Choice in shared resource problems: a social motives selection model

Joshua C. Skewes[1,2*]

1 – Department of Linguistics, Cognitive Science, and Semiotics, Aarhus University, Denmark

2 – Interacting Minds Centre, Aarhus University, Denmark

Abstract

Experimental research on choice in shared resource problems focuses on contextual and individual variation. Experimenters are interested in how factors including incentive structures, social dynamics, and cultural differences influence choices. By contrast, until relatively recently, theoretical research has focused on the development of formal models that are robust to these sources of variation. The aim of this paper is to provide a model of social decision-making that can be used to better integrate experiment and theory. This model is based on the idea that individual and contextual variation in the way people invest in shared resources is the product of variation in their social motivation. The paper builds on two well-developed theories of social choice behavior – the Experience Weighted Attractions model based on selfish reinforcement learning, and the Conditional Cooperation model based on social beliefs and preferences – with a view to understanding choice in terms of selection between the social motives represented by the two models. I show that this approach can accurately predict the social motives used to generate choice in artificial simulations, and that the model can outperform the EWA and CC models alone. I then apply the model to three prominent experimental cases. I show that increasing the return on public good investment decreases cooperative social motives; that allowing peer punishment in a public goods game increases cooperative motives; and that culture mediates the effects of punishment on social motive selection. The paper concludes with discussion of some of the theoretical and experimental avenues opened up by the approach offered.

Choice in shared resource problems: a social motives selection model

1. Introduction

Shared resource problems occur in almost all areas of human social life. Examples include the provision of funding for workers' unions; the management of scarce but renewable resources; and the provision of freely available infrastructure such as roads, hospitals, and libraries. Knowledge about how people make decisions when solving shared resource problems is of great importance. Not only is this knowledge useful for designing institutions that can promote cooperation in the use of shared resources, it also advances our conceptual understanding of social decision-making more generally (Smith, 2008).

Since the early 1980s, there have been two main lines of psychological research devoted to understanding decision-making in shared resource problems. Both lines of research are essential for theory development, and for designing institutions for regulating social decision-making around cooperative enterprises. Problematically, these lines of research have had divergent interests and agendas. The first line is primarily experimental. Its goals include conceptually rich psychological explanations that are informed by broader empirical analyses of social life and cognition. This tradition places emphasis on individual differences, and on the effects of situational factors on social decision-making (for a narrative review see Kopelman, Weber, and Messick, 2002; for a meta-analysis see Zelmer, 2003). The second line is primarily formal. Its goals include self-contained models that are robust to individual and situational variation (e.g. Camerer & Ho, 1999; Masel, 2007; Roth & Erev, 1995). This tradition of research places an emphasis on the generalizability of theoretical formalisms, and on the accuracy of model predictions.

This paper presents theoretical work that aims for a greater unification of these agendas. The focus is on experimental public goods games. Section 2 addresses the experimental and theoretical state of the art. We see how recent experimental work has focused on individual and contextual variation, and I review formal theoretical research designed to provide explanations that are robust to this variation. Section 3 seeks to overcome this divergence by treating alternative theories as formalizations of the alternative social motives that may be elicited when solving shared resource problems under different kinds of circumstances. I use latent mixture modeling to put forward a model of social motives selection. Section 4 presents a formal evaluation of this model, and the model is compared to related theoretical work. Section 5 aims to demonstrate the inferential power of this model by applying it to three prominent cases in the experimental literature. The focus here is on the effects of incentive structures, social dynamics, and cultural variation. Here I attempt to show how the model

may be used to answer substantive research questions about the role of social motives in these cases. Section 6 discusses the main theoretical and experimental avenues opened up by the approach developed in the paper.

2. Background

2.1 Experimental public goods games

In the standard public goods game, an experimenter endows a group of participants with some amount of tokens. Each person is asked to decide how many tokens they will set aside or "save", and how many they will contribute to a shared pool as an "investment" in a public good. The experimenter then multiplies by some fixed factor the number of tokens invested by the group; splits the pool evenly into shares; and redistributes the shares evenly to the participants. This procedure may then be repeated for a number of rounds, and at the end of the game, the tokens are cashed out and given to the participants.

This game models an interesting shared resource problem. The best strategy available to the group as a whole (i.e. the Pareto optimum) is in conflict with the best strategy available to each individual (i.e. the Nash equilibrium). The group's payout is maximized if each individual always invests their full endowment. If, for example, participants are endowed 10 tokens and the multiplication factor for the shared pool is 1.5, and if they all invest fully, each participant will earn the group maximum return of 5 extra tokens per person. However, assuming their partners all invest, each individual's payoff is maximized if they save all their tokens. If, for example, there are four people in each group, and one person saves when all the others invest, this "free rider" will earn 7.5 extra tokens – a 50% gain over the optimal group strategy. Importantly, if all the participants follow the best individual strategy, and save all their tokens, then no tokens will be pooled and multiplied, and nobody will make more than their initial endowment; the worst possible outcome for both the group and the individual. This makes the game an informative laboratory paradigm for studying cooperation, free riding, and other social dynamics involved in sharing common resources (Kopelman et al., 2002; Smith, 2008; Zelmer, 2003).

Decisions in experimental public goods games are usually analyzed by aggregating all participants' responses within a study – by combining all their contributions into an average or "typical" pattern of investments. Empirically, this typical pattern reflects neither the Nash equilibrium, nor the Pareto optimum, but something in between. Under normal circumstances, the "average" participant starts by splitting their endowment – or by investing about half of the tokens they are given – before their contribution declines over time until they are investing close to nothing by the end of the game (Ledyard, 1995).

A great deal of effort has been invested in explaining this typical pattern, and the next section discuss some of this work. Despite general theoretical interest, the well-known average result conceals substantial variability at the individual level (Charness & Rabin, 2002; Hichri, 2005; Janssen & Ahn, 2006; Kurzban & Houser, 2001). Some participants never contribute, but free ride the whole game; some contribute all of their tokens at the beginning, but reduce their contributions over time; some maintain a high level of contribution throughout the game; some make high contributions only in occasional trials; and there is reason to believe that most experiments show some combination of these individual patterns (Kurzban & Houser, 2001). It remains a challenge for theorists to formally explain this individual variability (Janssen & Ahn, 2006; Masel, 2007).

Experiments show that aggregate and individual patterns of contributions may also depend on the situation. A trivial example is the effect of the multiplication factor for the public good. Researchers have shown that if the multiplication factor for the public good is high enough, then investments will remain uniformly high throughout the game (Cartwright & Lovett, 2014; Zelmer, 2003). In fact, this should always happen whenever the public good is multiplied so that the marginal per capita return – or the amount a participant will receive in return for their investment even if nobody else contributes – is higher than one, because people are guaranteed some gain regardless of other people's decisions. This result is explored in the first case below.

A less trivial example is the effect of social dynamics on investments. Perhaps the best-known case of social dynamics influencing public goods contributions is the effect of the possibility of peer punishment in the game. If it is possible for individuals to choose to pay some of their own tokens to have others in the group lose tokens after a round, then the aggregate level of contributions remains high (e.g. Egas & Riedl, 2008; Fowler, 2005; Fehr & Gächter, 2002). The standard explanation of this is that participants may use this ability to "punish" individuals for giving lower than expected contributions, leading to greater (enforced) contribution. This result is explored in the second case below.

Many other factors have been shown to influence investment in the public good. These include group size (Isaac & Walker, 1988a), how the game is framed by the experimenter (Andreoni, 1995; Cartwright & Ramalingam, 2019; Fosgaard & Piovesan, 2015), the possibility of communication before and during the experiment (Dawes, van de Kragt, & Orbell, 1990; Isaac & Walker, 1988b; Nagatsu et al., 2018), prior expertise in professional public resource allocation (Butler & Kousser, 2015), expectations about the kinds of partners encountered in the game (Burton-Chellew, Nax, & West, 2015; Nagatsu et al., 2018), cultural background (Gerkey, 2013; Herrmann, Thöni, & Gächter, 2008; Nishi et al., 2017, see Balliet & van Lange, 2013 for meta-analysis), and other individual differences (Kopelman et al., 2002; Zelmer, 2001). I explore the effects of culture in the third

case below, and I discuss how the model might be applied to understand the effects of communication in the discussion of this paper.

2.2 Main theories of decision-making in solving shared resource problems

Early theory in social psychology was focused on explaining social decision-making in terms of individual differences. Early theorists interpreted each individual's patterns of investments to reflect facts about that individual that may be generalized to other social contexts. For example, Kelley and Stahelski (1970) argued that people's investment choices depend on a dimensional personality trait, which also underlies other aspects of their social behavior. At the one end of this dimension, the authors identified competitive personality types. These are people who rarely contribute to shared resources; tend to be high on related traits like authoritarianism; and perceive the actions of others as expressions of more competitive intentions. At the other end of this dimension, the authors identified cooperative personality types. These are people who readily contribute to shared resources; tend to be low on authoritarianism; and perceive the actions of others as expressions of both competitive and cooperative intentions.

Other scholars have formalized and expanded on this trait-based conception. For example, McClintock (1972) explained different patterns of investment as emerging from individual differences in how people experience reward in the game. He postulated that people's investment choices reflect one of four underlying social motives, which in turn reflect different utility preferences in social decision-making. People may be selfish, in which case they prefer choices that maximize their own outcome; they may be competitive, in which case they prefer choices which maximize the difference between their partners' and their own outcomes; they may be cooperative, in which case they prefer choices which maximize the joint outcome; or they may be altruistic, in which case they prefer choices which maximize the outcome for their partners.

These perspectives have been influential in experimental work, and subsequent research has attempted to build on and independently verify both trait and motivational theories. For example, building on the trait perspective, Bereczkei, Birkas, & Kerekes (2010) have argued for the existence of Machiavellian personality types. These are selfish or competitive individuals who use cooperative tactics to maximize their own outcomes (Czibor & Bereczkei, 2012), and who report lower affect when playing the game (Czibor, Vincze, & Bereczkei, 2014; see also Skatova & Ferguson, 2011; Utz, Muscanell, & Göritz, 2014; and Volk, Thöni, & Ruigrok, 2011; for more research on traits). Building on the motivations perspective, Chung, Yung, & Seong (2015) analyzed EEG signals using multi-feature pattern analysis, and showed that oscillations over the temporal-parietal cortex could be used to predict cooperation in a public goods game, which they interpreted as neural signals underlying

cooperative motivations (see also Akerlof & Kranton, 2005; Blanco, Engelmann, & Normann, 2011; Krawczyk & Le Lec, 2015; for more research on motivations).

Other research has attempted to sharpen the set of social motives used to explain decision-making in solving shared resource problems. Research by Burton-Chellew et al. casts doubt on altruism as an effective motive. Burton-Chellew and West (2013) showed experimentally that when people are told that their cooperative behavior benefits others, they choose to save more of their tokens for themselves rather than increasing their investments as might be expected. Following up on this result, Burton-Chellew et al. (2015) showed that payoff sensitivity, and not altruistic utility motives, was the best predictor of investment amounts. This research raises the possibility that not four, but three main social motives underlie decisions in shared resource problems.

Trait and motivational approaches were less influential in the formal work that immediately followed. Diverging from ongoing trends in experimental research, most of the prominent theoretical models developed from the 1980s to the early 2000s assumed that the psychological processes involved in social decision-making are geared toward maximizing one's own returns (see however Messick & Sentis, 1985). Seen from the broader motivational perspective, the dominant theoretical program during this time assumed that investment decisions are primarily the result of one particular type of social motive: a selfish motive.

Earlier research within this program focused on how an individuals' beliefs about other people's investment intentions should influence their own utility maximizing choices. For example, Kadane and Larkey (1982) argued that resources from Bayesian decision theory should be used to understand interactive choices, and that these resources should be supplemented by experimental research from cognitive psychology concerning how people form prior subjective beliefs about other people's behavior. Their argument, and the Bayesian cognitive approach it engendered, remains influential (e.g. Masel, 2007; Larrouy & Lecouteaux, 2017).

Later work drew on resources from learning theory. For example, Roth and Erev (Roth & Erev, 1995; Erev & Roth, 1998) assumed that people have latent propensities to choose each possible token amount in the game (i.e. 1, 2, 3, 4, 5, etc.), and that these propensities are updated in accordance with a learning rule, so that the history of payoffs for each token amount influences the propensity to invest that token amount again. For example, if an individual contributes all of their tokens, but the rest of the group free rides, then the marginal payoff they will receive will be low. For the investor, this will cause a reduction in the propensity for full investment relative to other amounts, and so they will be more likely to contribute less in the next round. For the free riders, the effect will be the opposite. They will receive higher payoffs, and so their propensity to free ride will be reinforced. Roth and Erev showed that because payoffs are dependent in this way on how much others invest, what appears

to be coordinated choice behavior may emerge in the group across rounds purely as a result of individual reinforcement learning.

Camerer and Ho (1999) developed a highly influential synthesis of belief-based and learning-based models, called the Experience Weighted Attraction or EWA model. They followed Roth and Erev in assuming that individuals have propensities – termed attractions in EWA – for each token amount, and that these are updated according to a learning rule. Expanding on Roth and Erev, they assumed that this rule also applies counterfactually to token amounts that could have been chosen, but were not. In this way, their model includes a process akin to belief-based regret, in the sense that the updating of attractions is partly influenced by payoffs received after a round, and partly by forgone payoffs that the individual believes they could have gained if they had invested differently. This model is described in full in section 3.

Since about 2000, there has been a resurgence in attempts to formally model social motivations other than selfish ones. One line of research has been focused almost exclusively on cooperative motives. This research has been broadly inspired by a game-theoretic solution to shared resource problems known as tit-for-tat (Axelrod, 1984; Lichbach, 1992). Players who use a pure tit-for-tat strategy seek always to match other's contributions in the game. Tit-for-tat based approaches thus differ fundamentally from reinforcement learning models. They ignore reward and utility. Thus they may be used to define social motives in terms of how people respond directly to others choices (e.g. Masel, 2007; Segal & Sobel, 2006).

A prominent examples of this kind of approach is the experimental schema for understanding conditional cooperation developed by Fischbacher and Gächter (2010). Similar to learning theories, the authors assumed that people have relative propensities for different token amounts. However, they did not assume that these are updated after feedback. Instead, they modeled propensities as a fixed set of preferences for contributing a given amount, dependent on the contributions of others. For example, a person might have wholly altruistic preferences, and always prefer to invest the maximum amount in the public good, regardless of what they believe others would do[1]. More frequently, people's decisions follow some form of conditional cooperation, leading them to prefer to invest the same – or somewhat less – than what they believe others will contribute. Fischbacher and Gächter therefore explain the stereotypical decline in investment over time not as a result of changing preferences due to learning, but as a result of the calibration of beliefs about what others are going to invest in a given trial. Building on this schema, I will specify what I will call the conditional cooperation or CC model in section 3.

[1] In line with the research of Burton-Chellew et al. (Burton-Chellew and West, 2013; Burton-Chellew, Nax, and West, 2015), such altruists were found to be empirically rare.

A separate line of research follows the tradition established by McClintock (1972) of defining social motives in terms of individuals' reward preferences. For example, Charness and Rabin (2002) proposed a three parameter utility function for joint outcomes in two-person games. The model includes two weighting parameters – one for the other player's outcome when the other player is earning less than oneself, and one for the other player's outcome when the other player is earning more than oneself. The third parameter is a shift parameter that changes the relative weighting of the other player's outcomes in response to instances of non-cooperation. This model is designed to capture three distinct social motives within the same utility function. The first is competitive motivation, which is defined using a parameter combination that yields higher utilities for greater differences between one's own and the other's outcomes. The second is social welfare motivation, which is defined using a parameter combination that yields higher utilities for greater joint outcomes (see also Andreoni & Miller, 2002). The third is difference aversion, which is defined using a parameter combination that yields higher utilities for smaller differences between one's own and the other's outcomes (see also Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000). Building on the learning theoretic approach, Janssen and Ahn (2006) incorporated this utility function into a version of the EWA model, thereby expanding the range of social motives that the theory can be used to account for. In this way, Janssen and Ahn's elaborated EWA model provides a reinforcement learning based alternative to models focused on conditional cooperation. This model is described in full in section 3.

In sum, many of the main experimental lines of research on decision-making in shared resource problems are focused on questions relating to individual differences and the influence of situational factors. Early theory was focused on understanding individual and situational variation in terms of traits or social motives. Subsequent theoretical work digressed from this trend, with theories increasingly developed as general models which assumed that individuals sought only to maximize their own utility; an ostensibly selfish social motive. More recently, other social motives have been incorporated into formal theory. Two main approaches have been prominent. The first ignores utility, and instead defines social motives more directly in terms of beliefs and preferences regarding the choices of others. This approach has yielded theories focused on cooperation. The second retains a focus on rewards and losses, and uses an expanded notion of utility to define a wider range of social motives.

2.3 The present project

This paper aims to build on and formally compare these two main approaches. Building on the theories described above, the paper presents a new model of individual differences in social decision-making in public goods games

in terms of the selection between competing social motives. In alignment with contemporary experimental work, these motives are not treated as fixed psychological constructs, but rather as flexible strategies that may be employed by different individuals in different circumstances for different reasons. I label this the Social Motive Selection or SMS model.

The current presentation of the model focuses explicitly on selfish and cooperative social motives. I employ Camerer and Ho's (1999) EWA model as an implementation of the selfish strategy (focused as it is on the optimization of individual utility), and I specify a novel formalization of Fishbacher and Gächter's (2010) CC schema, which I provide as an implementation of the more social cooperative strategy (focused as it is on matching investments to others in the group). Altruistic, competitive, and other motives may be easily incorporated into the model, but are left for future research (see section 6 for discussion).

To model individual differences in the application of selfish and cooperative motives, I embed the EWA and CC models within a Bayesian latent mixture model. I assume that the posterior for the mixture parameter represents the probability that a given individual has chosen to play either selfishly – i.e. that their choices are most likely generated by the EWA model; or cooperatively – i.e. that their choices are most likely generated by the CC model. I thus interpret the mixture parameter as representing the outcome of a social motive selection process in the mind of the individual decision-maker, where the social motives themselves are formalized in the component models. For examples of this approach in other domains in cognitive science, see Bartlema et al. (2014), Miettunen et al., (2016), and Kary et al. (2017).

I specify this model in detail in section 3. In section 4, I assess the model's ability to recover known social motives from simulated data. Also in section 4, I formally compare the Social Motive Selection model to the EWA and the CC models of which it is composed, and to Janssen and Ahn's (2006) approach of explaining social motives using the EWA model alone. I then seek to demonstrate the explanatory usefulness of the SMS model, by applying it to a set of empirical cases in section 6.

3. Model specification

3.1 Modeling selfish outcome maximization – the EWA model

The EWA model (Camerer & Ho, 1999) models public goods investments in terms of reinforcement learning. The model works by treating each token amount (e.g. 1, 2, 3, 4, etc.) as an independent "strategy" in a round, and by modeling learning in terms of updating the strength of relative preferences for each strategy. The authors term this relative strength of preference an "attraction", and a person's set of attractions can reasonably be

thought of as a set of utilities for each possible contribution amount. Public goods investments are then modeled as choices over attractions. The model thus consists of two main components: a learning rule for updating attractions to token amounts across rounds, and a choice rule for deciding how much to invest given one's attractions in any one round.

The learning rule is complex, and has three main components: a memory component, a reward processing and regret component, and an experience weighting component. These are included in a single learning rule, which is as follows.

$$A_{j,t} = \frac{\phi . N_{t-1} . A_{j,t-1} + (\delta + \Big((1-\delta).(c_{n,t} == j)\Big).(\frac{\pi(j + \sum c_{-n,t-1})}{nagents} - j)}{N_t} \quad (1)$$

The memory process is implemented in the first part of the model, $\phi . N_{t-1} . A_{j,t-1}$. Here new attractions $A_{j,t}$ for each token amount *j* in a trial *t* are updated from old attractions $A_{j,t-1}$, weighted by two parameters: an experience factor (N_{t-1}) (elaborated on below), and a free parameter ϕ, which models memory retention. When ϕ is higher, older attractions have more influence on decisions, and attractions are updated more slowly. ϕ is thus a rate parameter and so is bounded at (0,1).

For reward processing, the model assumes that attractions are updated on the basis of payoffs, and that all attractions – including attractions for both chosen and unchosen token amounts – are updated after a round. Attractions for *chosen* token amounts are assumed to be updated according to the actual payoff or outcome, and attractions for *unchosen* amounts are assumed to be updated according to imagined forgone outcomes. Regret is then represented as some weighted proportion of what the outcome would have been, had a given token amount been chosen.

This is implemented as follows. The expression $(\frac{\pi(j+\sum c_{-n,t-1})}{nagents} - j)$ is the calculation for the payoff from the public good – actual or imagined – in a trial. Here, *j* indicates a given token amount (i.e. 1, 2, 3, 4, etc.), $\sum c_{-n,t-1}$ indicates the sum of the actual contributions of the other players in the game (with the subscript –*n* representing the rest of the group), π is the multiplication factor, and *nagents* is the total number of players among whom the public good is divided.

Reward and regret are implemented in the expression $(\delta + \Big((1-\delta).(c_{n,t} == j)\Big)$, where δ is a free parameter representing the weighting of forgone payoffs for unchosen amounts, $(c_{n,t} == j)$ represents the amount which

was actually contributed by group member *n* in a trial *t*, and multiplication by $(1-\delta)$ ensures that weighting for the payoff received from the actual contribution is always one. Forgone outcomes are thus weighted as some proportion δ of the actual outcome, and so δ is bounded at (0,1).

For experience weighting, the model assumes that the effect of experience is to lock in preferences or attractions over time, such that new feedback and outcomes influence the updating of attractions less and less as the game goes on. This is implemented as a weighted trial counter N_t, which modulates the effects of learning across trials, such that as N_t increases, attractions update more slowly. In the model, this is implemented by including N_t as the denominator in the learning rule. This effect is governed by a free parameter ρ, which models the retention of previous experience, such that

$$N_t = \rho . N_{t-1} + 1 \tag{2}$$

The probability $p_{j,t}$ of choosing a token amount *j* on trial *t* is modeled using a decision rule over attractions[2]

$$p_{j,t} = \frac{e^{\theta A_j}}{\sum_{k=1} e^{\theta A_j^k}} \tag{3}$$

The final contribution, notated here as c_t^{EWA} to indicate choice from the EWA model, is then modeled as a sample from a categorical distribution with probabilities $p_{j,t}$.

$$c_t^{EWA} \sim Categorical(p_{j,t}) \tag{4}$$

3.2 Modeling cooperation – the CC Model

The CC model[3] models public goods investments in terms of degree of cooperative behavior. As such, and in contrast to the EWA model, the CC model has no learning process and no decision process. It also has no explicit representation of the multiplication factor or marginal per capita return for the game. Instead, the CC model has components incorporating beliefs about what the rest of the group will invest in the next trial;

[2] The authors do not see the choice rule in equation 3 as integral to the model. They emphasize that the rule that they use (softmax) is ancillary, and that the main theoretical content is the learning rule. We follow the paper in applying the same function for modeling choice.

[3] Strictly speaking, the model we present here was not developed as a cognitive model with free parameters representing psychological processes about which we can do inference. Instead, Fischbacher & Gächter (2010) used preliminary experiments to measure a number of individual parameter values directly. Thus, there is some degree of extrapolation from their original paper to the formalization we provide. However, our formalization is so directly dependent on the original work that we present it here as a direct citation.

preferences for one's own contribution, given what one believes the rest of the group will do; and a dynamic balance between the two. The model is in this sense blind to payoffs, and only concerned with the investments of others.

The model assumes that each individual has a set of token amounts that they would prefer to contribute in the game, depending on what they believe the average group contribution will be. For instance, one person might prefer to always match what they believe the group will contribute; while another person might prefer to always contribute slightly less than the group. The model also allows for people to have extreme preferences that are not conditional on the group contribution. For instance, a free rider might prefer to always contribute nothing, regardless of the group's contribution; while a perfect altruist might prefer to contribute all of their tokens in each round. Preferences are therefore assumed to follow the linear model[4]

$$P^{vec} = B_0 + B_p.token\ values \tag{5}$$

where P^{vec} is the vector of individual preferences for each possible contribution by the group, and the free parameters B_0 and B_p are the intercept and slope of the linear association between possible group contributions, and the individual's own preferred contribution.

This model captures the examples as follows. A person who prefers to consistently match others will have a B_0 of 0 and a B_p of 1. A person who prefers to undermatch will have a B_0 of 0 and a B_p of <1. A free rider will have a B_0 of 0 and a B_p of 0. And a perfect altruist will have a B_0 of the maximum token amount and a B_p of 0.

From this relationship, an individual's preferred contribution in a given trial P_t can be modeled as

$$P_t = P^{vec}_{Gb_t} \tag{6}$$

where Gb_t is an index that represents the individual's belief about what the average contribution will be on a trial. Beliefs themselves are formed by experience, and are updated in each trial using a learning rule

$$Gb_t = \gamma.(Gb_{t-1}) + (1-\gamma).(Ga_{t-1}) \tag{7}$$

[4] Fischbacher and Gächter (2010) also identified participants for whom this relationship did not adhere to a linear form. These so-called "triangle cooperators" had preferences which tracked beliefs up to a fixed point, before declining as other people's investments increased. Triangle cooperators are discussed in more detail in section 5.

where Gb_t is modeled as a weighted average of prior beliefs Gb_{t-1} and the observed average contribution Ga_{t-1}, and where the free parameter γ is the learning rate for updating beliefs about the group. γ is thus a rate parameter and so is bounded at (0,1).

With preferences and beliefs thus defined, the model assumes that an individual's contribution in a trial, notated c_t^{CC} to indicate choice from the CC model, is a weighted average of the two

$$c_t^{cc} = \omega_t.(Gb_t) + (1 - \omega_t).(P_t) \quad (8)$$

Fischbaher & Gächter (2010) also observed empirically that the balance between beliefs and preferences changes, with preferences dominating over time. This can be formalized by assuming that the weighting parameter ω_t undergoes decay across trials according to

$$\omega_t = \omega_{t-1}.\ (1 - \lambda) \quad (9)$$

where the free parameter λ represents the decay rate. Conceptually, this can be thought of as the rate at which individual preferences dominate over a starting strategy of conditional cooperation. ω_t and λ are both rate parameters and are bounded at (0,1).

3.3 Modeling social motive selection using latent mixtures – the SMS model

From here we can use Bayesian latent mixture modeling to infer whether individuals choose their investments according to a selfish or a cooperative motive. Latent mixture models assume that the data being analyzed may have been generated by one of multiple sub-populations represented in the sample, and that the sub-population to which each individual belongs is not known in advance, but is inferred from data (Bartlema et al., 2014). In this way, latent mixture models can be used to categorize individuals in terms of separable, unobservable variables or processes. Starting from an assumption that each individual's investment choices are generated by one of two latent processes – either the reinforcement learning process formalized in the EWA model, or the cooperative process formalized in the CC model – we may infer from data which of the two models is most likely to be generating choices for that individual. In this way, latent mixture modeling provides a method for (probabilistically) categorizing each individual in a sample as either a selfish utility maximizer, or a conditional cooperator. For similar uses of latent mixture models in cognitive science, see Bartlema et al. (2014), Mietunen et al., (2015), and Kary et al. (2017).

To do this, we start by modeling contributions in a trial c_t as a sample from a normal distribution

$$c_t \sim Normal(\nu c_t, \sigma c_t) \quad (10)$$

This represents an assumption that individuals have some preference for their contribution, which is fixed by the model or social motive they are implementing, but which is contaminated by Gaussian noise. For simplicity, I fix the precision of the noise parameter $\sigma c_t = 0.1$ for all applications of the model. The mean of the distribution in a trial νc_t represents the preferred contribution, and is defined as

$$\nu c_t = \begin{cases} c_t^{EWA}, & if\ Z = 0 \\ c_t^{cc}, & if\ Z = 1 \end{cases} \quad (11)$$

Equation 11 implements the latent mixture. Here, c_t^{EWA} represents the contribution predicted from the EWA model, and c_t^{cc} represents the contribution from the CC model. *Z* functions as an assignment parameter. Thus, *Z* indicates which model is more appropriate for explaining the investment choices of an individual, where *Z* = 0 indicates the selfish utility optimization motive implemented in the EWA model, and where *Z* = 1 indicates the conditional cooperation motive implemented in the CC model.

To infer the latent binary parameter *Z*, we can use a Beta-Bernouli model

$$Z \sim Bernouli(\psi) \quad (12)$$

$$\psi \sim Beta(1,1) \quad (13)$$

Thus the free parameter ψ represents the probability that a given individual is conditionally cooperating (i.e. when ψ is high) or choosing selfishly (i.e. when ψ is low) throughout the game. The parameter ψ may therefore be interpreted as reflecting the process of selection between selfish and cooperative motives, and it is the main target for inference throughout applications of the model. I label this the Social Motive Selection or SMS model, to distinguish it from the EWA and CC models of which it is composed.

3.3 Modeling social motives using reinforcement learning – EWA with elaborated utility

A latent mixture approach assumes that social motives are mutually exclusive – that one is either acting selfishly by focusing on one's own utility, or acting (conditionally) cooperatively by focusing on others' contributions. Janssen and Ahn (2006) have provide an approach based on reinforcement learning which does not make this assumption, and which therefore functions as an alternative model to the SMS. This model is an elaboration of the EWA model. It defines social motives strictly in terms of a utility function $U(c_{n,t}, Ga_{t-1})$, which relates one's own and others' outcomes in the public goods game

$$U(c_t, Ga_t) = \begin{cases} \chi_{win}.R_{other} + (1-\chi_{win}).R_{self}, & if\ R_{self} \geq R_{other} \\ \chi_{loss}.R_{other} + (1-\chi_{loss}).R_{self}, & if\ R_{self} < R_{other} \end{cases} \quad (14)$$

Here, R_{self} denotes the reward received on a trial, and R_{other} denotes the average reward received by others in the group. The χ parameters are weighing parameters, which regulate evaluations of others' outcomes relative to one's own, depending on whether or not one received more or less than the group average. Using this function, competitive players can be identified as those for whom $\chi_{loss} < \chi_{win} < 0$. People motivated by social welfare considerations can be identified for those whom $0 < \chi_{loss} < \chi_{win} < 1$. And people motivated by aversion to inequality can be identified as those for whom $\chi_{loss} 0 < \chi_{win} < 1$ (Janssen & Ahn, 2006).

To explain public goods behavior, Janssen and Ahn (2006) replace the payoff calculation in the standard EWA model with this utility function, such that the set of attractions is now defined as

$$A_{j,t} = \frac{\phi.N_{t-1}.A_{j,t-1} + (\delta + \Big((1-\delta).(c_{n,t} == j)\Big).U(c_{n,t}, Ga_{t-1})}{N_t} \quad (15)$$

4. Model Evaluation

4.1 Note on methods and data sources

In the following sections, the models presented above are applied to data and simulations using Hierchical Bayesian modeling. A Bayesian approach is required for this purpose. This is because, as a probabilistic model of the processes underlying choice behavior, the latent mixture model is Bayesian in nature (Lee & Wagenmakers, 2014; Bartlema et al., 2014). Gibbs sampling was used for all inference for all models reported in these sections. Gibbs sampling is a Markov Chain Monte Carlo algorithm commonly used in statistical inference. Sampling was implemented using JAGS software (Plummer, 2003) via the R2jags package (Su & Masano, 2012). All model code is available on the Open Science Framework https://osf.io/meh5w/.

Bayes Factors are reported for all hypothesis tests. Bayes factors were calculated using the Savage-Dickey density ratio, which is the ratio of the prior and posterior distributions at the value of interest (here the 0 value, which represents the point null hypothesis, see Lee & Wagenmakers, 2014). In short, a Bayes Factor represents the strength of evidence for or against a hypothesis.

All Bayes Factors of interest reported below are greater than 100, which represents strong evidence for the hypothesis, and which provides very good reason to believe that sufficient data are included in each data set for

the inferences being made. For parameter estimates, 95% Bayesian credible intervals are reported. Individual-level priors are described below, and summarized in table 1. All data used in this paper are freely available online, and the specific data used in each case is cited in the relevant methods subsection. Links to data are provided in the relevant references.

4.2 A priori model recovery

To evaluate the inferential reliability of the model, I conducted model recovery from forward simulations (Wilson & Collins, 2019; Heathcote, Brown, & Wagenmakers, 2015). I constructed groups of software agents to interact in a public goods game. Half the agents were randomly assigned to choose their contributions using the EWA model, and half to choose their contributions using the CC model. A fully Bayesian version of the latent mixture model was used to infer which model was most likely to generate the simulated data, and to compare the inferred assignment parameter to the true parameter used for the simulation. Using this method, we can say that if the inference reliably matches the true assignment, we may have more confidence in the SMS model's results when it is applied to data from real choices.

4.2.1 Method

For model recovery, 100 groups of ten agents were constructed to simulate ten rounds of a public goods game. Each agent was allocated 20 tokens, and the program simulated the number of tokens each agent contributed in each round. The multiplication factor π for the public good was fixed at $\pi = 1.5$. Agents were assigned so that either the EWA or the CC model determined their contributions, with the assignment based on samples from a Bernouli(0.5) distribution.

To ensure that model recovery was not sensitive to any specific combinations of EWA or CC model parameters, simulation parameters were set randomly wherever possible. For each agent in each group, all rate and weighting parameters for both the CC and the EWA models were sampled from Beta(1,1) distributions. The choice consistency parameter θ was sampled from a Gamma(0.1,0.1) distribution.

Some constraints were applied to agents assigned to the CC model. The intercept of the beliefs/preferences model B_0 was fixed at zero, reflecting a psychological assumption that conditionally cooperating agents will not contribute if they do not believe others will contribute too. The initial-weighting-for-beliefs parameter – here denoted ω_1 – was sampled from a Uniform(0.9,1) distribution. This reflects a psychological assumption that

cooperating agents will start with a higher tendency to contribute according to their beliefs rather than their own preferences. Agents' initial belief about the contributions of others – here denoted Gb_1 – was fixed at the maximum of 20 to simplify the simulations.

I then applied the latent mixture model described in the previous section to the simulation outputs. The same distributions used to generate model parameters were also used as priors in the latent mixture model, with three exceptions. For the CC model, initial beliefs about group contributions Gb_1 were assumed to be reflected in the first contribution made in the game; the prior for the intercept of the beliefs/preferences model B_0 was a uniform distribution across the range of possible contributions; and the prior for the initial-weighting-for-beliefs parameter ω_1 was (unrestricted) Beta(1,1). For the latent mixture, I assumed a default Beta-Bernouli model for the assignment parameter *Z*. All priors are presented in table 1.

Table 1: Priors used for Bayesian inference using social motives model

Experience Weighted Attraction Model		
	δ – Weighting of forgone payoffs	Beta(1,1)
	ρ – Discounting of previous trials	Beta(1,1)
	Φ – Memory for previous attractions	Beta(1,1)
	χ_m – Utility weight when others gain more	Beta(1,1)
	χ_l – Utility weight when others gain less	Beta(1,1)
	θ – Choice consistency	Gamma(0.1 , 0.1)
Conditional Cooperation Model		
	Gb_1 – Initial beliefs about others	20 (constant)
	ω_1 – Initial weighting for beliefs	Uniform(.9,1)
	λ – Decay rate in weighting for beliefs	Beta(1,1)
	γ – Learning rate for belief updating	Beta(1,1)
	B_0 – Intercept of beliefs/preferences model	Uniform(0,N. tokens)
	B_p – Slope of beliefs/preferences model	Beta(1,1)
Social Motive Selection Model		
	Z – Assignment parameter	Bernouli(ψ)
	ψ – Assignment prior	Beta(1,1)

4.2.2 Results

I compared the Maximum a Posteriori of the assignment parameter (*Z*) in the latent mixture model to the true assignment (i.e. the recorded sample from the assignment distribution in the simulations). The model could accurately recover whether simulated data had been generated by the EWA or the CC model for 92.9% of assignments in the simulation. The model correctly identified 94.8% of *CC* agents, however it mischaracterized 9.0% of EWA agents' responses as having been generated by the *CC* model. This suggests that the model is highly accurate, although it may be slightly biased towards the *CC* model in its categorization.

4.3 Model comparison

I applied separate Hierarchical Bayesian implementations of the the CC model, the SMS model, and both versions of the EWA model to a real data set, and compared the Deviance Information Criteria for all four. DIC is used in model comparison to estimate the prediction error for Bayesian hierarchical models. DIC includes a correction for model complexity, and models with lower DIC are preferred.

4.3.1 Method

The data used for model comparison were from an experiment reported by Fraser and Nettle (2020), and are freely available online (Fraser & Nettle, 2019). This study was designed to investigate the effects of hunger on social decisions, but this aspect of the dataset is ignored. The study also included two separate social interaction conditions: a standard public goods game, and a game in which individuals were able to pay to punish group members for failing to contribute enough (peer punishment). Only data from the standard game are analyzed here.

66 groups of four people participated in a ten-round game. For each round, participants were allocated 20 tokens, and they could choose how many to invest in the public good. The multiplication factor for the game was 1.4, yielding a marginal per capita return of 0.35 tokens. For further details on the experiment, see Fraser and Nettle (2020).

Separate versions of the four models were applied to the individual contributions recorded in the dataset. The subject-level priors for the models were identical to those used for model recovery (table 1).

4.3.2 Results

The DIC was 14546.1 for the EWA model, 14652.1 for the CC model, 14443.6 for the SMS model, and 14869.6 for the EWA model with social motives. Despite being the most complex model, the SMS model has the lowest DIC, and should therefore be selected as the best fitting model of the four. In the following section I apply this model across a range of well known cases, to show how it may be used to address substantive research questions.

5. Case studies

Having shown that the SMS model – and the social motive selection process it represents – can accurately recover known assignments from simulations using the EWA and CC models, and that it can more accurately predict actual investments in a real public goods game than either model alone, we turn now to applying the model in a series of cases.

The aim here is to show how the model can be used to infer the effects of well-studied contextual manipulations on social motives in public goods games, and to demonstrate that the approach offered should be reliable for asking novel research questions (Heathcote et al., 2015). In all cases, analyses are conducted using hierarchical versions of the latent mixture model tailored to the research question in the case. For hypothesis tests, reported Bayes Factors were calculated using the Savage-Dickey density ratio, and individual-level priors are the same as reported in table 1.

5.1 The effects of increasing marginal per capita return on cooperation

The first case investigates the effects on cooperation of increasing the multiplication factor in the game. As discussed in section 2.1, it is straightforward to derive a hypothesis concerning investment under this manipulation. We should expect that increasing the multiplication factor – and thereby the marginal per capita return for individuals – will incentivize people to contribute more in the game. It is less straightforward to derive a prediction of the social motive underlying this behavior. Cartwright and Lovett (2014) have interpreted higher contribution rates in these circumstances as evidence of incentivization of conditional cooperation. If this is true, then we should expect our ψ parameter – which represents the probability that an individual cooperated during the game – to increase when the multiplication factor is higher. However, it is also possible that higher investments are themselves directly incentivized by the increased multiplication factor. If this is the case, then

one might interpret the increase in investment in terms of selfish motives grounded in reinforcement learning, of the sort formalized in the EWA model. If this is true, then we should expect ψ to be lower when the multiplication factor is higher. These hypotheses are investigated in the present case.

5.1.1 Method

The data used for this case are from experiments reported by Burton-Chellew et al. (Burton-Chellew & West, 2013; Burton-Chellew, Nax, & West, 2015), which are freely available online (Burton-Chellew, Nax, & West, 2014). In the study, 116 participants performed a standard iterated public goods game in groups of four. Participants were given 40 tokens to save or invest in each round, and the game lasted for 20 rounds. In this experiment the groups were not stable, but randomly reshuffled in each round of the game, and participants were informed of this. The study also included conditions in which participants were informed they were playing against a "black-box" or algorithm instead of a group of other people, but we ignore this aspect of the dataset.

The same participants completed the experiment under two different incentive conditions. In the low incentive condition, the public good was multiplied by 1.6, for a marginal per capita return of 0.4 tokens (i.e. less than 1). In the high condition, the public good was multiplied by 6.4, for a marginal per capita return of 1.6 tokens. For a full description of the data, see Burton-Chellew and West (2013).

To test our main hypothesis, I used a hierarchical Bayesian within subjects t-test (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010) to compare the average ψ parameter between conditions. For each participant, the same latent mixture model was applied separately to both conditions. I re-parameterized the beta prior for the ψ parameter

$$\psi \sim Beta(Shape_1, Shape_2) \tag{16}$$

$$Shape_1 = \mu\psi_{s,c} + \sigma_s \tag{17}$$

$$Shape_2 = \left(1 - \mu\psi_{s,c}\right) + \sigma_s \tag{18}$$

so that $\mu\psi_{s,c}$ is the expected value or mean probability of cooperation for a participant in a condition, and σ_s is the subject-specific concentration parameter. I assume the standard normal distribution as a prior for σ_s. For the final group-level comparison, I probit transformed the expected value, such that

$$Probit\left(\mu\psi_{s,1.6}\right) = \mu_s \tag{19}$$

for the condition in which the public good is multiplied by 1.6, and

$$Probit(\mu\psi_{s,6.4}) = \mu_s + \alpha_s \quad (20)$$

for the condition in which the public good is multiplied by 6.4. The parameter μ_s is thus the probit transformed probability of conditional cooperation in the condition in which the public good is multiplied by 1.6, and α_s is the subject-level difference between the two conditions. I then modeled α_s as normally distributed at the group level, such that

$$\alpha_s \sim Normal(\alpha_\mu, 1) \quad (21)$$

This gives the mean group difference in the probability of conditional cooperation between the two conditions. I assume the standard normal distribution as a prior for α_μ.The Bayes factor is reported for the hypothesis test, and the Bayesian credible intervals for the difference parameter α_μ is reported as an estimate of the effect of incentive condition on elicitation of the cooperative social motive.

5.1.2 Results

The effects of the different multiplication factors on aggregate contributions across trials is represented in the left panel of figure 1. When the marginal per capita return for the game was 0.4, aggregate investment decreased across trials, in line with the typical pattern of contributions. When the marginal per capita return was 1.6 (i.e. greater than one), participants increased their contributions as expected.

The posterior for α_μ – the model parameter representing the probit transformed group-level difference in probability of conditional cooperation between the two conditions – is represented in the middle panel of Figure 1. The negative axis values indicates evidence that increasing the multiplication factor decreases cooperation, and selects for the selfish motive represented in the EWA model (BF > 100, CI = -2.87 to -1.22).

The right panel of Figure 1 represents maximum a posteriori estimates of the probability of cooperation ψ at the level of individual participants, after reversing the probit transformation. The figure shows that when the marginal per capita return on the game is more than one, all participants adopt the selfish social motive represented in the EWA model.

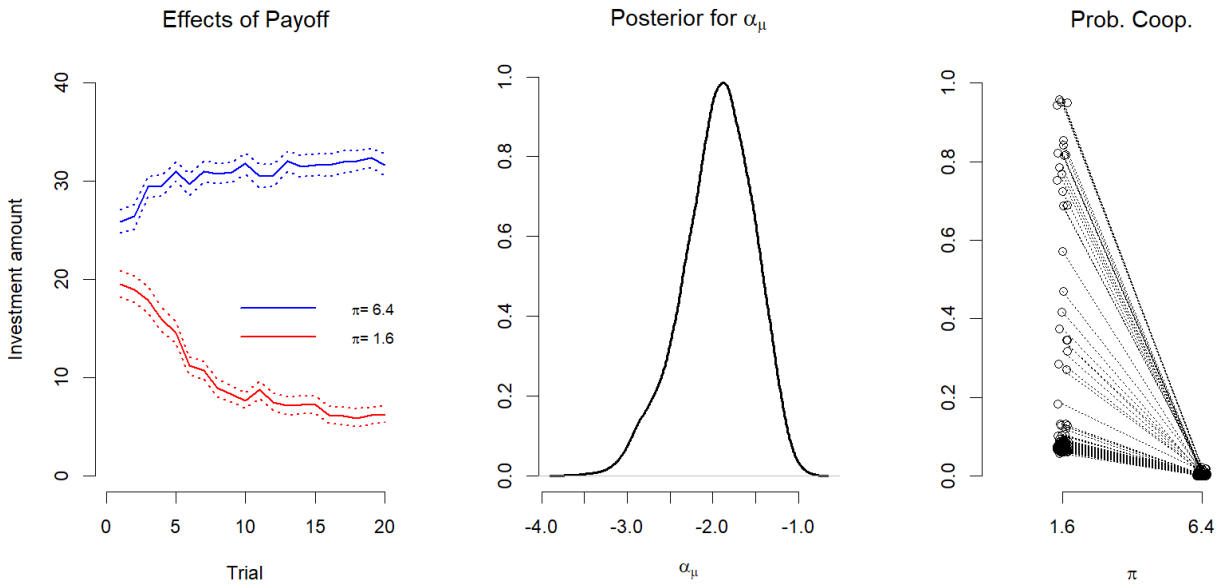

Figure 1: Results for the case comparing low ($\pi = 1.6$) and high $\pi = 6.4$) multiplication factors in the public goods game. The left panel shows the aggregate investment for each condition. The middle panel shows the posterior distribution for the group-level difference parameter α_μ, when comparing the probability of cooperation between the two conditions (Bayesian within subjects t-test). The right panel shows the individual-level maximum a posteriori estimate of the probability of cooperating across the two conditions.

5.2 The effects of possibility for peer punishment

The second case investigates the effects on cooperation of allowing peer punishment during the game. Peer punishment is possible if individuals may contribute a number of their own tokens in order to remove tokens from other people in the group. As discussed in section 2.1, this is typically done when individuals fail to contribute what the "punisher" considers to be a fair or appropriate amount. The effect of this possibility is typically to increase investments in the game, and we should expect to observe this effect in the present case.

Once again, it is not trivial to derive a hypothesis concerning the underlying motivational cause of this outcome, and two separate explanations are possible. The first relies on associative learning mechanisms (e.g. Fehr & Gächter, 2000). When punishment is possible, individuals realize that repeated punishments will reduce their

overall earning in the game. At the same time, they observe that others are investing more to avoid punishment. Thus, increased contribution is explained as a result of reinforcement learning and selfish motives. If this explanation is true, then in terms of the SMS model, the possibility of punishment should lead to a decrease in ψ, as the selfish social motive implemented in the EWA model is elicited.

The second explanation relies on cooperative mechanisms. When punishment is possible, failures to cooperate become more salient, and so the effect of punishment is to increase contributions by emphasizing and enforcing a norm of cooperativeness (e.g. Herrmann et al., 2008). If this explanation is true, then the possibility of punishment should lead to an increase in ψ, as the cooperative social motive implemented in the CC model is elicited. We investigate these hypotheses in the present case.

5.2.1 Method

The data used for this analysis are from the experiment reported by Fraser and Nettle (2020) and used in section 3.5 above. There I applied the model to the standard public goods condition only. Here, I compare model parameters between the standard version of the game, and the version with peer punishment. Both conditions are described in full in the Fraser and Nettle (2020) reference.

The dataset includes responses from 66 groups of four people participating in a ten-round game in which participants were allocated 20 tokens. The multiplication factor for the game was 1.4, yielding a marginal per capita return of 0.35 tokens. The same people participated in both conditions. Groups were stable throughout the experiment.

Because groups were stable, group membership was also represented in the model. To do this, each of the models were specified at both the individual and group level. To test our main hypothesis, I used the same hierarchical Bayesian t-test as applied in section 5.1. In this model, the probit transformed expected value for the parameter ψ in the standard game is given by

$$Probit\left(\mu\psi_{s,g,standard}\right) = \mu_{s,g} \tag{22}$$

and in the game with punishment it is given by

$$Probit\left(\mu\psi_{s,g,punish}\right) = \mu_{s,g} + \alpha_{s,g} \tag{23}$$

such that the parameter $\mu_{s,g}$ is the probit transformed subject-level probability of conditional cooperation in the standard game, $\alpha_{s,g}$ is the subject-level difference between the two games, and the subscript g indicates the

group to which the individual belonged. I then modeled $\alpha_{s,g}$ as normally distributed at the experiment level, as in section 5.1, again giving α_{μ} as a measure of the difference in the probability of conditional cooperation between the two conditions. I again report the Bayes factor for the hypothesis test, and the Bayesian credible intervals for the difference parameter α_{μ} as an estimate of the effect of incentive on cooperation.

5.2.2 Results

The main trend in aggregate contributions across trials is represented in the left panel of figure 2. In the standard game, aggregate investment decreased across trials, in line with the typical pattern of contributions. When peer punishment was possible, participants increased their contributions as expected.

The posterior for α_{μ}– the model parameter representing the probit transformed group-level difference in probability of conditional cooperation between the two conditions – is represented in the middle panel of Figure 2. The figure indicates strong evidence that allowing peer punishment increases cooperation, and selects for the cooperative motive implemented in the CC model (BF > 100, CI = 0.53 to 1.24).

The right panel of Figure 2 represents maximum a posteriori estimates of the probability of cooperation ψ at the level of individual participants, after reversing the probit transformation used for inference. The figure shows that when punishment is possible, almost all participants adopt the conditional cooperation motive. This is in contrast to the effect of increasing marginal per capita return. Therefore, the model reveals that even though aggregate investment behavior is the same in both cases, this behavior is produced by opposite social motives. The ability to infer such differences is a strength of the present approach.

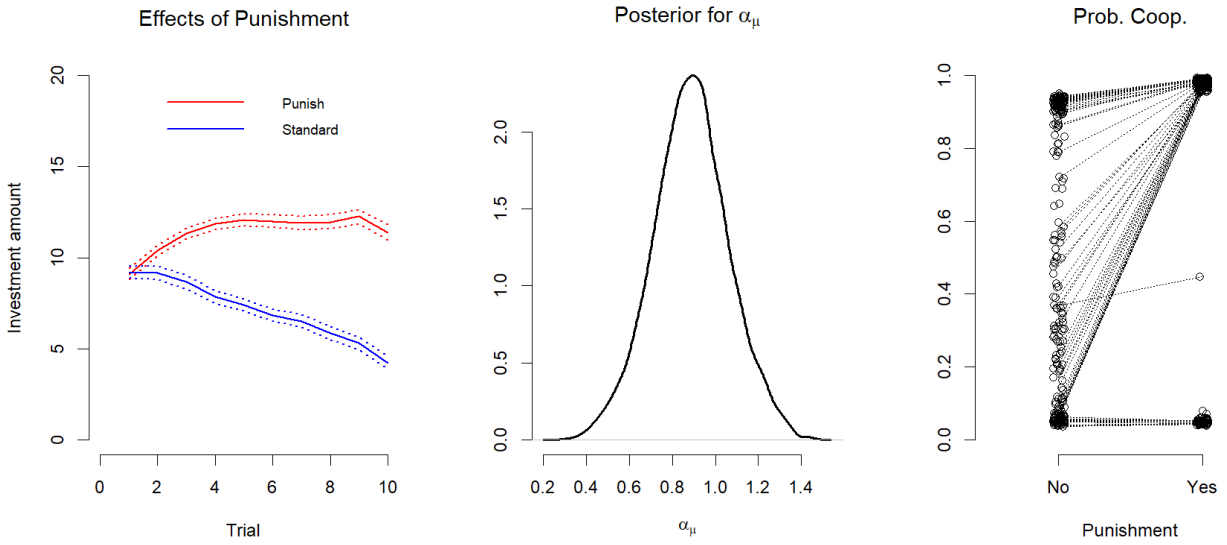

Figure 2: Results for the case comparing standard and peer punishment versions of the public goods game. The left panel shows the aggregate investment for each condition. The middle panel shows the posterior distribution for the group-level difference parameter α_μ, when comparing the probability of cooperation between the two conditions (Bayesian within subjects t-test). The right panel shows the individual-level maximum a posteriori estimate of the probability of cooperating across the two conditions.

5.3 Cultural influence on social motives – civic norms, punishment, and cooperation

The third case investigates cultural variation in the influence of peer punishment on promoting cooperation. In recent years, researchers have shown interest in the effects of culture on how people solve shared resource problems. For example, Nishi et al. (2017) have shown differences in trust and cooperation in public goods games between people from the USA and India; Gerkey (2013) has shown differential effects of institutional framing of public goods games in traditional societies in post-soviet Russia; and Xygalatas et al. (2013) have shown that extreme rituals increase charitable giving and contributions in trust games.

In one of the most far-reaching single studies on this topic, Herrmann, Thoni, and Gächter (2008) conducted public goods experiments across 15 culturally diverse countries. Participants took part in both the standard public goods game, and a version in which punishment was possible. The authors also reported an index of each country's adherence to norms of civic cooperation. The index – which had previously been measured in representative national samples in the countries investigated (for details and methodology see Knack & Keefer,

1997) – provides a measure of civic attitudes, including for example attitudes to tax evasion and evasion of fares on public transport. The authors found that in countries characterized by higher adherence to norms of civic cooperation, people were more likely to punish those who contribute less to the public good than they themselves did.

An important question left open by Herrmann et al.'s study is how civic norms are related to the selection of social motives – or whether greater adherence to civic norms increases the influence of punishment in promoting cooperation. In a recent meta-analysis, Balliet and van Lange (2013) raised the related question of whether the level of trust in a society increases or decreases the influence of punishment on public good investments. They framed the question in terms of two competing hypotheses. The first is that we might expect societies with low trust to rely more on punishment for enforcing cooperation, because in such societies cooperation is not incentivized by shared social norms. The second is that we might expect societies with high trust to rely more on punishment for enforcing cooperation, because social norms might be supported by the increased likelihood of peer punishment in such societies. Balliet and van Lange analyzed data from 18 societies and purported to show that the latter hypothesis is correct – that the possibility of punishment is more effective in increasing investment in high-trust societies. Following the study by Herrmann et al., a natural interpretation of Balliet and van Lange's result is that punishment increases investment via the increased elicitation of cooperative social motives in societies with high adherence to norms of civic cooperativeness. Here we apply the SMS model to Herrmann et al.'s data to directly investigate this hypothesis.

5.3.1 Method

Herrmann et al. (2008) asked people to play public goods games in 15 countries. These were Australia, Denmark, Oman, Russia, Korea, USA, the UK, China, Greece, Switzerland, Belarus, Ukraine, Germany, Turkey, and Saudi Arabia. In each country, the same participants played both the standard public goods game and the version of the game with the possibility of peer punishment. In all countries, people played in groups of four. They were given 20 tokens to keep or invest per round, and the multiplication factor for the public good was 1.4. Each game lasted 10 rounds. A total of 280 groups participated in the experimental across the 15 countries. Full details are given in Herrmann et al. (2008), and the data is freely available in Herrmann, Thöni, & Gächter (2017).

To test our main hypothesis, I used a hierarchical linear regression over the parameter $\alpha_{s,g}$, such that

$$\alpha_{s,g} = \beta_0 + \beta_c . civic_g \quad (24)$$

Where the subscript *s* denotes a subject, the subscript g denotes a group, and $\alpha_{s,g}$ represents the subject-level difference in probability of cooperation between the two conditions within groups (see equation 18). The vector $civic_g$ represents (standardized) national-level scores for adherence to civic norms for the group, and β_0 and β_c are the regression coefficients.

Data from the Oman groups were not included in the analysis, because no adherence to civic norms score is available for that country. Thirteen groups were removed, leaving 256 groups in the analysis, for a total of 1064 participants from 14 different societies.

5.3.2 Results

The results of the analysis are presented in Figure 3. The left panel shows the posterior distribution for the regression coefficient β_c, and indicates strong evidence that punishment is more effective at eliciting cooperative social motives in societies with stronger adherence to civic norms of cooperation (BF > 100, CI = .29 to .64). The other two panels describe this effect at the individual level. The middle panel shows the relationship between adherence to civic norms and the maximum a posteriori estimate for the probability of cooperation in the standard public goods game. The probability of cooperation is represented at the *group* level. This figure shows that regardless of adherence to civic norms, most groups are equally likely to choose their investments based on selfish or cooperative motives (with more variation possible at the individual level). The right panel shows the relationship between norms and the probability of cooperation in the game with punishment. The probability of cooperation is also represented at the group level in this panel. This figure shows that in societies with low adherence to civic norms, punishment has little effect on social motive selection, but that as adherence to civic norms increases, punishment elicits the cooperative social motive.

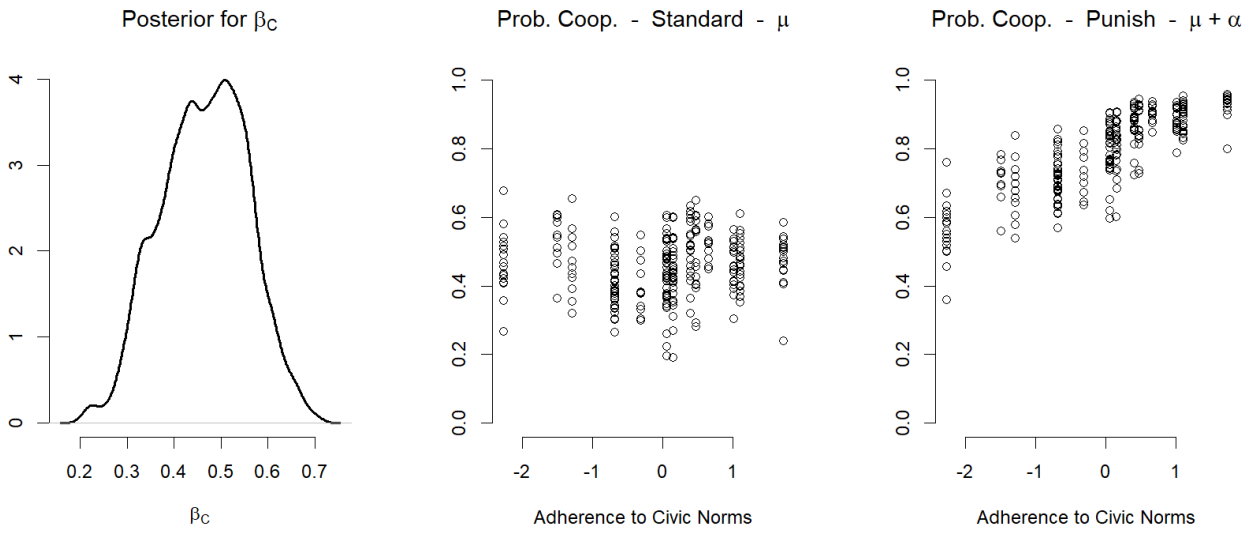

Figure 3: Results for the case investigating the relationship between adherence to civic norms and the probability that punishment increases cooperation, across culturally different societies. The left panel shows the posterior distribution for the group-level effect of adherence to civic norms β_C. The middle panel shows the relationship between civic norms and maximum a posteriori estimates of the probability of cooperating in the standard public goods game at the group level. The right panel shows the relationship between civic norms and maximum a posteriori estimates of the probability of cooperating in the public goods game in which peer punishment is possible.

6. Discussion

The main aim of this paper is to more tightly unify experimental and theoretical approaches to understanding choice in shared resource problems. To achieve this, I return to an old theoretical idea: that individual and contextual variation in the way people invest in shared resources is the product of individual and contextual variation in their social motivation. Drawing heavily on two well-developed formal theories of choice behavior in games – the EWA model based on selfish reinforcement learning, and the CC model based on conditional cooperation – we have seen how latent mixture modeling can be applied to understand investment behavior in terms of selection between the social motives represented by the two separate theories. We have seen that this approach can accurately predict the social motives used to generate choice in artificial simulations, we have seen that the latent mixture SMS model can outperform the EWA and CC models when they are applied alone, and

we have seen that the SMS model can outperform an alternative model of social motives framed only in terms of reinforcement learning and utility functions.

I have also applied the model to three prominent cases. The first concerns the effects of marginal per capita return on social motive selection. Here we have seen that when marginal per capita return is higher, people are more likely to maintain high investments; and this is more likely to be the result of selfish motives to maximize utility. The second case concerns the effect of peer punishment. Here we have seen that when peer punishment is possible, people are also more likely to maintain high investments, although this is more likely to be the result of cooperative motives. These cases are prima facie interesting, because they show how the model can be used to identify which kinds of social motives are elicited by task constraints (Kopelman et al., 2002; Zelmer, 2003) and social dynamics (Fehr & Gächter, 2002; Fischbacher & Gächter, 2010).

The results also have more general implications for experimental research. It is common practice to assume that if a group maintains a higher average rate of investment in an experimental condition, then this is because the group are cooperating more during the experimental treatment. If a group invests less in a condition, then this is because they are acting more selfishly. Thus, patterns of investment are often taken to indicate, or to be behaviorally identical to, latent social motives (e.g. Cartwright & Lovett, 2014; Fosgaard & Piovesan, 2015; Nagatsu et al. 2018). In learning that both cooperative and selfish motives can be equally effective in terms of increasing contributions in public goods games – and more generally that these different social motives can lead to almost identical aggregate behavior – we may begin to doubt this assumption. These cases therefore highlight the general usefulness of the latent mixture modeling approach presented in this paper.

The third case concerns the effect of culture on peer punishment. Here we have seen that in societies with greater adherence to civic norms of cooperation, peer punishment has a greater effect in eliciting cooperative motives. This case is interesting, because it suggests that the selection of social motives might be understood as a general mechanism for implementing individual and contextual variation in social decision-making. The first two cases support the basic idea that context creates variation in social decision-making by influencing which social motives produce choices at any given time. From the third case, we might go further and hypothesize that culture operates by more permanently strengthening certain motives instead of others. More theoretical work will be needed to flesh out the details of this hypotheses. Specifically, a richer cognitive model of the social motive selection mechanism may be required to fully understand how context, culture, and even personality are related to the selection of social motives. This remains a task for future research.

Two further avenues of theoretical research remain open from the present model. The first is the inclusion of a broader palette of social motives. The present model incorporates formalizations of selfish and cooperative

motives, but does not include competitive and altruistic motives, which have also been the focus of early theoretical work. The reason we have ignored these motives in the present paper is that they have been shown to be empirically rare (e.g. Burton-Chellew et al., 2015; Burton-Chellew & West, 2013). However, should altruism and competition be of theoretical or applied interest to any specific researcher, these motives should be easy to include within the present framework, given the modular structure of the latent mixture model.

The second open theoretical question concerns the functional form of preferences assumed in the CC model. With equation 5, we have assumed that the relationship between beliefs (about the investments of other people) and preferences (for one's own investment) is linear. Fischbacher and Gächter (2010) also discuss the existence of rare "triangle" cooperators in their data. These are individuals who prefer to match other people's investments up to a point, beyond which their preferences decrease below those of other people. From a social motives perspective, there are two possible explanations of this kind of behavior. The first is that the triangular form identified by Fischbacher and Gächter is a real special case of cooperation; that some individuals really do prefer not to cooperate beyond a certain investment threshold. If this is the case, then triangle cooperation should be incorporated in our formalization of the CC model. The second possible explanation is that triangle cooperators are not really cooperators, but have rather learned, through selfish payoff optimization, to favor specific contribution amounts. For simplicity, we have assumed the latter explanation in the present model. Further empirical research, using appropriately designed experiments, will be required to answer this question properly, and may be used to update the SMS model.

The approach we offer here also opens a number of lines of experimental research. Perhaps the most promising concerns the role of communication in shared resource problems. There is an extensive literature documenting the finding that communication leads to higher rates of investment within groups (Dawes, van de Kragt, & Orbell, 1990; Isaac & Walker, 1988b; Nagatsu et al., 2018; for discussion see Kopelman et al., 2002). This research is important, because it is of great relevance with regard to the design of institutions for optimizing solutions to shared resource problems (Smith, 2008).

Communication may be understood as a process of alignment in social motive selection, and the SMS model may provide a useful tool for understanding the social dynamics underlying different communicative mechanisms in shared resource contexts. For example, Nagatsu et al. (2018) report an experiment in which they divided people into groups of selfish or cooperative players. They then let group members communicate their expectations about other people's contributions. They found that in the selfish groups, the possibility of communication increased contributions. The authors provide two possible explanations of this empirical effect. The first assumes that the effect is due to social influence – that the communication of expectations establishes a

social context that elicits a preference to avoid disappointing those expectations. The second is due to payoff optimization and learning – that when others signal high expectations, they are also signaling a high level of commitment, such that cooperation may be the profitable strategy in the long run. The approach provided here should be ideal for identifying the underlying motive selection processes in these kinds of experiments, and I recommend its application in such settings. Similarly, we may expect the model to be useful in any experimental context in which scholars are interested in determining the latent motives causing choices in shared resource investments.

References

Andreoni, J. (1995). Warm-glow versus cold-pickle: The effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics, 110*(1), 1-21.

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

Akerlof, G. A., & Kranton, R. E. (2005). Identity and the economics of organisations. *Journal of Economic Perspectives, 19*(1), 9-32.

Axelrod, R. (1984). *The Evolution of Cooperation.* New York: Basic Books.

Balliet, D., & Van Lange, P. A. M. (2013). Trust, punishment, and cooperation across 18 societies: a meta-analysis. *Perspectives on Psychological Science, 8*(4). 363-379.

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology, 59,* 132-150.

Bereczkei, T., Birkas, B., & Kerekes, Zs. (2010). The presence of others, prosocial traits, Machiavellianism. A personality x situation approach. *Social Psychology, 41,* 238-245.

Blanco, M., Engelmann, D., Normann, H. T. (2011). A within-subjects analysis of other regarding preferences. *Games and Economic Behavior, 72,* 321-338.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review, 90,* 166-193.

Burton-Chellew, M. N., Nax, H. H., & West, S. A. (2015). Payoff-based learning explains the decline in cooperation in public goods games. *Proceedings of the Royal Society B: Biological Sciences, 282*(1801),

Burton-Chellew, M. N, Nax, H. H., & West, S. A. (2014), Data from: Payoff-based learning explains the decline in cooperation in public goods games *Dryad, Dataset*, https://doi.org/10.5061/dryad.cr829

Burton-Chellew, M. N., & West, S. A. (2013). Prosocial preferences do not explain human cooperation in public-goods games. *Proceedings of the National Academy of Sciences, 110*(1), 216-221.

Butler, D. M., & Kousser, T. (2015). How do public goods providers play public goods games. *Legislative Studies Quarterly, 40*(2), 211-240.

Camerer, C., & Ho, T. (1999). Experience weighted attraction in normal form games. *Econometrica, 67*(4), 827-874.

Cartwright, E. J., & Lovett, D. (2014). Conditional cooperation and the marginal per capita return in public goods games. *Games, 5*(4), 234-256.

Cartwright, E. J., & Ramalingam, A. (2019). Framing effects in public goods games: Choices or externalities? *Economics Letters, 179,* 42-45

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

Chung, D., Yun, K., & Jeong, J. (2015). Decoding covert motivations of free riding and cooperation from multi-feature pattern analysis of EEG signals. *Social Cognitive and Affective Neuroscience, 10,* 1210-1218.

Czibor, A., & Bereczkei, T. (2012). Machiavellian people's success results from monitoring their partners. *Personality and Individual Differences, 53,* 202-206.

Czibor, A., Vincze, O., & Bereczkei, T. (2014). Feelings and motives underlying Machiavellian behavioral strategies: narrative reports in a social dilemma situation. *International Journal of Psychology, 49*(6), 519-524.

Dawes, R. M., van de Kragt, A. J. C., & Orbell, J. M. (1990). Cooperation for the benefit of us – Not me, or my conscience. In J. J. Mansbridge (Ed) *Beyond Self Interest.* Chicago: Chicago University Press. pp 97-110.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences, 275*(1637), 871-878.

Erev, I., & Roth, A. E. (1998). Predicting how people play games. Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review, 88*(4), 848-881.

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

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415,* 137-140.

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

Fischbacher, U., & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review, 100*(1), 541-556.

Fosgaard, T. R., & Piovesan, M. (2015). Nudge for (the public) good: How defaults can affect cooperation. *PLoS One, 10*(12).

Fowler, J. H. (2005). Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences, 102*(19), 7047-7049.

Fraser, S, & Nettle, D. (2019). Data and code for Fraser and Nettle, 'Hunger affects social decisions in a Public Goods Game but not an Ultimatum Game' [Data set]. *Zenodo*.

Fraser, S., & Nettle, D. (2020). Hunger affects social decisions in a multi-round public goods game but not a single shot ultimatum game. *Adaptive Human Behavior and Physiology, 6*, 334-355

Gerkey, D. (2013). Cooperation in context: public goods games and post-soviet collectives in Kamchatka, Russia. *Current Anthropology, 54*(2).

Heathcote, A., Brown, S., & Wagenmakers, E. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann and E. J. Wagenmakers (Eds). *An Introduction to Model-Based Cognitive Neuroscience.* Springer. pp25-48

Herrmann, B., Thöni, C., Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362-1367.

Herrmann, B., Thöni, C., & Gächter, S. (2017). Data from: Antisocial punishment across societies. *Dryad, Dataset*

Hichri, W. (2005). The individual behavior in a public goods game. *International Review on Public and Non Profit Marketing, 2*(1), 59-71.

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

Isaac, R. M., & Walker, J. M. (1988b).Communication and free riding behavior: The voluntary contribution mechanism. *Economic Inquiry, 36*(4), 585-608.

Janssenn, M. A., & Ahn, T. K. (2006). Learning, signaling, and social preferences in public goods games. *Ecology and Society, 11*(2).

Kadane, J. B., & Larkey, P. D. (1982). Subjective probability and the theory of games. *Management Science, 28*(2), 113-120.

Kary, A., Hawkins, G. E., Hayes, B., Newell, B. R. (2017). A Bayesian latent mixture model approach to assessing performance in stock-flow reasoning. *Judgment and Decision Making, 12*(5), 430-444.

Krawczyk, M., & Le Lec, F. (2015). Can we neutralize social preferences in experimental games? *Journal of Economic Behavior & Organization, 117,* 340-355.

Kelley, H. H., & Stahelski, A. J. (1970). Social interaction basis of cooperators and competitors beliefs about others. *Journal of Personality and Social Psychology, 16*(1), 66-91.

Knack, S., & Keefer, P. (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *Quarterly Journal of Economics, 112(4),* 1251-1288

Kopelman, S., Weber, J. M., & Messick, D. M. (2002). Factors influencing cooperation in commons dilemmas: A review of experimental psychological research. In National Research Council (Ed). *The Drama of the Commons*. Wachington D. C.: The National Academies Press. pp. 113-156.

Kurzban, R., & Houser. D. (2001). Individual differences in cooperation in a circular public goods game. *European Journal of Personality, 15,* S37-S52.

Larrouy, L., & Lecouteux, G. (2017). Mindreading and endogenous beliefs in games. *Journal of Economic Methodology, 24*(3), 318-343.

Ledyard, J. O. (1995). Public goods: a survey of experimental research. In J. H. Kagel & A. E. Roth (Eds.) *The Handbook of Experimental Economics*. Princeton N.J.: Princeton University Press. pp. 111-194.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian Cognitive Modeling: A Practical Course.* Cambridge University Press.

Lichbach, M. I. (1992). The repeated public goods game: A solution using tit for tat and the Lindahl point. *Theory and Decision, 32,* 133-146.

Masel, J. (2007). A Bayesian model of quasi-magical thinking can explain observed cooperation in the public good game. *Journal of Economic Behavior and Organization, 64,* 216-231.

McClinktock, C. G. (1972). Social motivation: A set of propositions. *Behavioral Science, 17*(5), 438-455.

Messick, D. M., & Sentis, K. P. (1985). Estimating social and non-social utility functions from ordinal data. *European Journal of Social Psychology, 15*(4), 389-399.

Miettunen, J., Nordström, T., Kaakinen, M., & Ahmed, A. O. (2016). Latent variable mixture modeling in psychiatric research: a review and application. *Psychological Medicine, 46*(3), 457-467.

Nagatsu, M., Larsen, K., Karabegovic, M., Székely, M., Mønster, D., & Michael, J. (2018). Making good cider out of bad apples – Signaling expectations boosts cooperation among would-be free riders. *Judgment and Decision Making, 13*(1), 137-149.

Nishi, A., Christakis, N. A., & Rand, D. G. (2017). Cooperation, decision time, and culture: Online experiments with American and Indian participants. *PLoS One, 12*(2).

Martyn Plummer (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20–22, Vienna, Austria.

Roth, A., & Erev, I. (1995). Learning in extensive form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior, 8*, 164-212.

Segal, U., & Sobel, J. (2007). Tit for tat: Foundations of preferences for reciprocity in strategic setting. *Journal of Economic Theory, 136*, 197-216.

Skatova, A., & Ferguson, E. (2011). What makes people cooperate? Individual differences in BAS/BIS predict strategic reciprocation in a public goods game. *Personality and Individual Differences, 51*, 237-241.

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

Su, Y.-S., & Masano, Y. (2012). R2jags: A Package for Running jags from R (Version R package version 0.03-08) [Computer software]. *http://CRAN. R-project.*

Utz, S., Muscanell, N., & Göritz, A. S. (2014). Give, match, or take: A new personality construct predicts resource and information sharing. *Personality and Individual Differences, 70*, 11-16.

Volk, S., Thöni, C., & Ruigrok, W. (2011). Personality, personal values and cooperation predict preferences in public goods games: A longitudinal study. *Personality and Individual Differences, 50*, 810-815.

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60*, 158-189.

Wilson, R. C., & Collins, A. G. E. (2019). Tensimple rules for the computational modeling of behavioral data. *eLife*.

Xygalatas, D., Mitkidis, P., Fischer, R., Reddish, P., Skewes, J. C., Geertz, A. W., Roepstorff, A., & Bulbulia, J. (2013). Extreme Rituals Promote Prosociality. *Psychological Science*, *24*(8), 1602-1605.

Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, *6*, 299-310.