision the 1st mover could make. Decisions were made in private through the computer interface and participants were informed of the ID number of their partner in each round (after being previously informed that they would never be matched with the same person twice, across either task). The only difference between rounds was the magnitude of the benefit from mutual cooperation or defection (see Figure 2) and this order was randomly set to either be ascending or descending for both tasks. After completing both tasks, participants were asked about their expectations of other players' choices. To reduce cognitive load, these questions were limited to the actual 2nd mover decisions of other players with participants being asked to choose a decile corresponding to the percent of others choosing cooperation in the PD and defection in the mPGG. These predictions were incentivized by paying $1 per correct estimate in one randomly chosen round from each task ($4 possible[10]). Following these predictions, demographic information was collected before final payments were calculated and participants were compensated in private.

Results

2nd Movers

Figure 3 summarizes the decisions of participants choosing 2nd in both tasks. Of primary interest, 80.2% (77/96) of participants were behaviorally consistent as 2nd movers responding to 1st mover cooperation within all four rounds of each game.

[10] For each chosen round, $1 was paid for correctly predicting the percentage of 2nd movers who would cooperate in response to a cooperative 1st mover, and another $1 was paid for their estimate of how many 2nd movers would cooperate in response to 1st mover defection.

Specifically, if and when roughly 80% of participants cooperated within a task, they also cooperated in all other rounds in which it was less personally tempting to defect. In reference to the order of rounds in Figure 1, this means that these participants never switched from cooperation to defection more than once as the games got more tempting.

PD	Cooperate	Defect	mPGG	Public	Private	*k* to ≺ C and Public	β to ≺ C and Public
C	$9, $9	0, 10	Public	18, 18	9, 19	$k > \frac{1}{9}$	$\beta > \frac{1}{10}$
D	10, 0	1, 1	Private	19, 9	10, 10		
C	8, 8	0, 10	Public	16, 16	8, 18	$k > \frac{1}{4}$	$\beta > \frac{1}{5}$
D	10, 0	2, 2	Private	18, 8	10, 10		
C	7, 7	0, 10	Public	14, 14	7, 17	$k > \frac{3}{7}$	$\beta > \frac{3}{10}$
D	10, 0	3, 3	Private	17, 7	10, 10		
C	6, 6	0, 10	Public	12, 12	6, 16	$k > \frac{2}{3}$	$\beta > \frac{2}{5}$
D	10, 0	4, 4	Private	16, 6	10, 10		

Figure 2. Monetary payments resulting from all four possible pairs of choices in each round of both the abstractly labeled Prisoner's Dilemma (PD) and the investment Prisoner's Dilemma, which was a mini-Public Goods Game (mPGG). Payments are listed in pairs (x, y) with the first payment (x) referring to the row player's outcome and the second payment (y) referring to the column player's outcome. The rightmost columns indicate the level of sensitivity to the norm of cooperation (k) and the level of aversion to advantageous inequity (β) that would be necessary for an agent to strictly prefer cooperating with another cooperator in that game based on models by Bicchieri (2006) and Fehr & Schmidt (1999).

Furthermore, of the 77 participants with rationalizable[11] choice patterns, 57.1% (44/77) behaved consistently across tasks as well[12] - they switched from cooperation to defection at the same level of temptation in each task (i.e. acted as if they had a stable social preference/norm parameter). For example, in reference to Figure 1, if a person cooperated in the first game and then defected in the next three games of one task, consistency[11] required also cooperating in the first game and defecting in the following 3 of the second task. Table 1 depicts the frequency of choice patterns that were consistent both within and between tasks. Most notably, 81.8% (36/44) of fully consistent 2nd movers either defected in all 8 rounds or cooperated in all 8 rounds in response to a cooperative 1st mover.

Table 1.

Distribution of consistent 2nd movers in response to 1st mover cooperation

	Never Cooperated	Cooperated Only at Least Tempting Round to Defect	Cooperated Starting at 2nd Most Tempting Round to Defect	Always Cooperated
Frequency (%)	**22 (50%)**	**6 (13.6%)**	**2 (4.6%)**	**14 (31.8%)**

Note. 2nd mover data from the 3rd most tempting round to defect was lost due to a coding error

[11] Rationalizable/consistent in terms of traditional social preference models such as Fehr & Schmidt (1999) as well as the norm-based utility function of Bicchieri (2006)

[12] Of the 19 participants who were inconsistent within one task, only 1 displayed the same pattern of inconsistency across tasks as well.

PD	Cooperate	Defect	1st Cooperated	1st Defected	mPGG	Public	Private	1st Cooperated	1st Defected
C	9, 9	0, 10	**60.4%**	**11.5%**	*Public*	18, 18	9, 19	**59.4%**	**8.3%**
D	10, 0	1, 1	**39.6%**	**88.5%**	*Private*	19, 9	10, 10	**40.6%**	**91.7%**
C	8, 8	0, 10	**50***	**9.4**	*Public*	16, 16	8, 18	**NA***	**4.2**
D	10, 0	2, 2	**50***	**90.6**	*Private*	18, 8	10, 10	**NA***	**95.8**
C	7, 7	0, 10	**43.8**	**3.1**	*Public*	14, 14	7, 17	**40.6**	**2.1**
D	10, 0	3, 3	**56.3**	**96.9**	*Private*	17, 7	10, 10	**59.4**	**97.9**
C	6, 6	0, 10	**41.7**	**2.1**	*Public*	12, 12	6, 16	**30.2**	**4.2**
D	10, 0	4, 4	**58.3**	**97.9**	*Private*	16, 6	10, 10	**69.8**	**95.8**

Figure 3. Percentage of 2nd movers choosing to cooperate or defect in response to the decision of the 1st mover (elicited via the strategy method). *Due to a coding error, responses to 1st mover cooperation in the investment Prisoner's Dilemma with a multiplier of 1.6 were lost and a majority of these responses in the abstract Prisoner's Dilemma were also lost (50*% = 11 of 22 participants).

Empirical Expectations

Figures 4 and 5 display the distribution of participants' predictions regarding the choices of other 2nd movers in their experimental session based on their own 1st mover and 2nd mover choices in the abstract Prisoner's Dilemma and the mini-Public Goods Game. Overall, we find a trend whereby participants who defected as 1st or 2nd movers were more likely to believe that other participants would also defect as 2nd movers responding to a cooperative 1st mover. In the PD, a Kolmogorov-Smirnov test revealed that 1st movers who defected held significantly different empirical beliefs than cooperators in the least tempting round to defect, D = .34, p = .009, the 2nd least tempting, D = .32, p = .01, and the 2nd most tempting, D = .3, p = .02, but not in the most tempting condition ($6 from mutual cooperation vs. $4 from mutual defection), D = .24, p = .13. However, 2nd movers who defected only displayed significantly different empirical beliefs about other 2nd movers in the least tempting round to defect, D = .46, p < .001. This pattern of findings was reversed in the mPGG where only beliefs about the 2nd least tempting game differed among 1st movers, D = .33, p = .008, but beliefs among 2nd movers in all three games were significantly different between defectors and cooperators, D = .4, p = .002; D = .44, p < .001; and D = .32, p = .01 in order of decreasing temptation to defect.

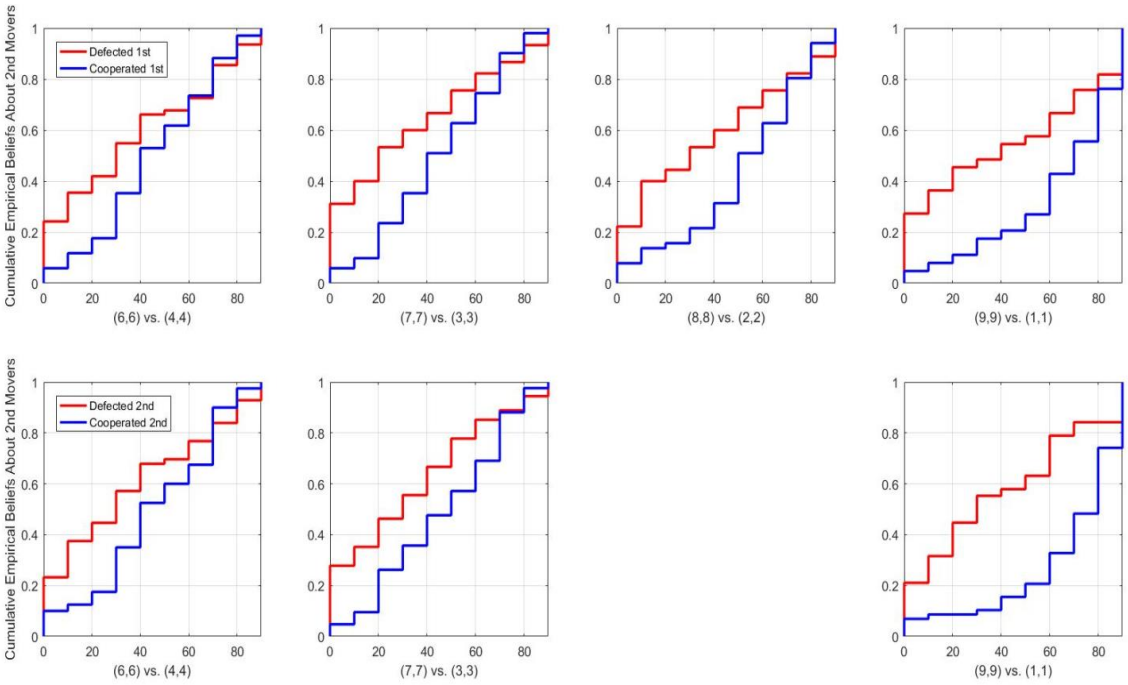

Figure 4. Cumulative distribution plots of empirical expectations as a function of 1st mover decision (top row) and 2nd mover decision (bottom row) in the abstract Prisoner's Dilemma. Beliefs correspond to the expected percentage of other participants choosing cooperation as the 2nd mover in response to a cooperative 1st mover in that round. Predictions were made by choosing one of 10 equally spaced deciles, e.g. 0 corresponds to the decile from 0 – 10% and 20 corresponds to the decile from 20 – 30%. The x-axis labels correspond to the payments from mutual cooperation and mutual defection for that round.

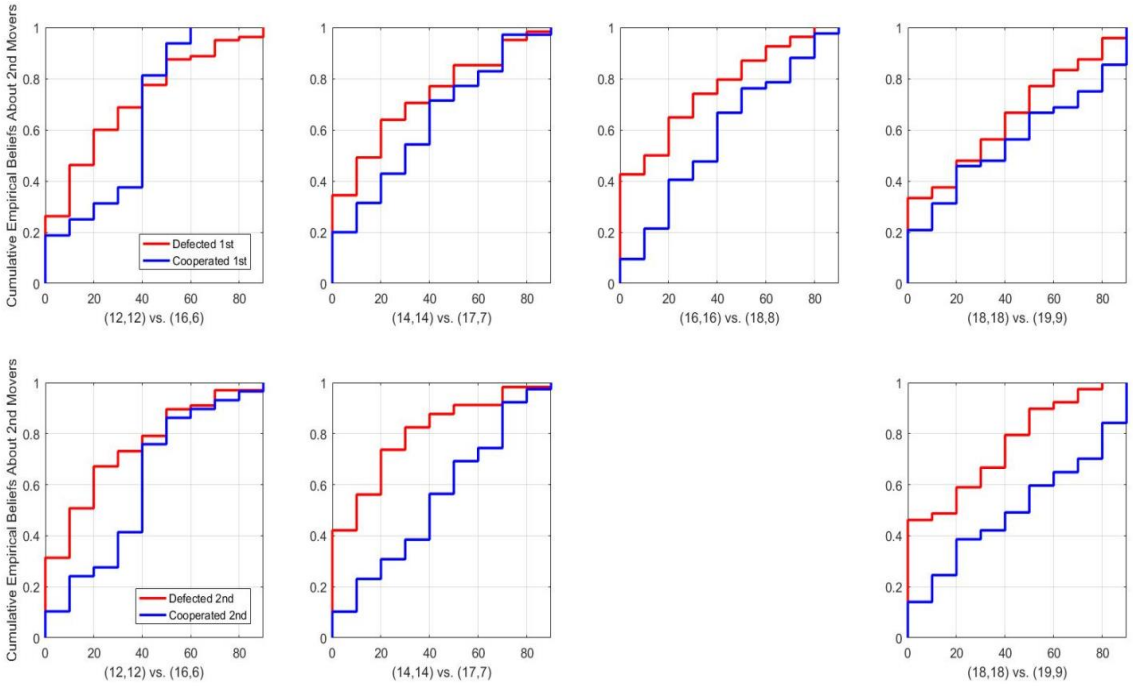

Figure 5. Cumulative distribution plots of empirical expectations as a function of 1st mover decision (top row) and 2nd mover decision (bottom row) in the mini-Public Goods Game. Beliefs correspond to the expected percentage of other participants choosing cooperation as the 2nd mover in response to a cooperative 1st mover in that round. Predictions were made by choosing one of 10 equally spaced deciles, e.g. 0 corresponds to the decile from 0 – 10% and 20 corresponds to the decile from 20 – 30%. The x-axis labels correspond to the payments from mutual cooperation and mutual defection for that round.

Discussion

Rational choice models have been invoked to explain cooperative behavior in experimental social dilemmas, but these theories are only tenable if decision making is internally consistent. Study 1 investigated this basic assumption both within and between two-person sequential-move Prisoner's Dilemmas that were framed differently yet remained structurally equivalent. Of particular interest was 2nd mover behavior in response to a cooperative 1st mover, because strategic motives/fears are eliminated under these circumstances. In response to cooperative 1st movers, the preferences of

2nd movers were fairly well-behaved within a particular social dilemma task (~80% consistent within both tasks), but less-so between tasks that had the same incentive structure (~57% of that 80%). The within-task consistency is arguably good news for rational approaches, especially in light of findings from related tasks that show, e.g., that significantly less than 100% of participants choose options which are both personally and socially dominate (Kümmerli, Burton-Chellew, Ross-Gillespie & West, 2010) – presumably due to confusion, cognitive errors, fatigue, boredom, etc. Likewise, the decrease in consistency across tasks could be explained by errors, novelty-seeking, moral licensing, or another competing decision making process. Supporting this view is the finding that 2/3 (22/33) of the participants who were consistent within both tasks but not between both tasks switched from cooperation to defection in the second task at only one level of temptation different from their switching point in the first task (and 5 participants switched from complete cooperation to complete defection!). A less favorable interpretation of these findings is that only 45% of the participants (44/96) behaved consistently within and between social dilemmas under conditions which should maximally encourage consistency (binary choice two-person dilemmas conducted in the same time period and presented in either ascending or descending order of temptation). Furthermore, the vast majority of completely consistent choosers either defected or cooperated in all 8 rounds which leaves little room for a more nuanced social preference/norm models to explain behavior.

Study 2

Although the results of Study 1 provide mixed evidence, even finding that strategic situations are amenable to rational choice analysis leaves open the question of whether pro-social actions reflect an underlying preference for fairness or simply a preference to gain/avoid social approval/disapproval. Dana et al. (2006, 2007) provided evidence in favor of one answer (conforming to expectations), while van der Weele et al. (2014) more recently found that pro-social behavior was unaffected by a manipulation that decreased the social image consequences of a self-interested choice. The authors attributed this finding to the fact that their decision makers gained morally relevant information about the other person through their interaction and concluded that preferences for fair outcomes may be more robust than previously estimated. However, van der Weele et al.'s (2014) design confounded the additional information gained about the other person with the reciprocal social context in which they interact – an element that was absent in Dana et al.'s original designs (2006, 2007) and may establish expectations of equal treatment (see, e.g., Bicchieri, Xiao & Muldoon, 2011). Therefore, it could be the fact that this decision is embedded into a richer socio-relational context, with its associated norms, that motivates reciprocity as opposed to the morally relevant information gained about the "type" of agent affected by their choice. Study 2 seeks to disentangle these explanations using a modification of the dictator game from Dana et al. (2006).

Method and Procedures

Participants (N=486, 44% female) were recruited from Amazon's Mechanical Turk Marketplace in exchange for a small payment[13]. Between-subjects, participants played a dictator game with either a random participant from an unrelated study or a friend whom they listed at the start of the survey. Those in the friend condition were asked to provide the e-mail address and first name of a friend to possibly receive a small electronic Amazon gift card in connection with the study, but no further information was provided about the task at that point. Participants also provided information on how long they knew their friend and how connected they felt to their friend on core values (based on Bartels, Kvaran & Nichols, 2013). In both conditions, instructions were provided through Qualtrics survey software and described the dictator game in the abstract followed by quiz questions on the rules. Next, participants were informed of the specifics - they were being given $10 to allocate between themselves and the other person (in $1 increments). In addition, between-subjects, participants were informed that their recipient would receive an e-mail message along with the allocation either explaining the circumstances leading to their payment or not (similar to Study 2 of Dana, Cain & Dawes, 2006). Specifically, the e-mail sent in the full information condition began with, "A friend[14] of yours recently participated in an online study where they were given $10 and asked to allocate that money between themselves and a friend.

[13] Qualifications were U.S. residence and an approval rate greater than or equal to 90% on prior work.
[14] The stranger condition e-mail read as follows, "In a recent online study, a participant was given $10 and asked to allocate that money between themselves and a randomly assigned person from a previous study, which was you. This person decided to keep $___ while allocating $___ to you." The wording in the "no information" condition was the same for both types of recipients.

You were chosen as their friend/recipient and they kept $___ while allocating $___ to you." Alternatively, the e-mail in the no information condition did not describe the circumstances leading to the recipient's payment, "We run research studies through Amazon and as a result of a recent study you have been chosen to receive $___ and are therefore [not] receiving an electronic Amazon gift card for this amount." This resulted in a 2 (recipient: stranger vs. friend) × 2 (information: full vs. none) between-subjects design. After making their allocation decision, participants were presented with an unexpected opportunity to "exit" the game by accepting a $9 personal bonus payment (with $0 going to the recipient) along with a guarantee that the recipient would never be sent a message about the task/their initial $10 allocation. Following the "exit" decision participants completed several demographic questions before submitting their survey.

If altruistic behavior in the DG is primarily caused by socially motivated preferences, the increase in morally relevant information about the recipient in the friend condition should result in an interaction whereby dictators only exit on strangers who are given no information about the reason for their gift card amount. However, if seemingly altruistic behavior in the DG is a function of the relationship context and the expectations/norms associated with it, dictators are predicted to exit the game when information about the task is provided to the recipient independent of whether the person is a friend or stranger, and not otherwise.

Results

Initial allocations

Figure 6 depicts the distribution of dictator allocations for each condition as well as the proportion who later exited the game for $9. Table 2 displays the conditional and marginal mean amount kept by dictators. A 2 × 2 factorial ANOVA conducted on the amount kept revealed no main effect of recipient type, $F(1, 482) = .33$, $p = .57$, no main effect of information, $F(1, 482) = .1$, $p = .753$, and no interaction, $F(1, 482) = .22$, $p = .64$. Similarly, a Kruskal-Wallis H test showed no significant difference between mean ranks, $\chi^2(3, N=486) = .81$, $p = .85$.

Table 2.

Average (standard deviation) and median kept by dictators

	Full Information	No Information	
Friend	*M*: 7.55 (3.17) *Mdn*: 10	*M*: 7.74 (2.87) *Mdn*: 10	7.65 ($3.02)
Stranger	*M*: 7.81 (2.48) *Mdn*: 9	*M*: 7.77 (2.20) *Mdn*: 8	7.79 ($2.34)
	7.69 (2.82)	7.76 (2.53)	

Note. Mean (standard deviation) and median of the amount of money kept, in dollars, by the dictators from the initial $10 endowment.

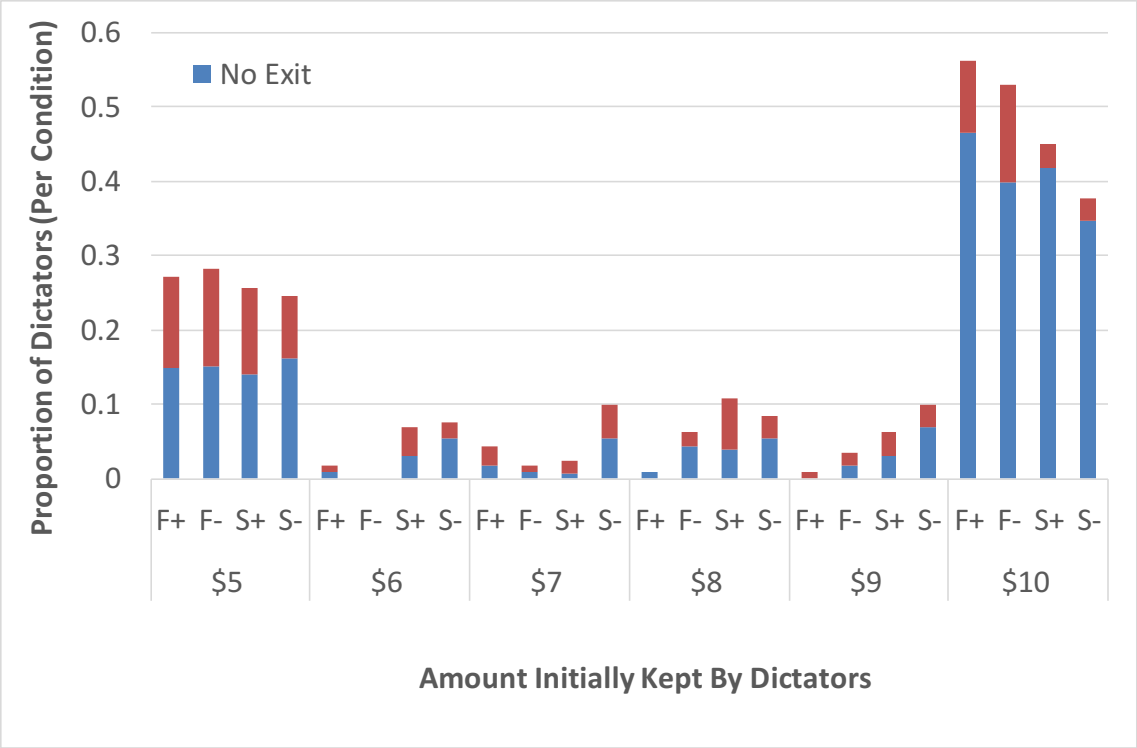

Figure 6. Frequency of dictator allocations (amount kept) by condition and exit decision. Recipient condition is denoted by F (friend) and S (stranger) while the information condition is denoted by + (full information) or – (no information). The 5% (24/486) of dictators who initially kept less than $5 are not shown for visual clarity, but the full distribution can be found in the Supplemental Materials.

Exit decisions

Table 3 reports the percentage of participants choosing the exit option in each condition. A logistic regression[15] confirmed that exiting did not differ significantly as a function of recipient type, OR = 1.16, 95% CI [.78, 1.71], the information provided to recipients, OR = 1.08, 95% CI [.73, 1.60], nor their interaction, OR = .61, 95% CI [.78, 1.71]. However, consistent with Broberg, Ellingsen & Johannesson (2007), we found

[15] Recipient variable was dummy coded with the stranger condition as the reference group and the information variable was dummy coded with no information as the reference group.

that participants choosing to exit kept significantly less for themselves in their initial dictator choices, M_{kept} = \$6.87 and \$8.08 for those who exited and did not exit respectively, $t(484) = 4.63$, $p < .001$.

Table 3.

Percent of dictators choosing the "exit" option

	Full Information	No Information	
Friend	29.0%	32.7%	30.8%
Stranger	31.0%	24.6%	27.8%
	30.0%	28.4%	

Note. Proportion of participants in each condition who chose to "exit" by accepting a \$9 personal payment instead of having their dictator choice enacted and a message sent to their recipient.

Discussion

The purpose of Study 2 was to examine strategic behavior when the influence of morally relevant information about one's interaction partner was isolated from the influence of the socio-relational context in which that information is typically encountered (as in van der Weele et al., 2014). While several patterns of behavior were hypothesized, the results do not seem to readily align with any simple explanations. Table 4 outlines predictions from several stylized accounts of social motivation in terms of expected initial dictator gifts ranked from the most generosity-inducing condition (1) to the least (4). The stars indicate situations in which a theory could predict exiting for \$9. There is a lot of gray area in the table, e.g., a norm-conformer may believe there are different norms for friends and strangers or not, may have internalized norms for friends even in the absence of expectations or not, etc. Likewise, a person motivated by reputational concerns may still give to a friend that receives no information about the

reason for their gift card because she can take credit for it after the experiment and still get the gain in esteem, or it may not be worth the effort. However, no one theory, no matter how loosely applied, appears to predict a failure to replicate Dana et al.'s (2006) exit findings in the stranger conditions as well as the relatively equal giving/exiting across all initial conditions.

Table 4.

Qualitative theoretical predictions for Study 2

	$-Maximizer		**Norm-Conformer**		**Max(Reputation)**		**Inequity Averse**	
	+	-	+	-	+	-	+	-
F	**4**	**4**	**1***	**3**	**1**	**2**	**2.5**	**2.5**
S	**4**	**4**	**2***	**4**	**3**	**4**	**2.5**	**2.5**
	Info-Based		**Moral Virtue**		**No Responsibility**		**Friend Lying**	
	+	-	+	-	+	-	+	-
F	**1.5**	**1.5**	**2.5**	**2.5**	**1***	**2.5***	**1, 4**	**1, 4**
S	**3.5**	**3.5**	**2.5**	**2.5**	**2.5***	**4***	**2.5**	**2.5**

Note. Info-Based refers to making a conditional decision based on the morally relevant information the dictator has about the recipient. Moral Virtue is meant to capture utilitarian or Rawlsian preferences. No responsibility refers to those who act altruistically to avoid feeling responsible for low allocations; and, Friend Lying refers to those who include their own (1) or a fake (4) e-mail in the friend condition[16].

[16] The measure of "connectedness" to the friend was included, in part, as a subtle of dishonesty but no correlation was found between how similar a person viewed themselves to their friend and their initial dictator allocation, r = -.07, p = .30, nor their exit decision, r = .02. , p = .73.

General Discussion

Psychological explanations for strategic behavior may be rooted in instrumental goals that are consciously or unconsciously pursued in social situations. If so, decision making should conform to assumptions regarding internal consistency and researchers need only map out this distribution of motives. However, the current work sheds doubt on the universality of this approach. In particular, under arguably ideal circumstances, only half of participants displayed rationalizable behavior both within and between social dilemmas in Study 1. Moreover, the consistent half was made up almost entirely of those that either always defected or always cooperated which leaves little room for more exotic models/motivations. However, this finding opens the door to the question of what decision process is actually generating the "inconsistent" data. While the potential for boundedly rational explanations is infinite, a research program organized around heuristic-based accounts (e.g., Rand et al., 2014) or perceptually-based ones (e.g., Jiang, Potters & Funaki, 2014) may prove more fruitful.

However, another alternative is that the rational choice approach requires a paradigm shift. In particular, the methodological individualism that has produced so many insights may be running up against the reality of a non-reductive world. For example, perhaps quantum approaches to game theory with their entangled preferences better approximate the psychology leading to social decisions (e.g., Eisert, Wilkens & Lewenstein, 1999). Or maybe it is time the field follows economist Vernon Smith in rediscovering first principles from Adam Smith's *Theory of Moral Sentiments*. Smith (2015) argues that social behavior is best described as rule-based, which is similar

to Bicchieri (2006) except she consciously acknowledges the move towards individualism and embeds rules into conditional preferences – choices made consistently conditional on the agent's beliefs about the social context of her choice. However, Smith (2015) adopts an older philosophical position whereby preferences are not just socially constructed in the sense that society provides the inputs into a person's then-individual preference, but that preferences can only be accurately understood within their social context. As Smith (2015) phrases it, "A rule maps context, inclusive of the available set of outcome payoffs, into an action, but the resulting outcome only has meaning in the context (circumstance) that led to the action and is not separable from the context. Equilibrium, if the concept applies, is in rule space and stems from empathy, but more significantly, from mutual empathy as in the *Theory of Moral Sentiments*" p. 186. So, when researchers attempt to measure non-strategic preferences via the dictator game (Study 2) or strategy method (Study 1), they may not be uncovering the "true" nature of the individual but rather merely observing the noise resulting from a choice out of context (or rules out of equilibrium).

References

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*(2), 737-753. doi:10.1111/1468-0262.00302

Bartels, D. M., Kvaran, T., & Nichols, S. (2013). Selfless giving. *Cognition,129*(2), 392-403. doi:10.1016/j.cognition.2013.07.009

Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*,*144*, 1-35. doi: 10.1016/j.jet.2008.01.004

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

Binmore, K. (2010). Social norms or social preferences? *Mind & Society*, *9*, 139-157. doi:10.1007/s11299-010-0073-2

Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior,72*(2), 321-338. doi:10.1016/j.geb.2010.09.008

Bowles, S., & Gintis, H. (2011). *A cooperative species: Human reciprocity and its evolution*. Princeton, NJ: Princeton University Press.

Broberg, T., Ellingsen, T., & Johannesson, M. (2007). Is generosity involuntary? *Economics Letters,94*(1), 32-37. doi:10.1016/j.econlet.2006.07.006

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. New York, N.Y: Russell Sage Foundation.

Dana, J., Cain, D., & Dawes, R. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational behavior and human decision processes*, *100*, 193- 201. doi: 10.1016/j.obhdp.2005.10.001

Dana, J., Weber, R., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*, 67-80. doi: 10.1007/s00199-006-0153-z

Eisert, J., Wilkens, M., & Lewenstein, M. (1999). Quantum games and quantum strategies. *Physical Review Letters,83*(15), 3077-3080. doi:10.1103/physrevlett.83.3077

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

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *American Economic Review,97*(5), 1858-1876. doi:10.1257/aer.97.5.1858

Jiang, T., Potters, J., & Funaki, Y. (2015). Eye-tracking social preferences. *Journal of Behavioral Decision Making,29*(2-3), 157-168. doi:10.1002/bdm.1899

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, *1*, 60-79.

Gintis, H. (2009). *The bounds of reason: Game theory and the unification of the behavioral sciences*. Princeton, N.J: Princeton University Press.

Gintis, H., &Helbing, D. (2015). Homo socialis: An analytical core for sociological theory. *Review of Behavioral Economics,2*(1-2), 1-59. doi:10.1561/105.00000016

Handgraaf, M., Van Dijk, E., Vermunt, R., Wilke, H., & De Dreu, C. (2008). Less power or powerless? egocentric empathy gaps and the irony of having little versus no power in social decision making. *Journal of Personality and Social Psychology*, *95*(5), 1136-1149. doi: 10.1037/0022-3514.95.5.1136

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience,10*(12), 1625-1633. doi:10.1038/nn2007

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica,47*(2), 263. doi:10.2307/1914185

Kummerli, R., Burton-Chellew, M. N., Ross-Gillespie, A., & West, S. A. (2010). Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games. *Proceedings of the National Academy of Sciences,107*(22), 10125-10130. doi:10.1073/pnas.1000829107

Lehmann, L., Keller, L., West, S., & Roze, D. (2007). Group selection and kin selection: Two concepts but one process. *Proceedings of the National Academy of Sciences,104*(16), 6736-6739. doi:10.1073/pnas.0700662104

Mao, A., Dworkin, L., Suri, S., & Watts, D. J. (2017). Resilient cooperators stabilize long-run cooperation in the finitely repeated Prisoner's Dilemma. *Nature Communications,8*, 13800. doi:10.1038/ncomms13800

Marr, D. (1982). The philosophy and the approach. In, *Vision: a computational investigation into the human representation and proceedings of visual information*. San Francisco: Freeman.

Oppenheimer, D. M., & Kelso, E. (2015). Information processing as a paradigm for decision making. *Annual Review of Psychology,66*(1), 277-294. doi:10.1146/annurev-psych-010814-015148

Pylyshyn, Z. (1999). What's in your mind? In E. Lepore & Z. Pylyshyn (Eds.), What is cognitive science. Oxford: Blackwell.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, *83*(5), 1281-1302.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications,5*. doi:10.1038/ncomms4677

Rilling, J. K., Demarco, A. C., Hackett, P. D., Thompson, R., Ditzen, B., Patel, R., & Pagnoni, G. (2012). Effects of intranasal oxytocin and vasopressin on cooperative behavior and associated brain activity in men. *Psychoneuroendocrinology,37*(4), 447-461. doi:10.1016/j.psyneuen.2011.07.013

Samuelson, P. A. (1937). A note on measurement of utility. *The Review of Economic Studies,4*(2), 155. doi:10.2307/2967612

Savage, L. J. (1954). The foundations of statistics. New York: John Wiley & Sons. doi:10.1002/nav.3800010316

Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary theory and the ultimate–proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science,6*(1), 38-47. doi:10.1177/1745691610393528

Sigmund, K., Hauert, C., & Nowak, M. A. (2001). Reward and punishment. *PNAS*, *98*(19), 10757-10762. doi: 10.1073/pnas.161155698

Smith, V. L. (1962). An experimental study of competitive market behavior. *Journal of Political Economy,70*(2), 111-137. doi:10.1086/258609

Smith, V. L. (2015). Adam smith: Homo socialis, yes; social preferences, no; reciprocity was to be explained. *Review of Behavioral Economics,2*(1-2), 183-193. doi:10.1561/105.00000028

Smith, A. (1793). *The Theory of Moral Sentiments*. Basil: Tourneisen.

Suleiman, R. (1996). Expectations and fairness in a modified ultimatum game. *Journal of Economic Psychology*, *17*, 531-554.

van der Weele, J. J., Kulisa, J., Kosfeld, M., & Friebel, G. (2014). Resisting moral wiggle room: How robust is reciprocal behavior? *American Economic Journal: Microeconomics,6*(3), 256-264. doi:10.1257/mic.6.3.256

van Honk, J., Montoya, E. R., Bos, P. A., Vugt, M. V., & Terburg, D. (2012). New evidence on testosterone and cooperation. *Nature,485*(7399), E4-E5. doi:10.1038/nature11136

Volk, S., Thöni, C., & Ruigrok, W. (2012). Temporal stability and psychological foundations of cooperation preferences. *Journal of Economic Behavior & Organization,81*(2), 664-676. doi:10.1016/j.jebo.2011.10.006

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., . . . Simunovic, D. (2013). Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organizational Behavior and Human Decision Processes,120*(2), 260-271. doi:10.1016/j.obhdp.2012.06.002

Yang, Y., Onderstal, S., & Schram, A. (2016). Inequity aversion revisited. *Journal of Economic Psychology,54*, 1-16. doi:10.1016/j.joep.2015.12.009

Part 2: A Social Species

There is evidence that from a very early age (Kovacs, Teglas & Endress, 2010), and possibly to a degree not observed in non-human primates (Horner & Whiten, 2005), people begin representing and being influenced by the beliefs and practices of humans around them. This tendency, along with others[1], leads to the emergence and maintenance of culturally specific norms – rules of behavior that are a function of what is commonly done and/or what is commonly approved and disapproved by relevant others (Cialdini, Reno & Kallgren, 1990). Norms dictate everything from the conventions of fashion to the rituals of religion, but the current work focuses on their particular influence in social dilemmas – situations in which personal interests are at odds with collective interests. In this domain, (internalized) social norms and norm-enforcement mechanisms may sustain cooperation towards collective welfare, but testing this claim requires an operational definition of norms and a model of their impact on behavior. The purpose of the present work is to test an operationalization of norms provided by Cristina Bicchieri in her 2006 book, *The Grammar of Society: The Nature and Dynamics of Social Norms*.

Norms operate by establishing shared expectations regarding the rights, duties, and actions/consequences of agents in social roles. Bicchieri (2006) refers to beliefs about the anticipated actions of others in a social role as *empirical expectations* and defines *descriptive norms* as those requiring only these beliefs to induce conformity. In

[1] See, e.g., Chudek & Henrich, 2011

practice, this means that descriptive norms are sufficient to facilitate coordination among aligned interests, such as which side of the street to drive on. Meanwhile, Bicchieri (2006) refers to beliefs about the rights and duties associated with social roles as *normative expectations*, which correspond to the anticipated approval or disapproval by relevant others towards actions in a role. *Social norms* are then defined as those requiring both empirical and normative expectations to induce conformity. Bicchieri (2006) argues that these are necessary for motivating cooperation among competing interests, such as resisting the urge to rubberneck once driving on the "right" side of the road. However, unlike descriptive norms to coordinate, social norms to cooperate require the additional condition that agents either view others' normative expectations as legitimate (i.e., have internalized the norm) or expect that conformity/violations will be rewarded/punished either symbolically through gossip and the gain/loss in social status, or non-symbolically through inclusion/exclusion and material rewards/sanctions (Bicchieri, 2006; Gintis & Helbing, 2015; Andrighetto, Grieco & Tummolini, 2015).

More formally, Bicchieri (2006) defines a social norm for a given population *P* as a behavioral rule *R* for situations of type *S* (where *S* can be represented as a mixed-motive interaction) if there exists a sufficiently large subset of conditional norm followers, $P_{cf} \subseteq P$, such that, for each individual $i \in P_{cf}$:

1) *i* knows that a rule *R* exists and applies to situations of type *S*
2) *i* prefers to conform to *R* in situations of type *S* on the condition that:
   a) *i* believes that a sufficiently large subset of *P* conforms to *R* in situations of type *S* (*empirical* expectations);

   and either
   b) *i* believes that a sufficiently large subset of *P* expects *i* to conform to *R* in situations of type *S* (normative expectations);

or

b') *i* believes that a sufficiently large subset of *P* expects *i* to conform to *R* in situations of type *S*, prefers *i* to conform, and may sanction behavior.

While the above conditions are necessary for a social norm to exist, a social norm *R* is followed by population *P* if there exists a sufficiently large subset of norm followers, $P_f \subseteq P_{cf,}$ such that, for each individual $i \in P_f$, conditions 2(a) and either 2(b) or 2(b') are met for *i* and, as a result, *i* prefers to conform to *R* in situations of type *S*. In words, P_{cf} is the subset of a group who know about a norm and have a conditional preference for conforming to it, and P_f is the subset of conditional followers who believe their empirical and normative expectations have been met and actually do conform, though these thresholds can be heterogeneous in the population and may vary across norms within the same person. The definition of a descriptive norm is the same as above except conditions 2(b) or 2(b') do not need to be met for norm compliance.

Having defined norms, Bicchieri (2006) imbeds the concept into the rational actor model of individual decision making to predict behavior. Rational choice theories model decision making as a function of an individual's beliefs (subjective priors), preferences, and constraints but traditionally do not attempt to explain the underlying source of these beliefs and motives (Binmore, 2010). The framework of analytical game theory is then used to analyze the behavior of rational agents in strategic interactions - those where each individual's outcome is a function of both their own and others' choices, such as in social dilemmas. Empirical tests of game theoretic predictions began with the simplifying assumption that preferences were motivated by narrow short-term

self-interest, but a recent trend in behavioral economics has sought to improve the descriptive accuracy of these models by defining which *social preferences* can be explicitly modeled as arguments in an individual's objective function (e.g., Rabin, 1993; Fehr & Schmidt, 1999). Alternatively, Bicchieri (2006) argues[2] that the behavioral regularities in lab games that the social preference approach seeks to model could be capitalizing on a different source of behavioral consistency – that of following social norms. This is to say that, within a particular social role or social group, and over a short enough time scale, (internalized) social norms will manifest themselves as stable social preferences that conform to the tenets of rational choice theories, but that this approach will systematically fail at predicting behavior across domains, reference groups, and time periods.

Because Bicchieri's definition of norms assumes that motivation is conditional on a person's beliefs about what is commonly done and what is commonly approved of, her approach necessarily falls under the purview of psychological game theory instead of traditional game theory (see Geanakoplos, Pearce, & Stacchetti, 1989; Battigalli & Dufwenberg, 2009). The defining characteristic of psychological games is that payoffs depend on beliefs (about others' choices and beliefs) and not just on the actions players take. This means that tests of Bicchieri's (2006) model require measuring both empirical and normative expectations along with an individual's sensitivity to the active norm(s) in the environment (i.e. their threshold for conforming). With this in mind, Bicchieri and

[2] A position held by Binmore (2010), Gintis (2010), Kimbrough & Vostroknutov (2013), as well as Gintis & Helbing (2015).

colleagues have experimentally elicited beliefs with a focus on tasks involving norms of fairness, trust, and reciprocity. A sampling of the findings from this group includes results suggesting that reciprocity is a norm but trusting is not (Bicchieri, Xiao, & Muldoon, 2011), that empirical expectations for fairness are more motivating than normative ones when they are at odds with one another (Bicchieri & Xiao, 2009), that social norms of fairness are distinct from personal norms of fairness (Chavez & Bicchieri, 2013), and that norms can be reinforced through 3rd party rewards and/or punishments (Chavez & Bicchieri, 2013). Moreover, the influence of social expectations in pro-social settings has been supported by independent research groups in a variety of different cultures[3].

For instance, Hauge (2016) employed the dictator game (DG) to measure the influence of beliefs in a non-strategic setting. In the DG, participants are organized into pairs and one person, the dictator, is given an endowment of money, 120 Norwegian Krone (NOK) in this case, which she may divide between herself and the other person, the recipient. The recipient makes no decisions in this task and has no means of protesting the allocation. Because participants are paired anonymously and payments are made in private, any deviation by the dictator from keeping all of the endowment is often viewed as a measure of her social preference. However, Hauge (2016) had dictators make allocation decisions first in the absence of expectations and then as a function of the empirical expectation (the average dictators would give) and normative

[3] See, e.g., Dana, Cain & Dawes, 2006; Dana, Weber & Kuang, 2007; Dufwenberg, Gachter, & Henning-Schmidt, 2011; Kimbrough & Vostroknutov, 2013; Andrighetto et al., 2013; Yamagishi et al., 2013; and Andrighetto, Grieco & Tummolini, 2015

expectation (the "morally right" amount to give) of their recipient[4] in two subsequent games (order counterbalanced, new recipient each game). Specifically, dictators provided a conditional allocation for 3 cases in which the recipient either expected [or reported that the morally right thing was for] dictators to: 1) Give nothing (low belief); 2) Give something but less than 50% (medium belief); or 3) Give 50% (high belief). At the end of the experiment, the dictator allocation associated with the recipient's actual beliefs was enacted. The main finding was that dictators were sensitive to both empirical and normative expectations. When the recipient's beliefs were unknown, dictators gave on average 48.8 NOK, but this average decreased to 29.7 and 23.4 for low empirical and normative beliefs respectively, while increasing to 39.5 and 36.4 for medium beliefs, and 46.8 and 48.1 for high beliefs.

While Hauge (2016) measured empirical expectations (plus personal normative beliefs) and explicitly provided this information to dictators, Xiang, Lohrenz & Montague (2013) demonstrated the power of descriptive norms learned behaviorally. In particular, the authors trained participants in a neuroimaging study on a specific descriptive norm during a repeated $20 ultimatum game (UG). The UG has the same structure as the DG, except the recipient ("responder") of a dictator ("proposer") allocation has the opportunity to "reject" the allocation and leave both parties with $0 (if not, they "accept" the allotted division, as is). In one condition, the first 30 offers in the UG were drawn from a normal distribution with a very unfair mean ($4), but the

[4] Recipients were incentivized to provide accurate empirical expectations of average dictator behavior, but no such procedure exists for eliciting personal normative beliefs so the authors paid each recipient a small fee to report thoughtfully.

next 30 offers were drawn from a medium mean distribution ($8). In the other condition, the first 30 offers were drawn from a hyper-fair mean ($12) while the next 30 were drawn from the same medium mean as in the first condition. As predicted by a norm-based account, when faced with the second 30 offers drawn from the medium distribution, participants who learned the unfair norm more frequently accepted offers in the range of $6-$8 than those trained on the hyper-fair norm. This type of variable pro-sociality is difficult to accommodate in traditional social preference models that focus strictly on the distribution of outcomes in predicting choice (because these outcomes are the same across conditions in the final 30 rounds).

The current work seeks to extend these tests to the norm of cooperation by examining whether first and second order beliefs about norms mediate two often cited effects in the social dilemma literature that are also problematic for some models of strategic decision making[5]. Study 1 examines a situation where norms may be influenced by the social labels of the task and Study 2 explores a situation where norms may differ by group affiliation. The contribution of these studies is the elicitation and testing of beliefs. Having a measure of participants' beliefs is essential to understanding decision making in psychological games but was absent in the original research. Although the work above has informed this debate, previous work related to social dilemmas has not measured both the empirical and normative expectations necessary for the instantiation of a norm in Bicchieri's (2006) model.

[5] In particular, those focusing solely on distributional concerns

Study 1

Study 1 investigates whether cues in the environment that have been shown to affect cooperative behavior do so via their systematic effects on empirical and/or normative expectations about others' behavior and beliefs. An often cited effect in the social dilemma literature is that socially labeling a two-person version, known as the prisoner's dilemma (PD), as either a "Community Game" or a "Wall Street Game" shifts behavior towards cooperation or defection respectively (see, e.g., Ross & Ward, 1996; Liberman, Samuels, & Ross, 2004). Ellingsen, Johannesson, Mollerstrom & Munkhammar (2011) replicated this effect and then tested the influence of expectations behaviorally by implementing additional PD designs which restricted the choices, the observability of choices, or the order in which choices were made. All three manipulations eliminated the framing effect which led the authors to conclude that social labels/frames act as coordination devices instead of affecting genuine concern for the welfare of the other person or the anticipated esteem gains from cooperation. Study 1 seeks to test this same hypothesis not by manipulating beliefs, but through measuring empirical and normative expectations as defined by Bicchieri (2006).

Method and Procedures

Participants (N=258, 58% female) were recruited from Amazon's Mechanical Turk Marketplace in exchange for a small payment[6]. Following several pilot studies (see

[6] Qualifications were U.S. residence, an approval rate greater than or equal to 70%, and having completed less than or equal to 50 assignments on MTurk. The last qualification was included in response to concerns raised by Chandler, Mueller & Paolacci (2013) suggesting that more experienced MTurk workers may be less susceptible to subtle framing effects due to repeated exposure to online games.

Supplemental Materials), the original social labeling manipulation (Ross & Ward, 1996) was augmented to strengthen its effect and improve statistical power. In particular, after introducing the social labels for the game (see Figure 1), as in the original manipulation ("Wall Street Task" vs. "Community Task"), participants were shown a short video clip documenting either the competitive nature of day trading or the societal focus of a community action group before receiving instructions on the task itself. In addition, not only was the title of the game socially framed, but the choices themselves were labeled, "Stock A" vs. "Project A", and an image of either a stock ticker or a community circle was transparent in the background. After reading instructions about the simultaneous-move Prisoner's Dilemma, participants completed quiz questions on the rules and possible outcomes from task. Contrary to the usual motivation to minimize experimenter demand in comprehension checks, the two questions testing knowledge of the payoff matrix highlighted the dominate money maximizing strategy to ensure equal knowledge of the tradeoff from cooperating[7].

[7] Viewed as particularly important given the (presumed) inexperience of the participants sampled

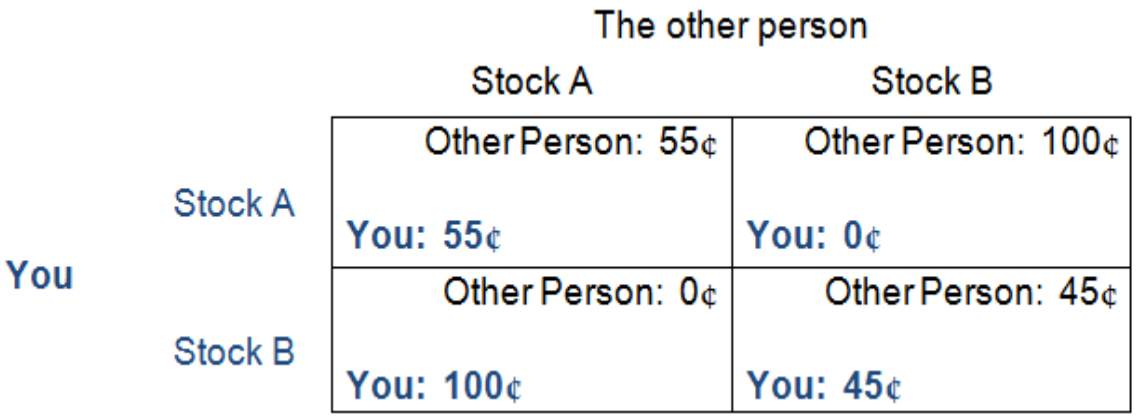

Figure 1. An example of how the payoff matrix in Study 1 was socially framed in the Wall Street condition. The same payoffs were used in the Community condition, but the labels were replaced with "Project A" and "Project B"

Following the quiz, participants made their Prisoner's Dilemma decision and then answered norm elicitation questions. Empirical expectations were assessed by having participants predict the percent of others choosing each option, and honest beliefs were incentivized via a small cash payment for estimates within 5% of the actual value. Normative expectations were assessed first by eliciting the participant's personal normative beliefs about whether there was a "right thing to do" in the task and then by having them predict the normative beliefs of other participants using the same incentive procedure as before. The order of the expectation questions was counter balanced and they were followed by a counterfactual question asking what the participant would have chosen if they knew the decision of the other player before making their choice. For someone sensitive to the norm of cooperation in this context, knowing that the other person has cooperated provides information on their underlying type (as a potential fellow norm-conformer) and induces role/rule following in response. After the eliciting social beliefs, participants answered several questions about their general concern for

acting appropriately and then completed a demographics questionnaire before submitting their responses.

Results

Social Framing

Figure 2 displays the proportion of participants choosing each option in the Prisoner's Dilemma for each social frame. While the current design resulted in the strongest social framing effect in pilot testing, we found no effect of frame on the rate of cooperation in the fully powered sampled, $\chi^2(1, N=258) = .19$, $p = .66$. Furthermore, empirical expectations did not differ significantly across conditions, $t(256) = -.76$, $p = .45$, nor did normative expectations for cooperation, $t(256) = -.76$, $p = .45$, defection, $t(256) = -1.61$, $p = .11$, and for no norm being in effect, $t(256) = 1.78$, $p = .08$. Likewise, personal normative beliefs did not significantly differ across context as well, $\chi^2(2, N=258) = 2.84$, $p = .24$. Average expectations and the distribution of personal beliefs are displayed in

Table 1.

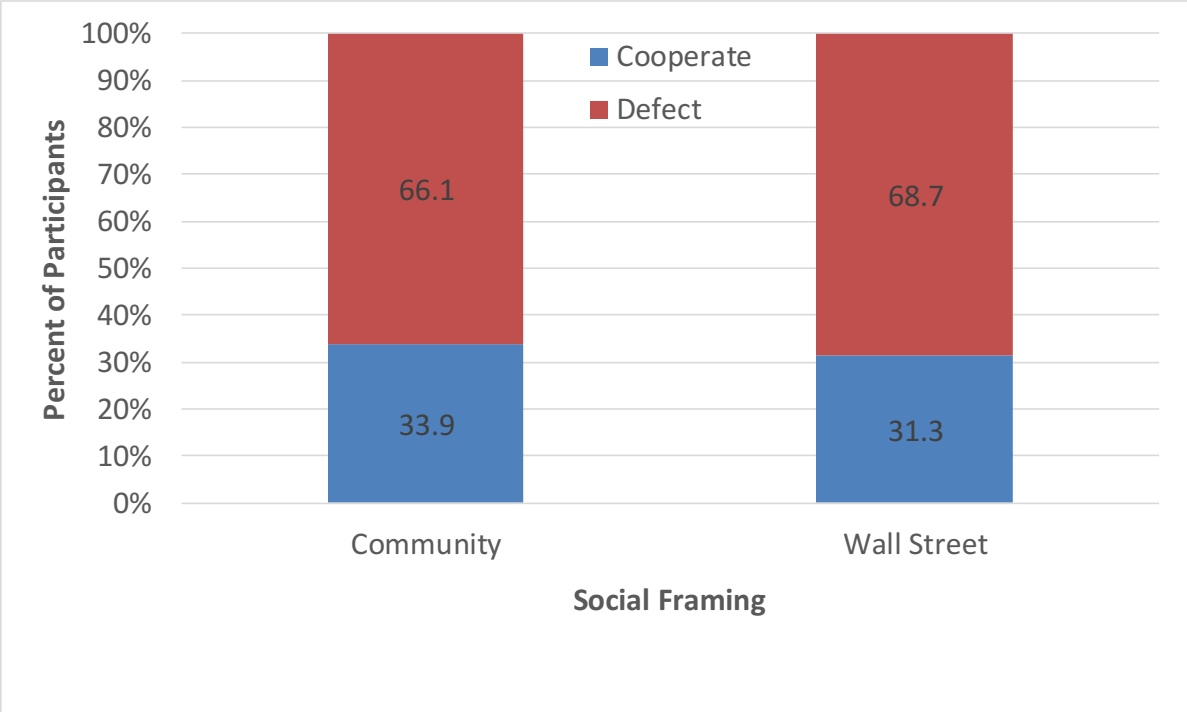

Figure 2. Proportion of participants choosing cooperation and defection in the Prisoner's Dilemma when framed as a "Community Task" (N=127) or a "Wall Street Task" (N=131).

Table 1.

Average expectations of descriptive and social norms alongside personal normative beliefs

	Empirical Expectations: Cooperation	**Normative Expectations: Cooperation**	**Normative Expectations: Defection**	**Normative Expectations: No Norm**	**Personal Normative Beliefs**
Community	43% (27.1)	45.7% (27.9)	44.8% (28.1)	28.7% (29.1)	C: 36% D: 36% NN: 28%
Wall Street	40.5% (26.7)	40.2% (26.89)	50.9% (27.2)	29.2% (27.1)	C: 27% D: 43% NN: 30%

Note. Average predictions (and standard deviations) about what others would choose in the task (empirical expectations) as well as what others thought was the "right thing to do" (normative expectations). The participant's own answer to the latter question is their personal normative belief about whether the right thing to do is to cooperate (C), defect (D), or whether they believe there is no normative (NN) response in the task.

Expectations and Personal Normative Beliefs of Cooperators and Defectors

Figure 2 examines empirical and normative expectations as a function of the participant's PD decision and reports pairwise comparisons. With the exception of beliefs about there being no norm active, cooperators thought others would cooperate more than defectors did, and they also thought that others would believe cooperating was the right thing to do while defection was not to a higher degree than defectors. Table 2 displays the frequency of personal normative beliefs as a function of PD decision and also reveals that people who actually cooperated were more likely to believe that that was the right action to take, whereas defectors viewed defection as more often the right choice, $\chi^2(2, N=258) = 104.7$, $p < .001$.

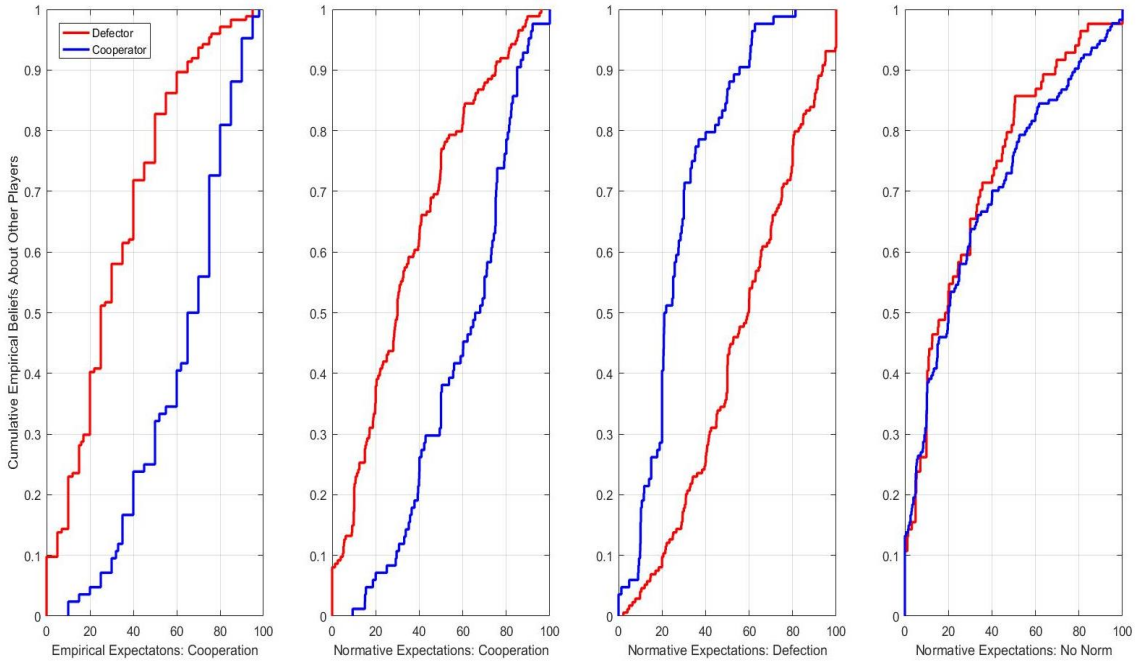

Figure 2. Cumulative distribution plots of beliefs about other players' behavior and beliefs as a function of the person's choice in the Prisoner's Dilemma. A Kolmogorov-Smirnov test revealed that cooperators and defectors had different beliefs about the behavior of others, D = .52, $p < .001$, as well as about others' beliefs regarding whether cooperation, D = .45, $p < .001$, or defection, D = .55, $p < .001$, was the right thing to do in the task. There was no difference in beliefs about whether there was no norm in the environment, D = .09, $p = .77$.

Table 2.

Personal normative beliefs of cooperators and defectors

	Personal Normative Belief		
	Cooperate	**Defect**	**No Norm**
Cooperators	71% (60)	4% (3)	25% (21)
Defectors	12% (21)	56% (98)	32% (55)

Note. Percentage (frequency) of cooperators/defectors who reported each type of normative belief.

Table 3 includes the zero-order linear correlations between the different expectations as well as the multiple correlations for regressions predicting each descriptive/normative belief as a function of the person's personal normative beliefs[8]. All expectations were significantly correlated except for beliefs about there being no norm in the task. Lastly, Table 4 includes coefficients from a binary logistic regression[9] predicting PD choice as a function of all belief measures (social and personal). The model was significant, $\chi^2(6) = 138.8$, $p < .001$, and explained 58% of the variance in PD choices (Nagelkerke R^2). Holding other beliefs constant, increases in the belief of a descriptive norm to cooperate was associated with increased cooperation, while beliefs that others' thought the normative choice was to defect were associated with more actual defection. Both the personal normative belief to defect and the belief in their being no norm were associated with increased defection compared to the personal belief that cooperating was the right action.

[8] Personal beliefs were dummy coded with the norm to cooperate as the reference group.
[9] All empirical and normative expectations were centered and personal normative beliefs were dummy coded with the norm to cooperate as the reference group.

Table 3.

Correlations among empirical and normative beliefs

	Empirical Expectations: Cooperation	**Normative Expectations: Cooperation**	**Normative Expectations: Defection**	**Normative Expectations: No Norm**	**Personal Normative Beliefs**
Empirical Expectations: Cooperation		.74**	-.69**	0	.61**
Normative Expectations: Cooperation	.74**		-.63**	.05	.6**
Normative Expectations: Defection	-.69**	-.63**		-.13*	.62**
Normative Expectations: No Norm	0	.05	-.13*		.56**
Personal Normative Beliefs	.61**	.6**	.62**	.56**	

Note. Pearson correlation coefficients between beliefs about the prevailing norm; multiple R reported for personal normative beliefs. * is significant at .05 level (two-tailed) and ** at .01

Table 4.

Binary logistic regression predicting cooperation in the PD

	Odds Ratio	95% CI [OR]	Wald	Significance
Empirical Expectations: Cooperation	1.03	[1.01, 1.05]	7.78	.005
Normative Expectations: Cooperation	.99	[.98, 1.01]	.31	.579
Normative Expectations: Defection	.98	[.96, 1.00]	3.81	.051
Normative Expectations: No Norm	.99	[.98, 1.01]	.60	.437
Personal Normative Belief: Defect	.03	[.01, .14]	23.46	< .001
Personal Normative Belief: No Norm	.35	[.14, .90]	4.78	.029

Note. Results from a binary logistic regression predicting cooperation in the PD as a function of (centered) beliefs about the salient norm and personal normative beliefs (dummy coded with cooperative beliefs as the reference group).

Discussion

Study 1 sought to test whether differences in social dilemma behavior due to social labels were caused by differences in beliefs about the prevailing norm in the environment as defined by Bicchieri (2006). Despite significant pilot testing, Study 1 did not replicate the social labeling effect. However, beliefs about norms were also no

different across conditions and were associated with actual behavior in the social dilemma. In particular, beliefs about the descriptive norm, normative norm of defection, and one's personal normative beliefs were associated with cooperation rates over and above their shared variance with other beliefs.

Study 2

Whereas Study 1 tested whether Bicchieri's (2006) conceptualization of norms could account for anticipated differences in behavior resulting from a social cue in the environment, Study 2 extends this test to a situation in which norms may change as a function of group membership. Specifically, another often-cited effect in the social dilemma literature is that the formation of groups, even minimal ones based on an arbitrary factor, can promote cooperation with in-group members in mixed-motive games (see, e.g., Dawes & Messick, 2000). However, findings by Charness, Rigotti & Rustichini (2007) challenge the effect of minimal-group manipulations in experimental games with adults while supporting the effect of group-membership if it has been made salient through public observation of decision making or a shared fate in payoffs. Study 2 sought to replicate the effect of group membership on cooperation in a social dilemma using salient self-selected groups at a large U.S. university – fraternity members. Of particular interest is whether differences in the descriptive and social norms for interacting with an in-group vs. an out-group member mediate the anticipated difference in cooperation across these conditions.

Method and Procedure

Participants (N=124, 100% male) were recruited from fraternities located at a large U.S. university for $5[10] and the opportunity to earn more based on their own and others' choices. The fraternities were enlisted via an e-mail solicitation and four houses responded with overlapping availability. Research teams attended the beginning of each fraternity's weekly meeting and conducted a simultaneous-move Prisoner's Dilemma experiment in tandem with all four houses. Instructions were read aloud and provided in writing (see Supplemental Materials) which outlined that the fraternity members would be playing a Prisoner's Dilemma (see Figure 3) with either someone from their own house or someone from a different house whose fraternity was listed. Following the instructions, an abstract quiz on the rules was administered and questions were answered in private. Next, participants made their official choice before completing norm elicitation questions similar to those in Study 1. However, empirical expectations were assessed separately for both one's own fraternity as well as the other fraternity[11] and were incentivized via a small financial payment for predictions within 10% of the actual value. Likewise, normative expectations and counterfactual questions were asked both for fraternity members in the participant's condition as well as those in the other fraternity who were given the same instructions.

---

[10] Some fraternities agreed ahead of time to contribute the $5 show-up fees to their fraternity's treasury, but all compensation from choices within the task went to the individual fraternity members in private envelopes after their meetings.

[11] This was included for exploratory purposes as the norms of a group one does not belong to or wish to belong to should, theoretically, have little influence on behavior. There were no differences between expectations of a person's own fraternity members and the expectations of members of a different fraternity (see Supplemental Materials).

		Someone from Alpha Chi Rho: Yellow	Someone from Alpha Chi Rho: Blue
You	Yellow	Other person: \$7 **You: \$7**	Other person: \$10 **You: \$0**
	Blue	Other person: \$0 **You: \$10**	Other person: \$3 **You: \$3**

Figure 3. An example of the monetary Prisoner's Dilemma used in Study 2. This example is from the out-group condition as a member of Pi Kappa Phi was the row player. In the in-group condition, the column label read, "Someone else from Pi Kappa Phi"

Results

Group Membership

Figure 4 displays the proportion of participants choosing each option in the Prisoner's Dilemma for in-group and out-group pairings. Even with self-selected groups, and in an context where participants were surrounded by only in-group members, we found no effect of grouping on the rate of cooperation, $\chi^2(1, N=124) = .04$, $p = .85$.

Regarding participants' beliefs about others, only normative expectations for cooperation among fellow fraternity members differed across conditions with fraternity members matched within their own house thinking that more of their other members would report that cooperation was the right thing to choose, $t(121) = 2.28$, $p = .02$. However, there was also a marginal effect of condition on empirical expectations in the hypothesized direction, expecting more cooperation within one's fraternity, $t(121) = 1.84$, $p = .07$. Neither normative expectations regarding defection, nor expectations regarding the lack of a norm were different across conditions. Yet, personal normative beliefs did marginally interact with condition such that participants matched with an

out-group member were more likely to report that there was no norm in the task, $\chi^2(2, N=123) = 5.81$, $p = .06$. Average expectations about one's own fraternity members (who were in the same grouping condition), and the distribution of personal beliefs are displayed in Table 5.

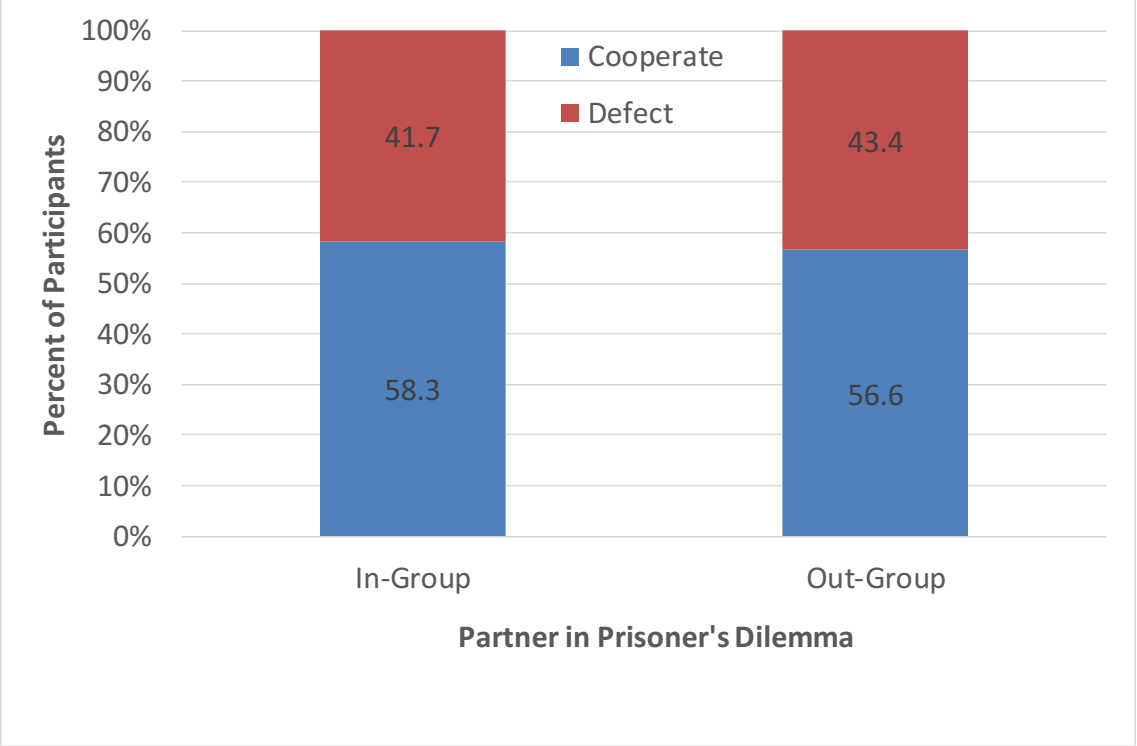

Figure 4. Proportion of participants choosing cooperation and defection in the Prisoner's Dilemma when matched with someone from their own fraternity, In-Group (N=48), or someone from another fraternity on campus, Out-Group (N=76).

Table 5.

Average expectations of descriptive and social norms alongside personal normative beliefs

Partner	Empirical Expectations: Cooperation	Normative Expectations: Cooperation	Normative Expectations: Defection	Normative Expectations: No Norm	Personal Normative Beliefs
In-Group	**59.6%** (24.8)	**55.6%** (28.6)	**28.5%** (26.2)	**17.3%** (20)	**C: 66%** **D: 19%** **NN: 15%**
Out-Group	**50.7%** (26.7)	**44%** (26.9)	**34.1%** (22.7)	**21.5%** (23.1)	**C: 54%** **D: 12%** **NN: 34%**

Note. Average predictions (and standard deviations) about what others would choose in the task (empirical expectations) as well as what others thought was the "right thing to do" (normative expectations). The participant's own answer to the latter question is their personal normative belief about whether the right thing to do is to cooperate (C), defect (D), or whether they believe there is no normative (NN) response in the task.

Expectations of Cooperators and Defectors

Figure 5 examines empirical and normative expectations as a function of the participant's PD decision and reports pairwise comparisons. As in Study 1, with the exception of beliefs about there being no norm active (which were marginally significant in this task), cooperators again thought others would cooperate more than defectors did, and they also thought that others would believe cooperating was the right thing to do while defection was not to a higher degree than defectors. Table 2 displays the frequency of personal normative beliefs as a function of PD decision and reveals, as in

Study 1, that people who actually cooperated were more likely to believe that that was the right action to take, $\chi^2(2, N=123^{12}) = 36$, $p < .001$

Table 7 includes the zero-order linear correlations between the different expectations as well as the multiple correlations for regressions predicting each descriptive/normative belief as a function of the person's personal normative beliefs[13]. All expectations were significantly correlated, even for beliefs about there being no norm in the task. Lastly, Table 8 includes coefficients from a binary logistic regression[14] predicting PD choice as a function of all belief measures (social and personal). The model was significant, $\chi^2(6) = 77.5$, $p < .001$, and explained 63% of the variance in PD choices (Nagelkerke R^2). Holding other beliefs constant, only increases in the belief of a descriptive norm to cooperate among fellow fraternity members was associated with increased cooperation.

[12] One participant did not complete the personal normative belief question
[13] Personal beliefs were dummy coded with the norm to cooperate as the reference group.
[14] All empirical and normative expectations were centered and personal normative beliefs were dummy coded with the norm to cooperate as the reference group.

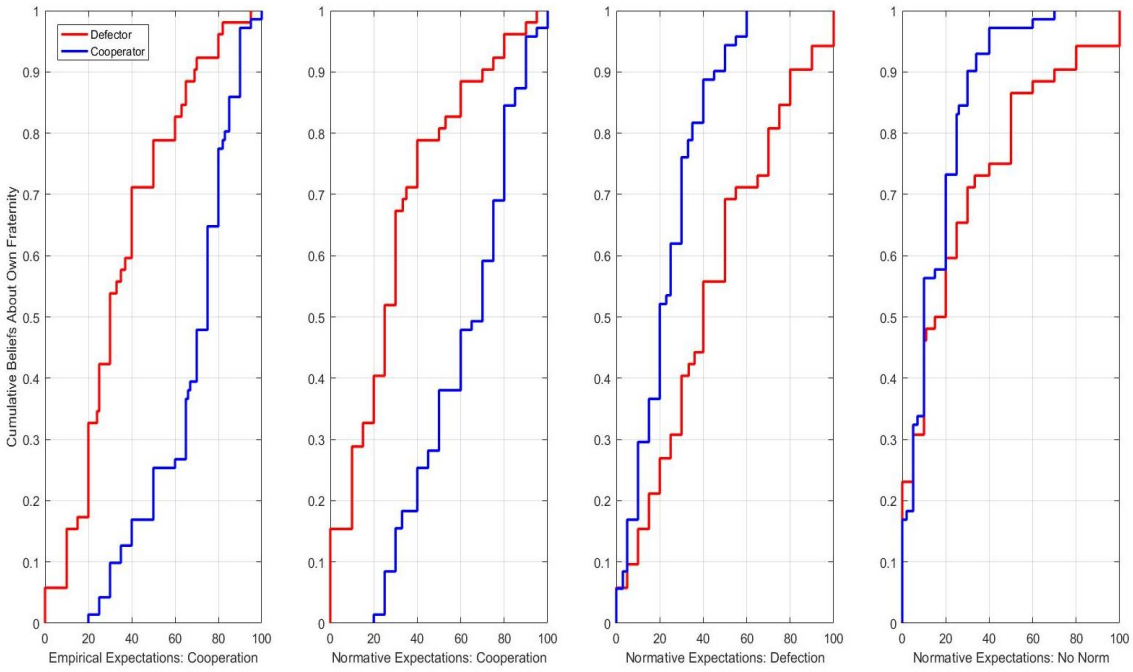

Figure 5. Cumulative distribution plots of beliefs about one's own fraternity members' behavior and beliefs as a function of the person's choice in the Prisoner's Dilemma (beliefs were only elicited about fellow fraternity members in the same grouping condition as the responder). A Kolmogorov-Smirnov test revealed that cooperators and defectors had different beliefs about the behavior of other members of their own fraternity, D = .58, $p < .001$, as well as about other fraternity members' beliefs regarding whether cooperation, D = .53, $p < .001$, or defection, D = .39, $p < .001$, was the right thing to do in the task. There was only a marginal difference in beliefs about whether there was no norm in the environment, D = .22, $p = .09$.

Table 6.

Personal normative beliefs of cooperators and defectors

	Personal Normative Belief		
	Cooperate	**Defect**	**No Norm**
Cooperators	77% (55)	0% (0)	23% (16)
Defectors	33% (17)	34% (18)	33% (17)

Note. Percentage (frequency) of cooperators/defectors who reported each type of normative belief.

Table 7.

Correlations among empirical and normative beliefs for one's own fraternity's members

	Empirical Expectations: Cooperation	**Normative Expectations: Cooperation**	**Normative Expectations: Defection**	**Normative Expectations: No Norm**	**Personal Normative Beliefs**
Empirical Expectations: Cooperation		.68**	-.58**	-.20*	.35**
Normative Expectations: Cooperation	.68**		-.66**	-.51**	.55**
Normative Expectations: Defection	-.58**	-.66**		-.26**	.40**
Normative Expectations: No Norm	-.20*	-.51**	-.26**		.56**
Personal Normative Beliefs	.35**	.55**	.40**	.56**	

Note. Pearson correlation coefficients between beliefs about the prevailing norm; multiple R reported for personal normative beliefs. * is significant at .05 level (two-tailed) and ** at .01

Table 8.

Binary logistic regression predicting cooperation in the PD

	Odds Ratio	95% CI [OR]	Wald	Significance
Empirical Expectations: Cooperation	1.05	[1.02, 1.07]	12.21	< .001
Normative Expectations: Cooperation	1.02	[.95, 1.08]	.23	.63
Normative Expectations: Defection	.99	[.94, 1.07]	.001	.97
Normative Expectations: No Norm	.99	[.93, 1.06]	.13	.72
Personal Normative Belief: Defect	.00	[.00, --]	.00	.99
Personal Normative Belief: No Norm	.58	[.15, 2.17]	.67	.41

Note. Results from a binary logistic regression predicting cooperation in the PD as a function of (centered) beliefs about the salient norm among members of one's own fraternity and personal normative beliefs (dummy coded with cooperative beliefs as the reference group).

Discussion

Study 2 sought to test whether differences in social dilemma behavior due to group membership were caused by differences in beliefs about the prevailing norm within one's reference group/fraternity as defined by Bicchieri (2006). Although we did

observe some differences in beliefs as a function of group membership, Study 2 did not replicate the grouping effect. One limitation of the current design was that it did not elicit fraternity members' opinion of the out-group fraternity their members were paired with. It is possible that fraternities did not see themselves in competition, or thought of themselves as part of a superordinate in-group of fraternity members/university students, which served to counteract the intended manipulation. However, beliefs about norms were again associated with actual behavior in the social dilemma. In particular, beliefs about the descriptive norm of one's other fraternity members in the same grouping condition were associated with cooperation rates over and above its shared variance with other beliefs.

General Discussion

A scientific understanding of norms and norm-based decision making is necessary for both theoretical advances and real-world applications. For example, developing effective policies to address actual social dilemmas and coordination problems will likely benefit from an understanding of the current institutional environment, which often entails understanding a community's current norms. To the extent that the existing informal rules legitimize the formal ones, the cost of enforcing the new formal rules will be lower and vice versa (Boettke, 2012). By operationalizing the concept of norms, Bicchieri (2006) provides a framework for testing the existence and impact of norms as well as a guideline for changing behavior through changing beliefs. Specifically, it is the belief that others are conforming to a norm and also expect

you to conform that motivates norm compliance (though sometimes rewards/sanctions or an internalization process is essential for conformity/enforcement).

Although unable to conceptually replicate two well-known findings, beliefs about what others would do (empirical expectations), what others thought was the right thing to do (normative expectations), and personal normative beliefs all correlated with the actual behavior of participants in social dilemmas (even though estimates of the expectations of others were incentivized via truthful elicitation procedures). In addition, while beliefs about descriptive and injunctive norms were highly correlated with each other, as well as with an individual's personal beliefs, a consistent finding was that empirical expectations were especially indicative of cooperative behavior. By Bicchieri's (2006) definition of the norms needed to induce conformity, this suggests that lab games designed as social dilemmas may actually be represented as coordination games in the minds of many subjects (which fits with Bicchieri's theory that social norms transform games into ones of coordination with fellow norm-conformers). However, future work would be beneficial in a couple directions.

The first direction might be improving the psychometrics of studying norms. For example, devising enough measures to conduct factor analyses with sufficient degrees of freedom to include multiple correlated latent factors - such as descriptive norm, social norm, and personal norm. With this in mind, establishing a task that reliably manipulates norms (unlike Studies 1 and 2) would provide a test bed for new quantitative measures that could go into a later analysis. For example, normative

expectations might be partially defined as/indicated by an agent's beliefs about what others' believe to be the normative expectations of the group.

Another direction would be to outline a theory of what norms to expect and, more crucially, when to expect them. For example, Alan Fiske's (1992) theory of four fundamental social relationships (communal sharing, authority ranking, equality matching, and market pricing) would go a long way to constraining the set of possible expectations an agent could have at any given time. However, this only gets one so far before the question becomes, what determines which social relationship is adopted during decision making at a specific point in time? This is where Gintis & Helbing's (2015) argument of a general social equilibrium model may be useful. Gintis & Helbing (2015) enrich the Walrasian general equilibrium model of economic theory to capture the distribution of social roles as well as their content. In equilibrium, the content (descriptive and social norms) associated with social roles is public knowledge and no actors have an incentive to change roles. However, out of equilibrium, expectations are represented as a statistical distribution over the content of roles which is assumed to be a subjective, yet networked, probability distribution. This idea suggests a method for predicting the salient norm for a given agent through measuring the joint beliefs of agents in their network weighted, potentially, by their social distance from the agent of interest (Gintis & Helbing, 2015).

References

Andrighetto, G., Brandts, J., Conte, R., Sabater-Mir, J., Solaz, H., & Villatoro, D. (2013). Punish and voice: Punishment enhances cooperation when combined with norm-signalling. *PLoS ONE,8*(6). doi:10.1371/journal.pone.0064941

Andrighetto, G., Grieco, D., & Tummolini, L. (2015). Perceived legitimacy of normative expectations motivates compliance with social norms when nobody is watching. *Frontiers in Psychology,6*. doi:10.3389/fpsyg.2015.01413

Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*,*144*, 1-35. doi: 10.1016/j.jet.2008.01.004

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

Bicchieri, C., & Chavez, A. K. (2013). Norm manipulations, norm evasion: Experimental evidence. *Economics and Philosophy*, *29*, 175-198. doi: 10.1017/S0266267113000187

Bicchieri, C., & Xiao, E. (2009). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making*, *22*, 191-208. doi: 10.1002/bdm.621

Bicchieri, C., Xiao, E., & Muldoon, R. (2011). Trustworthiness is a social norm, but trusting is not. *Politics Philosophy Economics*, *10*(2), 170-187. doi: 10.1177/1470594X10387260

Binmore, K. (2010). Social norms or social preferences? *Mind & Society*, *9*, 139-157. doi:10.1007/s11299-010-0073-2

Boettke, P. J. (2012). *Living Economics Yesterday, Today, and Tomorrow.*. Chicago: Independent Institute.

Chandler, J., Mueller, P., & Paolacci, G. (2013). Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods,46*(1), 112-130. doi:10.3758/s13428-013-0365-7

Charness, G., Rigotti, L., & Rustichini, A. (2007). Individual behavior and group membership. *American Economic Review*, *97*(4), 1340-1352. doi: 10.1257/aer.97.4.1340

Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology*, *39*, 268-277. doi: 10.1016/j.joep.2013.09.004

Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences,15*(5), 218-226. doi:10.1016/j.tics.2011.03.003

Cialdini, R. B., Raymond, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*, 1015-1026.

Dana, J., Cain, D., & Dawes, R. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational behavior and human decision processes*, *100*, 193-201. doi: 10.1016/j.obhdp.2005.10.001

Dana, J., Weber, R., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*, 67-80. doi: 10.1007/s00199-006-0153-z

Dawes, R. M., & Messick, D. M. (2000). Social dilemmas. *International Journal of Psychology*, *35*(2), 111-116.

Dufwenberg, M., Gachter, S., & Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, *73*, 459-478. doi: 10.1016/j.geb.2011.02.003

Ellingsen, T., Johannesson, M., Mollerstrom, J., & Munkhammar, S. (2012). Social framing effects: Preferences of beliefs. *Games and Economic Behavior*, *76*, 117-130. doi: 10.1016/j.geb.2012.05.007

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

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review,99*(4), 689-723. doi:10.1037//0033-295x.99.4.689

Gintis, H., &Helbing, D. (2015). Homo socialis: An analytical core for sociological theory. *Review of Behavioral Economics,2*(1-2), 1-59. doi:10.1561/105.00000016

Hauge, K. E. (2016). Generosity and guilt: The role of beliefs and moral standards of others. *Journal of Economic Psychology,54*, 35-43. doi:10.1016/j.joep.2016.03.001

Horner, V. & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). Animal Cognition, 8, 164-181. doi: 10.1007/s10071-004-0239-6.

Kovács, A.M., Téglás, E. & Endress, A.D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. Science, 330, 1830-1834. doi: 10.1126/science.1190792

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, *1*, 60-79.

Liberman, V., Samuels, S. M., & Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and Social Psychology Bulletin*, *30*(9), 1175-1185. doi: 10.1177/0146167204264004

Kimbrough, E. O., & Vostroknutov, A. (2015). Norms make preferences social. *Journal of the European Economic Association,14*(3), 608-638. doi:10.1111/jeea.12152

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In T. Brown, E. S. Reed & E. Turiel (Eds.), *Values and Knowledge* (pp. 103-135). Mahwah, NJ: Lawrence Erlbaum Associates.

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., . . . Simunovic, D. (2013). Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organizational Behavior and Human Decision Processes,120*(2), 260-271. doi:10.1016/j.obhdp.2012.06.002

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *The Journal of Neuroscience*, *33*(3), 1099-1108. doi: 10.1523/JNEUROSCI.1642-12.2013