CONFORMITY, CONTEXT, SELF-IMAGE: A MULTIFACETED STUDY OF SOCIAL ATTITUDES IN DECISION MAKING

A DISSERTATION
SUBMITTED TO THE DOCTORAL SCHOOL
IN COGNITIVE AND BRAIN SCIENCES
OF THE UNIVERSITY OF TRENTO
BY

Folco Panizza

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Tutor: Giorgio Coricelli

Advisor: Alexander Vostroknutov

June 2020

Acknowledgments

I would like to thank my family for their unconditioned support, even from a long distance. My mother Giovanna for her dedication and attention, my father Paolo for his care and curiosity, and my sister Alina for her ability to listen and always give good advice. I thank my girlfriend Virginia for the certainty she gives me with her presence in my ups and downs, for her way of looking at things head on, and her efforts to sort things out whenever I feel lost.

I thank my lifelong friends, Tiziano, Giulia, Veruska, Matteo, Alessio, Alessandra, and especially Andrea and Francesco: you have been of comfort and relief even at the most unsuspected times. A big thank you to all the friends I had the privilege to meet in Trento: Daniele, Dmytro, Maria, Giorgio, Nicholas, Lisa, Anna Maria, Gaia: life would have been much less interesting if you had not pushed me to discover how much Trentino has to offer. A special thanks goes to my fellow players, Rachele, Alessandro, Lorenzo, Daniela, Nicholas. It has always been a lot of fun, and I am sure it will be again, no matter how far away we are.

I thank my flatmates, and in general all those with whom I lived under the same roof, like Valeria and Fabiano: I thank you for your infinite patience, but most importantly for your company, it is a pleasure to live together with someone you feel attuned to. I also thank my psychologist, Dr. Tricoli, for the new ways he gave me to look at myself and the world.

I am grateful to my colleagues who have accompanied me in these years for a short or long time: Nadège, Benjamin, Luca, Joshua, Tobias, Doris. You have been of great help to me with your ingenuity and insight, but I thank you also for the joviality that you have shown me both inside and outside the laboratory. I thank my tutor, Giorgio Coricelli, for giving me this unique opportunity to learn and test my skills while doing things that I am passionate about. Most importantly, I thank my advisor, Alexander Vostroknutov: without his presence, perseverance, and attention, this work simply would not have existed. Thank you, Sasha.

Abstract

Social attitude is the approach of a person displayed towards other individuals or groups. Social attitude comprehensively affects the way we perceive, behave in, and interact with, the surrounding world; it is simply not possible to understand complex social behaviour such as strategic thinking without first knowing the attitude of the parties involved. Several disciplines contribute to the complex study of social attitude (social preferences in economics, social value orientation in psychology), but only recently have these disciplines started to communicate and develop comprehensive definitions and models. In particular, the current research debate focuses on pinpointing the nature of social attitude (e.g., what its defining components are), the factors that influence it (e.g. context, other individuals), as well as its consequences (e.g., its relevance for self-image representation). This thesis aims to answer to some of the open questions in the literature by testing and comparing the proposed competing explanations. The studies presented are based on a series of behavioural experiments coupled with established but also newly developed measurement tools concerning social norms and personal preferences. In addition, we try to uncover the mental processes underlying decisions with the help of computational models. The thesis is structured as follows. In Chapter 1, We outline a brief summary of the theories on social attitude from the economic and psychological literature, and describe the main tasks and models employed in the thesis. Chapter 2 explores how social attitude is influenced by others' behaviour. We conduct a systematic comparison of the possible mechanisms driving attitude conformity using various experimental conditions, computational models, and control tasks (e.g., norm elicitation). We find that participants conform due to both peer influence (by learning from others about how salient a norm is) and compliance to authority (i.e. experimenter demand effects). Chapter 3 studies the effect of context in a task eliciting social attitude. We specifically test the effect of unavailable choices, that we call "meta-context", on participant's decisions. We find that participants' concerns about social norms, as well as their choices, depend on the currently available options, but also on meta-context. In Chapter 4, we study whether individuals tend to selectively forget about their morally questionable choices, and information related to it, such as the context in which the choice was made. We find that participants recollect less correctly selfish or anti-social choices compared to pro-social ones, but we find no memory bias concerning the context of the choice. Moreover, we uncover some potential evidence of a second memory bias related to choice frequency: people are generally more pro-social than antisocial, which means antisocial choices are more rare and thus more difficult to remember correctly. Finally, in chapter 5 We summarise the main findings of the thesis and present some conclusions. We try to integrate the various results to propose an empirically-informed model of social attitude to be applied in future research on the topic.

Contents

List of Tables

List of Figures

List of Boxes

1 Theoretical and Methodological Introduction

1.1 Social Attitude

Some people are willing to incur great personal costs and to take risks in order to help–or, conversely, to hurt–strangers, even without any apparent personal benefit. Some others might not even take pleasure in their actions, but rather may not feel at peace with their conscience if they do not commit themselves to this endeavour, whether it is sacrificing one's own life to assist the most in need (MacFarquhar, 2016), or organising the deportation and extermination of entire ethnic groups (Arendt, 2006). Such selfless interest in strangers' well- or ill-being has particularly puzzled social scientists and in particular economists, who for a long time have theorised that individuals' behaviour is driven by self-interest alone (Henrich et al., 2005). Researchers refer to the perplexing motives behind these behaviours as social preferences (Fehr and Krajbich, 2014) or social value orientation (Murphy and Ackermann, 2014). Given the broad and differentiated range of mechanisms that fall under these appellations, in this thesis we prefer to call them more generically with the term *social attitudes*.

The aim of the present thesis is to explore the relation between social attitude and other relevant aspects of social decisions: are we influenced by others' social attitude, and if yes, what drives this influence? (Chapter 2); if a social action is possible in principle but not in practice, will it still influence the way one ultimately chooses to behave? (Chapter 3); Does the need to protect self-image distort the memory of selfish or antisocial actions? (Chapter 4). These and other questions we ask in this thesis try to address just as many limits of the research on social attitude.

While we leave the discussion of each specific question to the respective chapters, here we introduce some common theoretical and methodological grounds on

which our studies were built. Notions and tools related to the study of social attitudes have evolved separately in various disciplines; nevertheless, they share striking similarities if not overlaps. Thus, in the following sections we will try to organise and integrate these contributions to delineate the development of the main experimental tasks and cognitive models of social attitude that we employ and attempt to improve in this thesis.

1.2 The Evolution of Social Attitude

The study of social attitude traces back to economics (Geanakoplos et al., 1989; Bolton, 1991; Rabin, 1993), psychology (Messick and McClintock, 1968; Griesinger and Livingston Jr, 1973), as well as sociology (Sawyer, 1966). Until recently (see for instance Murphy and Ackermann, 2014), this research developed mostly in parallel, without much intersection of contributions from other fields. As an example of this, consider the very definition of social attitude. In economics, social attitude has been linked to preferences. Economic preference is understood as the way a good (or an action) is valued relative to possible alternatives (Richter, 2008). In this sense, displaying some social attitude reveals that one does not only value a good or an action based on personal gains or losses, but also based on those of other individuals. Coincidentally, psychologists developed the concept of social value orientation (SVO, McClintock and Allison, 1989; van Lange, 1999), which is almost identically defined as a person's preference on how to allocate resources among oneself and others.

The Dictator Game. Beyond the similarity in definitions, economics and psychology rely on similar tasks to study how these social 'preferences' influence people's choices: economic games. These games generally consist of players choosing how to distribute some resource between themselves and other participants (Camerer,

2011), the most common variation of these games being the Dictator Game (DG for short, Kahneman et al., 1986b; Forsythe et al., 1994, Figure 1.1). In the Dictator Game, one player, called the dictator, is endowed by the experimenter with a sum of resources (usually money), and can choose whether to allocate part of it to another unknown player, the recipient. Two relevant features of this game are that the recipient cannot react to the dictator's decisions, and that identity of players is never revealed. Under these conditions, any non-zero allocation to the other player has been taken as evidence of a positive social preference that is traded off against monetary self-interest.

Choice-based social attitude. The plainest conceptualisation of social preference assumes that one's attitude depends on the willingness to increase or decrease the other person(s)' wealth. To express the subjective value of an outcome, which economists refer to as utility U, we thus consider the personal returns of the dictator (d) and of the recipient (r) (1, Griesinger and Livingston Jr, 1973; Liebrand, 1984; Loewenstein et al., 1989; Murphy et al., 2011)[1]:

$$U(\pi_d, \pi_r) = \pi_d + \omega \cdot \pi_r, \tag{1}$$

Where U is the dictator's utility of the allocation as a function of one's own (π_d) and the recipient's (π_r) gains, and ω is the weight of the recipient's payoff relative to one's own gains[2]. This formulation possesses intuitive properties that make social attitude easily understandable and comparable across individuals: when ω is

[1]To represent social attitude, here and in the next formulae we will consider the case of the Dictator Game with only two agents, although most of the models have been extended to multiple agent situations.

[2]Other choice-based models use different utility specifications. Team reasoning theories (Sugden, 2000), for instance, formulate utility as the sum of payoffs of each player and regardless of how they are distributed, as if players were part of a single team. Similarly, the model of *homo moralis* (Alger and Weibull, 2013) theorises a strategy where participants try to maximise the payoffs across players.

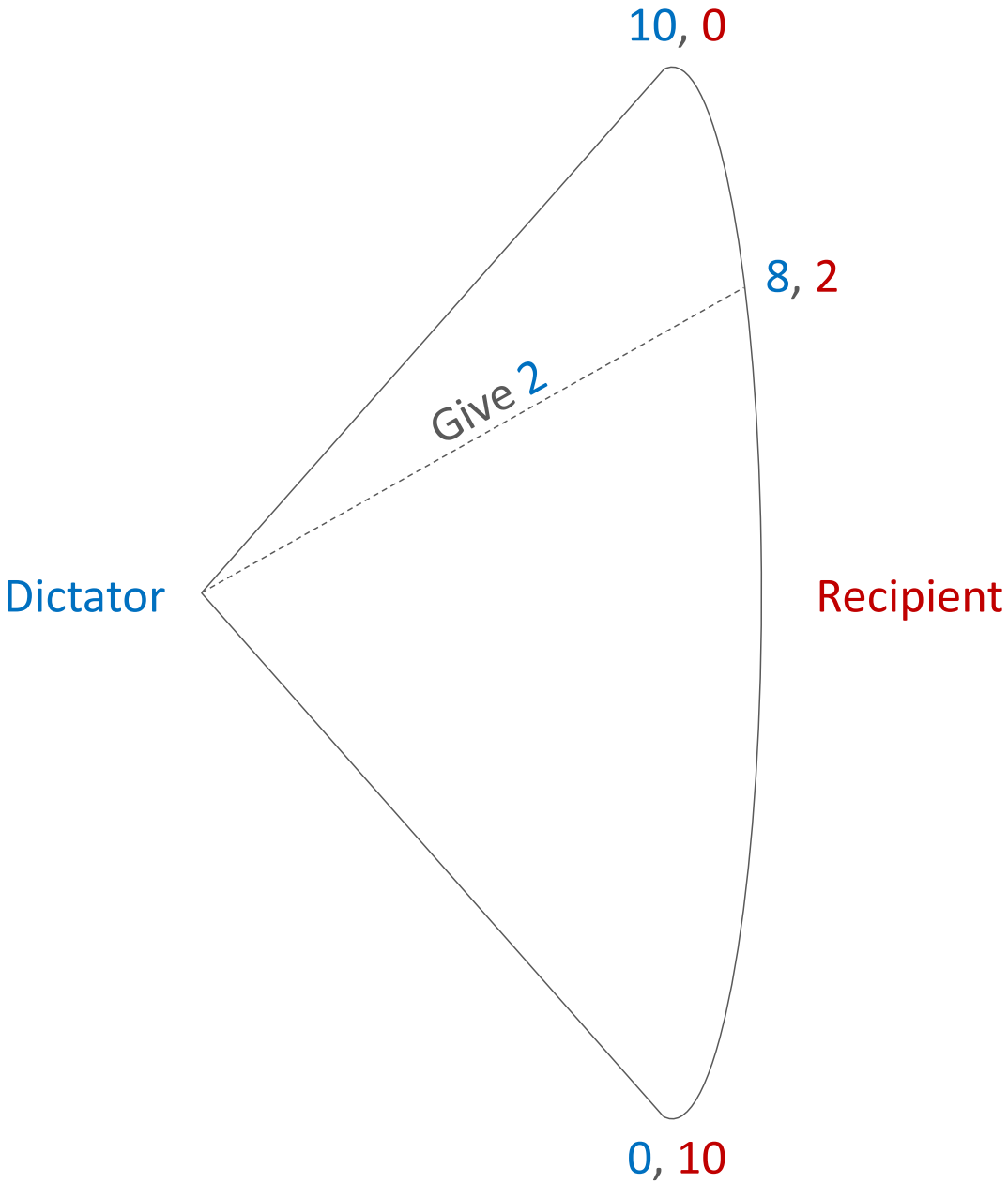

Figure 1.1: Example of a Dictator Game. Dictator's choices may vary from giving nothing (0) to giving everything (10) to the other. In this case, the dictator chooses to send 2 tokens to the recipient.

positive (negative), the higher (smaller) amount for the other the better, and that the greater the value of ω, the greater the weight on the other's gains. Simplicity however comes with several limitations, one of the most prominent being that ω

does not take into account relative gains; in other words, the utility of a dictator is always proportional to an increase (decrease) of the recipient's money, no matter what the final resource distribution between players is. Such indifference to outcomes limits the adoption of this model in settings when payoffs of participants remain largely comparable. We adopt a variation of choice-based social attitude ($\omega = \tan\alpha$) in Chapter 2, where this condition is met.

Outcome-based social attitude. To express the idea that social attitude may depend not only on the initial choice, but also on the final state of agents, researchers developed several theories linking social preferences to the outcome of players. One of the most prominent families of outcome-based models is developed around the notion of inequity aversion. Inequity aversion posits that more equitable distributions of resources are generally preferred to less equitable ones, and that agents can have different preferences based on whether they end up with more or less resources than the other player. The simplest formulation of an inequity aversion model takes the following form (2, Fehr and Schmidt, 1999; Charness and Rabin, 2002, but see Bolton and Ockenfels (2000) for an exception):

$$U(\pi_d, \pi_r) = \begin{cases} \pi_d + \omega \cdot \pi_r, \text{ if } \pi_d \geq \pi_r \\ \pi_d + \beta \cdot \pi_r, \text{ if } \pi_d < \pi_r \end{cases}, \tag{2}$$

Where ω and β represent the weight on the recipient's payoff conditional on the relative gain. More complex formulations add other properties to social attitude, such as accounting for how increasing one of the two payoffs affects the value of the other payoff (constant elasticity of substitution, Andreoni and Miller, 2002; Falk and Fischbacher, 2006; Andreozzi et al., 2013).

The classic Dictator Game requires some design improvement to be able to test outcome-based theories, and variants of the game as the ones developed in the

SVO literature serve this purpose. Some of these Dictator variations allow participants to choose between allocations differing in the total sum of resources (i.e. with different efficiency: for instance 10/0 and 6/6; Andreoni and Miller, 2002). In another version of the task, participants are forced to choose among a discrete set of allocations, rather than the whole range; versions of this task are called mini-DG (Figure 1.2, see for instance Engelmann and Strobel, 2004).

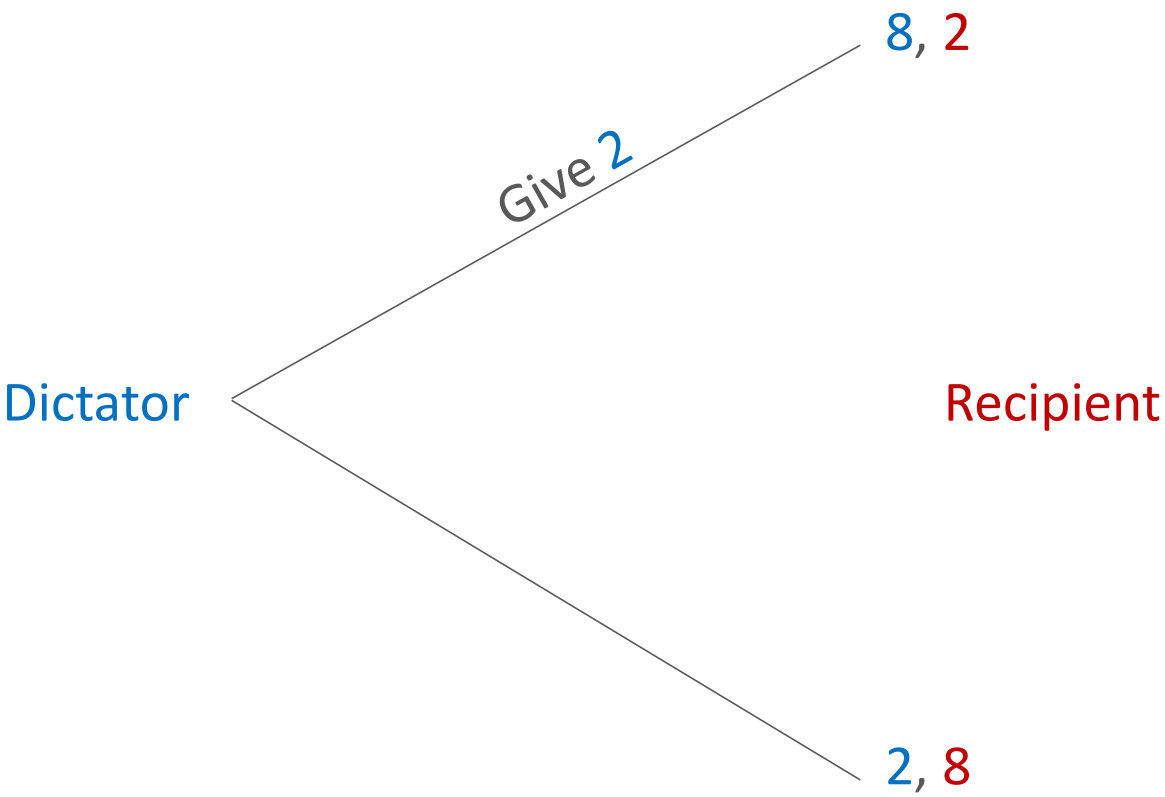

Figure 1.2: Example of a mini Dictator Game. Dictator's choices are limited to a subset of all possible resource distributions, generally two or three.

Belief-based social attitude. Another class of social preference models expands on previous theories by incorporating beliefs about the recipient's intentions or expectations into the utility of the dictator. Beliefs build upon concepts such as reciprocity (Rabin, 1993), or emotions like guilt (Battigalli and Dufwenberg, 2007), shame (Tadelis, 2011), anger or frustration (Battigalli et al., 2019). Belief-based models rely on the framework of psychological games (Geanakoplos et al., 1989;

Box 1.1: Neural correlates of the Dictator Game

Given the widespread interest in economic social preferences, a number of functional magnetic resonance imaging (fMRI) studies has tried to capture their underlying brain processes. This research has mostly employed Dictator Game variations focusing on the role of the dictator at the time of decision; findings have highlighted a brain network associated with the expression of social attitude, including prefrontal areas (ventromedial, vmPFC; dorsomedial, dmPFC; dorsolateral, dlPFC; anterior cingulate, ACC), the striatum, the amygdala, and the anterior Insula (AI).

Early work focused on charity donations as a framework to explore propensity to give. In these experiments, participants were endowed with a sum of money and decided whether to donate some of it to a charitable organisation. Many of these studies evidenced the role of the ventral striatum–an area involved in reward processing–in giving behaviour: ventral striatum activity was stronger in free compared to mandatory donations (Harbaugh et al., 2007; Hare et al., 2010), or even compared to when the participant herself received similar rewards (Moll et al., 2006). Harbaugh et al. (2007) additionally found that, across participants, striatal activation during free donations to charity–as compared to activation during personal gains–positively correlated with the amount donated. However, such association between ventral striatum and donations was not confirmed in a later study by Hare et al. (2010); instead, the authors found that subject's donation size correlated with activity in their vmPFC, an area involved in value computation. This vmPFC activity suggests that evaluation of needs of others (such as a charity) is processed in the same areas that compute value for personal non-social goods (such as personal earnings). vmPFC activity further correlated with a self-report questionnaire about engagement in voluntary activities (Moll et al., 2006); social value computations in the vmPFC have therefore been associated also with real-life engagement in pro-social activities.

More recent work on charitable giving has focused on separating neural correlates of donation choice and evaluation of the outcome (Kuss et al., 2013): in fact, when the outcome of a donation is certain, both events (decision to donate and outcome for the recipient) are thought to be processed nearly simultaneously in the brain. To study separately decision and outcome, the authors introduced a 50% chance that the selected choice in the charitable game was unsuccessful. Through this manipulation, it was found that the expected reward for charity minus the actual outcome (a measure called the reward prediction error) correlated with activity in the Nucleus Accumbens (NAc), one of the nuclei of the ventral striatum. In line with the findings by Harbaugh and colleagues, NAc activation across participants correlated to their

proportion of costly donations, which in turn was associated with the frequency of real-life prosocial activities as measured by a self-report questionnaire.
The same team of researchers tried to study the role of the reward prediction error using actual Dictator games with other participants rather than charity donations (Kuss et al., 2015); the study however failed to replicate the previous findings, leading the authors to speculate that, in a more direct social setting such as the dictator/recipient relation, the rewarding experience is locked to the choice rather than to the outcome. No evidence has been brought however to support this hypothesis, suggesting caution with the analogy between charitable donations and dictator donations. That being said, the fMRI study on Dictator Games yielded two additional noteworthy findings. Firstly, vmPFC and dmPFC activity was stronger during non-costly prosocial choices as compared to self-interested choices. The joint activation of these two areas has been interpreted as an engagement of reflective cognitive resources (dmPFC) together with integration of value of personal and receiver payoffs (vmPFC). Secondly, participants classified as selfish (according to a social value orientation questionnaire) showed additional vmPFC activation compared to prosocial individuals in the same contrast (non-costly prosocial $>$ self-interested), suggesting that making a prosocial choice for 'selfish' participants was more cognitive demanding than for prosocials.
Other fMRI studies on Dictator Games have specifically focused on the choice process. Zaki et al. (2011), employed a variation of the game in which the dictator had to choose between allocations where only the dictator or the reicipient, but not both, could earn money. Orbitofrontal cortex (an area within the vmPFC) was more activated when participants chose the highest payoff (irrespective of who received it). Across subjects, the more often a dictator chose the highest payoff (regardless of who received it), the smaller the activation was in the Anterior Insula (AI), a region associated with increased awareness. Güroğlu et al. (2014) also explored the Dictator Game choice context, but with an interest on asymmetric endowments. Participants in the experiment could choose between an equal split of money and an unequal split favouring either oneself or to the recipient, two alternative forms of inequity (advantageous and disadvantageous) theorised by the inequity aversion models. When choosing the unequal option (irrespective of who was favoured) several frontal regions where more active than when choosing the equal split, including the dlPFC, AI, and the dorsal ACC, a network of areas that has been previously associated with responses to unfairness. When focusing on decisions that involved an advantageous choice for the other (as compared to equitable choices), there was an additional activation of the right AI, and (in accordance with the charity donation

literature) vmPFC and ventral striatum, suggesting that advantageous and disadvantageous splits are processed engaging different cognitive mechanisms.
Neuroimaging research has also tried to test directly social attitude models such as inequity aversion. Tricomi et al. (2010) searched for neural correlates of inequity aversion by assigning different amounts of money to different participants and by then making them rate how appealing were additional money transfers to either themselves or others. Participants with an originally smaller endowment displayed a greater activity in the ventral striatum and vmPFC when observing a potential transfer to oneself as compared to when observing a potential transfer to the other, richer player. This evidence has been brought to support the prediction of inequity aversion models that people prefer equitable distributions of resources. Along the same lines, Haruno and Frith (2010) explored whether aversion to unequal payoffs correlated with brain areas. Inequity was defined as the magnitude of payoff difference in the linear regression (3):

$$U(\pi_d, \pi_r) = \beta_1 \cdot \pi_d + \beta_2 \cdot \pi_r + \beta_3 \cdot |\pi_d - \pi_r| + \varepsilon, \tag{3}$$

The β_3 weight correlated with brain activity in the amygdala for participants who were classified as prosocials (according to a social value orientation task) but did not in participants classified as individualists. The authors hypothesised that activation of the amygdala, a set of nuclei associated with negative emotional responses, could reflect a rapid intuitive response. This intuitiveness hypothesis was corroborated by a behavioural replication of the experiment: participants played the task under cognitive load to prevent deliberate thinking, and whereas the group of individualists did behave differently under this condition, no change was observed in the behaviour of prosocial participants.
Finally, one recent neuroscientific study using Dictator games has tried to integrate response time of dictators' decisions to study the brain correlates of generous choices (Hutcherson et al., 2015). In the task, participants played a series of mini-DG where they chose between two alternative allocations. One of the two alternatives was identical across all trials (default prize), while the other varied (proposed prize). Crucially, the authors theorised that the decision process of the dictator could be represented as a continuous comparison in time between two alternatives (4):

$$RV_t = RV_{t-1} + \beta_1 \cdot (\pi_d^2 - \pi_d^1) + \beta_2 \cdot (\pi_r^2 - \pi_r^1) + \varepsilon_t, \tag{4}$$

Where RV_t is the relative difference of values at time t, and 1 and 2 indicate the two alternative allocations presented in each trial. The model predicts that a choice is

instantiated once enough relative value difference is 'accumulated' in favour of one of the allocations. The authors regressed the proposed prize amounts for oneself and for the other player against neural activity. Several regions involved in different stages of social decision-making correlated with the amount for oneself (ventral striatum, vmPFC), and with the amount for the other (temporoparietal junction, precuneus, vmPFC). Notably, a conjunction analysis revealed that a portion of the vmPFC was associated with both the amount for oneself and for the other. In line with previous results (Hare et al., 2010; Kuss et al., 2015), this finding was taken as evidence that vmPFC could be involved in the integration of different value signals. In addition, with the use of the computational model it was possible to confirm that the rate of generous choices was greater in participants with lower weight for personal gains β_1 and higher weight for the other's gains β_2.

Despite the multiplicity of frameworks and models trying to capture social preferences, there appear to be numerous converging findings. Both the ventral striatum and the ventromedial PFC have been associated with the execution of a generous (or equitable, see Tricomi et al., 2010) actions, and have been shown to discriminate across subjects in their prosocial behaviour. These findings suggest that these highly interconnected areas play an defining role in computing the social value in decision-making. Ventromedial PFC, together with Dorsomedial areas, has also been linked to socially reflective/strategic behaviour, contrasting with amygdala activation that was found instead for supposedly automatic responses. These findings fit in the narrative that social attitude could require different cognitive resources depending on the decision context. Finally, several findings point at vmPFC, and more specifically at the OFC, as a general hub for integrating value signals such as personal gains together with rewards for others. While this hypothesis has been supported under various decision domains, these results contribute by proposing that such mechanism is also common to social choices (Ruff and Fehr, 2014, **but see Ugazio et al. (2019)**).

We need to stress however that many of the neuroimaging results presented base their analyses on limited sample sizes, raising doubts about their generalisability. Güroğlu et al. (2014), for instance, based one of their analyses on just 10 subjects; Harbaugh et al. (2007), as well as Haruno and Frith (2010), and Kuss et al. (2013) chose to split their sample for some analyses, using data from only 9, 16 and 14 subjects, respectively; Zaki and Mitchell (2011) ran a linear regression based on only 19 data points. Some of these analyses currently would be considered outdated, if not outright unreliable. To solve this concerning issue, neuroscience researchers have proposed to increase the minimum sample size, also by means of cooperation

through different institutions (Button et al., 2013). An additional effort that could help resolving the problem of under-powered designs is the employment of new techniques, such as multivariate analyses (e.g., Representational Similarity Analysis, Kriegeskorte et al., 2008; Kriegeskorte and Kievit, 2013) that could reduce the need for sample splitting, promoting instead a classification of participants' attitude on a continuum (van Baar et al., 2019, is an example), as suggested by most of the utility models that we covered.

Battigalli and Dufwenberg, 2009, 2019), namely that the very representation of payoffs in a game like the Dictator is affected by the players' beliefs. To represent the utility of a belief-based model, let us consider a mini-dictator game where the dictator feels guilt about not meeting the recipient's expectations (5, Attanasi and Boun My, 2016; Battigalli and Dufwenberg, 2009, 2007):

$$U(\pi_d^1, \pi_r^1, \pi_r^2) = \pi_d^1 - \theta \cdot \eta \cdot \max\{\pi_r^2 - \pi_r^1, 0\}, \tag{5}$$

Where π_d and π_r are respectively the amount of points for the dictator and for the recipient in the two allocations, indicated by the superscript 1 and 2. Parameter θ represents the dictator's sensitivity to guilt (i.e., how much she cares about the recipient's expectations), whereas η indicates the dictator's second-order belief about the probability that the recipient expects to receive the higher π_r of the two allocations (i.e., according to the dictator, the recipient expects to receive the higher payoff with probability η).

Ties-based social attitude. Two central assumptions of the models that we discussed so far are that social preferences are not affected by a) the identity of the other player nor by b) past interactions. Ties-based models take a different approach by tailoring social preferences on the specific features of the other player (Attanasi et al., 2014, 2016; Van Winden et al., 2008; Levine, 1998). Ties depend on the quality and quantity of characteristics dictator and recipient have in com-

mon[3], as well as on the affective history between the two individuals (Van Winden et al., 2008). Ties vary in strength and can be either positive or negative, such that we may display a strongly positive social preference towards a close friend, and a mildly negative social preference towards a person who just jumped the queue we are in. Attanasi et al. (2014) provide us with an example of ties-based utility (6):

$$U(\pi_d, \pi_r) = (1 - \lambda)\pi_d + \lambda(\pi_d + \pi_r), \tag{6}$$

Where λ represents the tie of the dictator with the recipient. Importantly, this formulation is different from choice-based utilities in that the weight λ not only is specific to the given recipient, but can also change with time depending on the interactions the dictator has with her.

Belief-based and ties-based models try to overcome other limits of choice-based and outcome-based models, namely that the identity of the recipient and her beliefs are not taken into account. Rather, dictators and receivers might have differing needs with regard to the resource to be allocated. These models critically introduce the idea that economic preferences can depend on both the social and non-social context of the interaction.

1.3 Limits of social preferences

A challenging limitation of economic social preferences is the scarce evidence demonstrating their stability through time. Studies in economics have tested the temporal stability of preferences, but with mixed results. Blanco et al. (2014) reported that inequity aversion is relatively stable at a sample level, but not at an individual level when participants play a series of distinct economic games. Brosig et al. (2007) and Nax et al. (2015) found that participants playing multiple times the Dictator and

[3]Such characteristics should define the players' social identity (as in Tajfel et al., 1979).

other economic games generally increase the proportion of selfish responses the more games they play. Volk et al. (2012) did find some degree of preference stability at the individual level: participants played three consecutive Public Goods games (Kahneman et al., 1986a), and their behaviour in the games was grouped into three categories (conditional cooperators, free-riders, others). Based on this categorisation, half participants in the study displayed a stable pattern of behaviour across the three games. Field studies (De Oliveira et al., 2012; Carlsson et al., 2014) found evidence that donation behaviour in the laboratory predicts real-life donations in a considerable portion of participants, even after long spans of time. Personality traits like agreeableness and justice sensitivity seem also to correlate with consistent patterns of behaviour in social settings (Volk et al., 2012; Lotz et al., 2011, 2013; Fetchenhauer and Huang, 2004). Findings on the field and personality studies suggest that some individuals could be more resilient than others to situational variations, but do not exclude the influence of context and time on the preferences of others.

Perhaps the most compelling support for preference stability comes from work in psychology on social value orientation. van Lange (1999), for instance, measured participants' preferences in a series of Dictator games, and found temporal consistency between repetitions of the game. A general review of work in this field, however, acknowledges that research on stability is lacking and produces conflicting results (Bogaert et al., 2008). Furthermore, literature on social value orientation generally conceives social attitudes as a substantive trait of personality, classifiable in predefined categories. This tendency to categorise attitude has led until recently (see Murphy et al., 2011) to favour the use of simplified measures at the expense of more fine-grained estimates. Since categories could miss noticeable temporal variations in subjects, they set a limit on the generalisation of these results. Most relevantly, social attitude estimation in SVO research generally employed a fixed

sequence of Dictator Games for all repeated measures, meaning that temporal stability could alternatively be explained by familiarity with the task, since familiarity in turn could provide participants with cues on how to behave. Task sensitivity is an issue known in other types of economic preferences such as risk aversion, where it has been shown that even apparently similar tasks yield contrasting estimates of participants' risk attitude (Pedroni et al., 2017; Zhou and Hey, 2018). Variability in estimation is particularly concerning in Dictator Games, since they have gone through several design changes across the literature (Table 1.1).

Resource Endowment	windfall		earned
Dictator	complete anonymity		wiggle room
Recipient	unknown		interaction history
Choice Set	give		take
Allocations	fixed sum (efficiency)		half-half split
Choice	continuous	2/3 allocations (mini-DG)	default allocation

Table 1.1: Some examples of how features of the Dictator Game can vary.

Indeed, small modifications to the experimental setting or even seemingly irrelevant information seem to play a significant role in the way participants choose in a Dictator Game. A first feature that impacts on dictators' decisions is the way in which they have been endowed with resources (Chlaß and Moffatt, 2012; Engel, 2011). In the most typical form of a Dictator Game, the dictator receives resources for free, making the decision to give to the recipient relatively effortless; if dictators have to complete some demanding task to earn their resources, then giving is strongly reduced (Cherry et al., 2002). Cost of giving could also be influenced by the set of options from which the dictator can choose (List, 2007). If for instance dictators are given the additional possibility of taking resources from the recipient (a so-called give-take Dictator Game), then not only dictators give less, but some do actually detract resources from the recipient (List, 2007; Bardsley, 2007; Cappelen et al., 2013). Dictators's behaviour depends also on the traceability of their

actions: donations drop in experiments where the experimenter is also unaware of participants' identities (i.e. double blind paradigms) (Franzen and Pointner, 2012a; Hoffman et al., 1996), whereas cues of being observed, even if extremely subtle–such as eye spots drawn in front of the dictator's work station–appear to increase giving (Haley and Fessler, 2005).

It seems clear then that the social attitude measured in Dictator Games does not necessarily reflect inherent social preferences of people, as it is assumed to some degree by most preference models presented so far. A number of researchers have even proposed that the Dictator Game setting is too artificial and vacuum-like to infer any meaningful information about individuals (Smith, 2010; Oechssler, 2010; Chlaß and Moffatt, 2012): participants may be confused despite–or even because of–the task simplicity, making participants dependent on contextual cues. Conversely, however, increasing the "naturalness" or recognisability of the game does not necessarily reduce the risk of misinterpretation by participants, who instead may be more opinionated in their behaviour (Jimenez-Buedo and Guala, 2016). Furthermore, claims of artificiality do not take into account real-life situations that do resemble the most classic version of the Dictator Game, as for example when the resource to be shared is time (e.g., voluntary work).

1.4 Norm-based social attitude

If social attitude is not a direct expression of some personal tendency, what can be inferred from behaviour in tasks like the Dictator Game? Participants' interest in interpreting contextual cues gives a hint about what might be actually understood. According to new research in experimental economics (Kimbrough and Vostroknutov, 2016), context informs participants about what rules are more likely to apply in the specific situation, who else will be informed about their actions, and what roles would these other people play. As Jimenez-Buedo and Guala (2016) el-

egantly put it, it appears that individuals "care about others' opinions, more than (or in addition to) others' welfare". Belief-based social preferences partly address this problem by factoring in the beliefs about the recipient, but the impact of this information is limited compared to the general social setting.

One prominent theory that has been proposed recently to account for contextual information is that social attitude may be driven by social norms, defined as the collectively shared beliefs on the appropriate behaviour in a specific situation. Norm-based models (López-Pérez, 2008; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2015) posit that people do not gain utility from giving (or not) to the other person, but rather from abiding (or not) by the expected norm. Several studies have indeed shown how appropriateness concerns play a role in economic games like the DG (López-Pérez, 2008; Krupka and Weber, 2009, 2013; Kimbrough and Vostroknutov, 2015, 2016, 2018b)[4]. In theory, norm-based models can be expressed with the formula (7):

$$U(\pi_d, \pi_r) = \pi_d + \gamma \cdot N(\pi_d, \pi_r), \tag{7}$$

Where $N(\pi_d, \pi_r)$ is the normative evaluation of the allocation, and γ is the person's sensitivity to following the expected social behaviour. A variation of norm-based utility is presented in Chapter 3.

Whereas norm-based models fill a problematic gap in the study of social attitude by providing an explanation to context-derived variability in decisions, they also retain several limitations. In contrast to social preference models, norm-based models disregard any recipient-related utility, even though it could still affect social attitude in several circumstances (e.g., when agents have a history of interac-

[4]We notice that Bernheim (1994) has proposed a similar approach using a belief-based model. The difference from Bernheim's approach that we want to highlight here is the relevance of context in defining the appropriateness of an action: *ceteris paribus*, the same behaviour can be seen as appropriate or inappropriate depending purely on the specific situation in which it is enacted.

tions). A complex but possibly successful approach to this dichotomy would be to integrate norm-based and belief-based models of social attitude; we elaborate on the potential ramifications of this approach in the general discussion of this thesis (Chapter 5). A second complication of norm-based models is the measurability of norm-related components, such as the definition of the norm function or the individual sensitivity to norm-following. The proposed solutions that we use in the present work, presented in Chapters 2 and 3 (Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2018b), are still limited in their applicability and in the precision of their estimates.

1.5 Methodological Contributions

In the last section of this theoretical introduction, we briefly summarise the thesis' contributions to still unresolved theoretical and methodological problems related to the study of social attitudes.

Experimenter Demand. A first issue that we want to address is the relation of dictators with the experimenter in the Dictator Game. Whereas field research has partly replicated findings on social attitude (see for instance Franzen and Pointner, 2012b; Henrich et al., 2005), it is still plausible–as it has long been noted in psychology (Orne, 2009; Rosenthal and Rosnow, 1977)–that the experimenter could involuntarily influence the laboratory results. It has indeed been proposed that social attitude in tasks like the Dictator Game could be explained by a desire to fulfil experimenter's expectations rather than responding according to one's preferences or normative beliefs (Zizzo, 2009; Jimenez-Buedo and Guala, 2016). The experimenter in fact possesses relevant information about the participant's behaviour; participants seem to care about this knowledge, at least according to the above-mentioned finding that complete anonymity reduces giving.

This phenomenon has been called Experimenter Demand Effect, or EDE for short, and takes two forms (Zizzo, 2009, 2013). The first has a social character in that participants are conditioned by the experimenter's authority, and therefore they are willing to comply to his/her presumed expectations or wishing to convey a certain self-image. This social conditioning was illustrated in a study where Dictator giving was greater when the experimenter was a professor compared to when the experimenter was an unknown collaborator (Brañas-Garza, 2007). The second type of EDE, cognitive EDE, is driven instead by participants' interest to find out what the experiment is about. By trying to guess the presumedly hidden intentions of the experimenter, participants form an opinion about what they are expected to do and use this information to complete the game.

Given the pervasiveness of EDE, not only in terms of social but also of cognitive expectations, appropriate experimental controls should be in place to detect and reduce the impact of the experimenter's influence. A previous proposed method to control for EDE has been to elicit participants' tendency with the use of a separate task (Fleming and Zizzo, 2015). This method however suffers from similar limitations of other preference estimation methods, in that in independent settings, influence of experimenter demand could vary. The present method that we delineate in Chapter 2, is instead of obfuscating the goals of the experiment (Zizzo, 2009) by giving conflicting cues about the experimenter's intentions. Whereas this method does not prevent participants from having ideas about the purpose of the experiment, it allows to detect who is more prone to react to the information about the experiment goals, and therefore provides us with a way to categorise participants according to their EDE tendency *within* the task itself. An indirect but appealing result is that we are able to separate the vertical social influences caused

by the presence of an authority (Zizzo, 2013) from horizontal influences related to the social context, and hence to social norms (Guala and Mittone, 2010).

Stability of social attitude. In the previous section we discussed how evidence in support of temporal stability in social attitudes is lacking. The fact that attitude changes with time would cast concerning doubts on the generalisability of any behavioural or neural result concerning differences in social attitude. Such concern could be partly mitigated by considering contextual information in the computation of attitude, as in norm-based or belief-based models; these models however leave open the question as to whether there is a stable trait of the individual at the core of social attitude, as some scholars maintain.

One potential method to test for attitude stability is to test different theories about the choice process, that either assume or do not assume attitude stability. An example of a choice model which assumes temporal stability, and one of the most widely adopted models in cognitive science is the softmax (Sutton and Barto, 1998). Softmax choice models posit that people have a clear order of preferences over their possible actions (such as being selfish rather than altruistic), but that the probability of choosing the preferred option is not equal to 1 because of cognitive miscalculations when comparing the options available. If we apply such choice model to the simple case of a mini-Dictator Game with only two allocations possible, we get (8):

$$\Pr(1) = \frac{1}{1 + e^{-(U_1 - U_2)/\tau}}, \tag{8}$$

Where U_1 and U_2 are the utilities for the two allocations, and τ is the decision noise that should explain why participants sometimes choose the less-preferred allocation relative to their social attitude: the smaller (larger) τ is, the higher (lower) the probability of choosing consistently with one's own attitude.

To consider a class of choice models which does *not* assume temporal stability of attitude we turn to random preference models (Koppen, 2001; Regenwetter et al., 2010, 2011). Random preference models assume that preferences are not stable, but rather change continuously over time following a pre-specified probability distribution. Parameter values (defining the person's preferences) are assumed to vary according to this distribution; hence, at different times these values will push the participant to choose different alternatives. This decision process can be estimated using the cumulative density function of the probability distribution (9):

$$\Pr(1) = \text{CDF}(T, \theta), \tag{9}$$

Where CDF is the joint cumulative distribution function of the utility parameters, θ is the parameter vector defining the CDF, and the threshold T is the value of the utility parameters under which the person should be indifferent about which option to choose ($U_1 \approx U_2$). The most common adaptation of this model posits that preferences are not completely random but rather remain the expression of an underlying trait of the person (Loomes et al., 2002; Gul and Pesendorfer, 2006; Moutoussis et al., 2016; Lu and Saito, 2018, but see He et al. (2019)). To express this central tendency, these models assume the probability distribution to be normal with parameters μ expressing the tendency of the participant and σ expressing the variability in behaviour. The practicality of the normal distribution resides in its flexibility: a small σ represents a strongly consistent attitude, whereas a large σ indicates that the person tends to behave randomly, suggesting no stable underlying attitude. We adopt this version of random preference together with softmax

in Chapter 2, where we test which choice model is able to explain the attitude expression in Dictator Game-like scenarios better.

Context. The notion of context has broad connotations when studying social behaviour, and as a result it is ridden with ambiguities. We touch on this concept in two chapters of this thesis, with two different goals. In Chapter 3, we study the different scopes of context that characterise a setting like the Dictator Game. We can conceive in fact at least two levels of context in a game like the DG: a 'proper' context concerning the immediate situation, and a broader one, 'meta-context', relating to all similar but not identical scenarios, namely where some component of the task differs, such as the possible actions available. Whereas there has been some indirect evidence for meta-context in the literature (see for instance Chlaß and Moffatt, 2012; Thomsson and Vostroknutov, 2017), this hypothesis has never been tested explicitly; to create a 'meta-context' in our experiment, we employ a series of mini-DGs: while dictators choose in one of the games, they are also aware of the possible outcomes of other related games. This design choice thus allows testing whether information about unavailable choices influences dictators decisions.

In Chapter 4 we study another property of contextual information, namely its persistence in memory. Previous studies in fact suggest that, in tasks like the Dictator Game, participants show less vivid memories of their behaviour when they behaved antisocially or selfishly compared to when they behaved more altruistically (Carlson et al., 2018; Saucet and Villeval, 2019). What lacks in this research however is whether also memory of contextual information (for instance, the allocations from which the dictator chose) is affected by this bias. Thus, we test this hypothesis based on a specific memory task designed to test memory of contextual information related to Dictator Games.

1.6 Summary

Social attitude is a multifaceted concept that has attracted and still attracts the interest of researchers across social and life sciences. Its conceptualisation has gone through multiple evolutions and increasingly complex assumptions, reflecting a greater attention to its several components. Yet, the study of social attitude requires several advancements both theoretically and methodologically to face the still numerous problems in the exploration and interpretation of the behavioural and brain-related findings. In the following chapters, we will try to establish a robust framework to get a more stringent picture of this fascinating phenomenon.

2 Drivers of Conformity in Interpersonal Decision Making

2.1 Introduction

Recent years have seen a growing concern with online discourse promoting violence, such as cyber-bullying or hate speech (Malmasi and Zampieri, 2017). Increasing exposure to uncivil commenting, besides taking substantial psychological and societal toll (Mohan et al., 2017), is thought to reinforce users' political polarisation (Anderson et al., 2016), or their perception of political divide (Hwang et al., 2014). Conversely, viral trends can also lead to pro-social outcomes: learning about others' donation choices increases individuals' willingness to give to charity (Agerström et al., 2016; Nook et al., 2016). Evidence suggests that fund-raising success of charitable initiatives is predicted by how much they are shared by social network users (Bhati and McDonnell, 2019), or by how concerted the network structure is (Lacetera et al., 2016). If people's attitude becomes more charitable or more malevolent in these contexts, this is at least partly due to social conformity (Wood, 2000; Cialdini and Goldstein, 2004; Toelch and Dolan, 2015).

Insights on the cognitive mechanisms behind anti- and prosocial conformity come from the literature on attitude alignment and preference learning (Shamay-Tsoory et al., 2019; Charpentier and O'Doherty, 2018). These studies have spanned a variety of domains such as attractiveness ratings (Zaki et al., 2011), food (Templeton et al., 2016), risk preferences (Chung et al., 2015; Suzuki et al., 2016; Devaine and Daunizeau, 2017), effort (Devaine and Daunizeau, 2017), and inter-temporal decisions (Garvert et al., 2015; Moutoussis et al., 2016; Apps and Ramnani, 2017; Calluso et al., 2017; Devaine and Daunizeau, 2017). At a brain level, learning about others' attitudes or preferences appears to alter the value representation of choices

(Zaki et al., 2011; Garvert et al., 2015; Apps and Ramnani, 2017) or even reward signals (Suzuki et al., 2016), while not necessarily affecting one's private preferences (Chung et al., 2015). In addition, features such as choice variability (Moutoussis et al., 2016) or attitude extremeness (Calluso et al., 2017) seem to be significant predictors of conformity.

While this research helps untangling the cognitive and brain bases of preference conformity, it remains largely unclear why exactly people shift their attitude in the direction of others' behaviour in general, and their social attitude towards other individuals in particular. In this paper we test several competing mechanisms that were proposed as explanations of attitude conformity. We consider five competing hypotheses that could explain attitude conformity. The *time-dependence* hypothesis predicts that people change their social attitude even in the absence of any observation. Indeed, there is some evidence that during strategic interactions, participants' behaviour becomes more self-oriented with time (Nax et al., 2015). The *contagion* hypothesis (Suzuki et al., 2016) posits that attitude conformity is the result of some kind of automatic mimicry. This hypothesis predicts that conformity will occur regardless of whether the observed agent is human or non-human. The *compliance* hypothesis states that participants could change attitude due to the mere presence of an authority, in our case the experimenter (Zizzo, 2009). This hypothesis predicts that a portion of participants would change their attitude in any context where they think they are expected to, rather than actually reacting to others' behaviour. The design of the present study tries to account for this mechanism by measuring its effect and separating it from actual changes in participants' attitude. We do so by measuring a collateral phenomenon, namely we expect complying participants to display both pro- and antisocial behaviour while expressing their social attitude. It is necessary to keep in mind however, that authority compliance is not a mere confound, but a phenomenon that is analogous to conformity,

as it links attitude change to vertical influences, as opposed to peer observation. The *preference learning* hypothesis (Moutoussis et al., 2016) posits that people are unsure about what their own preferences are, but they can learn them from the behaviour of others, assuming that the agent's and observer's preferences come from a common distribution. Other people's choices can thus be used to learn how one wants to behave, rather than how one ought to behave. Since preference learning should decrease in the process of learning others' choices, a second prediction of this hypothesis is that participants' behaviour should become more consistent after learning.

Our main hypothesis, *norm learning*, states that attitude conformity stems from learning what behavior is socially appropriate or how much social appropriateness matters in a given context. We conjecture that in many real life situations there is a considerable amount of uncertainty about what constitutes a social norm or how salient it is. Furthermore, many studies have shown that people have a strong preference to follow norms conditional on others following them as well (e.g., Bicchieri, 2016; Kimbrough and Vostroknutov, 2016). Thus, observing others's behaviour should reveal either information about what others believe is "the right thing to do" or how serious they are about following the norm (Cason and Mui, 1998; Bardsley and Sausgruber, 2005). This hypothesis makes two separate predictions: that participants conform by changing their beliefs about which norms are in place (*norm uncertainty*), or that participants conform by learning how salient the norm is, rather than learning about the norm itself (*norm salience*).

To disentangle the predictions of these five hypotheses, we use a series of between-subjects experimental conditions. In all conditions participants play a resource-allocation game where in each round they choose between two money allocations to themselves and another unknown participant. Halfway through the game, participants are asked to predict and learn the choices made by another agent in the

same task. Depending on the condition, the agent is either a computer, a previous participant, or a group of previous participants (in the Baseline condition participants do not observe anyone). After the main part of the experiment, we administer other tasks that measure authority compliance and normative beliefs.

To analyse the behaviour in the resource-allocation game, we use a series of cognitive models of participants' decisions. These models link the behaviour to the mental processes associated with the different hypotheses. Testing of the competing mechanisms behind social conformity is then performed by comparing social attitudes and choice consistency before and after the manipulation phase, and using additional evidence collected after the main task. Model estimation is essential for distinguishing different sources of attitude variability, as for instance participants' own variability in behaviour and attitude changes induced by learning. We also develop a new procedure to account for authority compliance and to integrate it in the estimation of participants' decision making (Box 2.2). This procedure, which we validate by means of a separate measure (Box 2.3), provides a new way to categorise participants according to the strength of this influence.

We find that participants in the experiment do shift their social attitude towards others' behaviour. Our data corroborate two hypotheses—compliance and norm learning—and reject all others. Specifically, a minority of participants is likely susceptible to what they think the experimenter expects them to do. The choices of the rest of participants are consistent only with the norm learning hypothesis, and more specifically with norm salience. We observe in fact that normative beliefs seem not to be affected by others' choices, which brings us to the conclusion that participants do not change attitude because they learn about what norms are in place, but rather how consistently they are followed and how salient it is to follow them.

Our study contributes to understanding of social learning and social decision making by showing that the change in anti- and prosocial attitudes is mostly brought about by mechanisms related to social norms, which also include compliance to (presumed) authority expectations, and not by potential non-social influences like contagion or time-dependence. We also believe that social norms may be responsible for preference change in other types of preferences studied in the literature, though this needs to be put to a test.

2.2 Methods

Participants were recruited through the recruitment system of the Cognitive and Experimental Economics Laboratory (CEEL) at the University of Trento and contacted via e-mail. Payments were made in cash and varied depending on participants' choices. No particular exclusion criteria were defined, with the only exception that participants should not have taken part in other experiments involving a similar task. The study was pre-registered on the open science framework (osf.io/th6wp); amendments to the original protocol are presented in Supplementary Material 6.1.1. All procedures were approved by the University of Trento Ethical Committee. All participants gave their informed consent for participating in the experiment.

To determine the sample size necessary to detect a change in social attitude, we conducted a power analysis using G*Power (Faul et al., 2007) aiming to obtain .95 power, .05 α probability, and at least a small-to-medium effect size (Cohen's d = 0.35 for all tests). This effect size figure was recently proposed as a plausible mean prior for experiments in social psychology (Gronau et al., 2017). Given the directionality of our hypotheses (i.e., participants' attitudes shift towards the agent's attitude), all tests considered were one-tailed one-sample t-tests against constant, and one-tailed paired t-tests (for post-hoc pairwise comparisons between condi-

tions, uncorrected). Calculation yielded a sample size of 90 participants in order to achieve the required power across all conditions. Due to an unforeseen limit in the size of the recruitment pool however, the samples of the last two conditions in order of acquisition were smaller than this pre-specified size (74 and 66 participants)[5].

376 participants (age 22 ± 2 (S.D.), 209 females) took part in the experiment. Participants' data were excluded A) if in any part of the experiment more than 10% of the answers were missing (i.e., no answer): this measure was necessary to have enough individual data for model fitting; or B) if participants failed to predict correctly at least 17 out of the last 20 (85%) choices made by the observed agent: this criterion was set in order to exclude participants who did not correctly learn about the agent's preferences and therefore did not properly undergo through our manipulation. These criteria led to the exclusion of 32 participants. In addition, data from nine more participants had to be excluded due to technical issues, instructions misunderstanding, or failures in pre-screening[6]. Analyses were thus conducted on 335 participants, however analyses for the full sample are also present in the Supplementary Result 6.2.1.

The experiment was organised in four between-subjects experimental conditions: Baseline (N = 132, 76 females), Computer (N = 65, 37 females), Individual (N = 52, 30 females), and Group (N = 86, 41 females). The main task of the experiment consisted of three parts: choice before, prediction, and choice after. In the choice parts of the task, participants played a resource-allocation game. In the manipula-

[5]The unbalance in sample size across experimental conditions is not particularly concerning for our analyses, as the tests that we run in the results to compare conditions are non-parametric and therefore do not require the assumptions typically achieved with samples of similar size such as homoscedasticity.

[6]Technical issues: software crash; instructions misunderstanding: participants who asked for clarifications halfway through the study demonstrating clear misunderstandings of the instructions; failure in pre-screening: the recruitment system manager did not exclude from re-enrolling two participants who previously took part in a pilot version of the task.

tion phase, participants predicted and learned the choices made by another agent in the same game. The nature of the observed agent depended on the experimental condition. After the main task in the Computer, Individual, and Group conditions, participants completed the norm elicitation task designed to measure normative beliefs (Krupka and Weber, 2013). Participants earned on average €7.58 in the Baseline condition, €11.54 in the Computer condition, €11.41 in the Individual condition, and €12.12 in the Group condition. All earnings were based on participants' choices as there was no show-up fee. The duration of the experiment was around 80 minutes. Some information (see section 2.2.2) was not disclosed to participants to make manipulations effective and to avoid biassing their expectations; aside from not disclosing certain details about the manipulation, the experiment involved no deception or misinformation.

The main tests that we use to analyse the data include within-subjects comparisons (one-tailed t-tests against constant to detect attitude change) and between subject comparisons (one-way ANOVA to detect whether attitude change is different across conditions plus post-hoc pairwise comparisons). We adopt non-parametric tests (Wilcoxon signed-rank, Kruskal-Wallis) in case the data do not meet parametric assumptions. All tests are corrected for multiple comparisons using a Benjamini-Hochberg procedure, and all confidence intervals (noted within square brackets) are at 95% level. Statistical tests are conducted using base R (R Core Team, 2018) and ggstatsplot (Patil, 2018). Non-parametric statistics are log-transformed for conciseness.

2.2.1 Resource-Allocation Game

During each trial of the task, participants observed an allocation of points (1 point = 0.10€) distributed between themselves and an unknown other person. Participants were then asked whether they preferred the current allocation of points or

a default allocation (100 points to oneself, 50 points to the other person, Figure 2.1B). At the end of the experiment, participants were randomly paired, and one participant in each pair was randomly selected: one of the selected participant's decisions was randomly sampled and implemented for payment (i.e. the selected participant earned the points for herself and the non-selected participant earned the points for the other).

All allocations (default and alternatives) were drawn from the set of integer allocations closest to the circumference of radius 50 centred at $(50, 50)$ (Box 2.1). The default allocation was the option with the highest payoff for the participant (100 points) among all possible alternatives. Thus, a person with selfish attitude—who derives no utility from the points that the other gets—should refuse all alternatives. Half of the alternative allocations were more advantageous for the other player than the default allocation ("prosocial" trials), whereas the other half left the other worse off ("antisocial" trials). Both before and after the manipulation phase, participants played 101 trials (102 for Baseline condition).

To estimate attitude towards others, we assume that participants can attribute to each allocation of points a unique subjective value. Value of an allocation is computed according to equation (10):

$$V(\pi_y, \pi_o) = \pi_y + \tan(\alpha) \cdot \pi_o, \tag{10}$$

where π_y and π_o are respectively points for oneself (you) and for the other, and α represents the "social value orientation" or social attitude of the participant (Messick and McClintock, 1968; Murphy and Ackermann, 2014). The attitude defines how much and in what way the amount of points for the other plays a role in the participant's decisions; in fact, $\tan(\alpha)$ represents how much one point for the other person is worth in terms of one's own points (e.g., when $\alpha = 30°$ one point for the

BOX 2.1: TRIAL SELECTION

The use of the circumference as a way to select trials is based on two considerations. First, the circumference has been used in the previous literature as a measure of social orientation (Liebrand, 1984, is the seminal paper), and should yield comparable results. Second, many models have been adopted in the literature to describe how people value options in social decision-making (e.g., van Lange, 1999; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Falk and Fischbacher, 2006); by using allocations around a circumference, most of these models make the same predictions. Agreement in model predictions allows us to use a very simple utility model (formula (10) in the main text), with only one variable defining the attitude of the decision-maker. A simple model greatly simplifies computations and thus helps testing our cognitive predictions concerning attitude conformity and choice consistency.

Allocation options were selected as follows: first, all integer coordinates within one point tolerance from the circumference were considered (i.e. all values between a circumference of radius 49 and a circumference of radius 51); second, only points between $112.5°$ and $-112.5°$ were included for the analyses (Figure 2.1A); this range eliminated allocations that are too extreme (e.g. $(15, 15)$ or $(0, 50)$). Third, we excluded allocations with more points to oneself than the default option, because gains for the player risk to overshadow the difference in points for the other. Likewise, we excluded allocations with the same points to the other as the default option, because no information about the attitude towards others can be estimated in this case. Finally, we excluded allocations that give more than 100 or less than 0 to the other player. This procedure yielded 406 allocations in total; these were then divided in four subsets, two of 101 and two of 102 trials, all evenly distributed around the arc of the circle. The two subsets of 102 trials were used in the Baseline condition (in the choice parts of the task), whereas the remaining subsets of 101 trials were used in the other conditions. During the task, participants also observed 9 trials where there is gain for the participant (more than 100 points) and no difference for the other person (50 points, as in the default option). These latter trials, which concerned a prospective fMRI follow-up study, were not included in the analyses.

other is roughly equal to 0.58 points for the self). If α is positive (negative), then the higher (smaller) amount of points for the other makes the player better off. A participant with a positive α is said to be *prosocial*, whereas a participant with a negative α is said to be *antisocial*.

Social attitude—together with other parameters relevant to the decision—is estimated twice, for choices before (α_{before}) and choices after (α_{after}) the manipulation phase. A separate estimation allows measuring any change in attitude that ensues from the manipulation phase (Box 2.2).

2.2.2 Manipulation Phase

In the Computer, Individual, and Group conditions, after the first part of the resource-allocation game, participants were asked to predict the choices of an agent in a different set of alternative allocations (Figure 2.1C). The behaviour of the observed agents in the Computer, Individual, and Group conditions were based on real choices of participants in the Baseline condition taken from either the first or second part of the resource-allocation game. In the Individual condition, the agent was a single previous participant. In the Group condition, the agent consisted of a group of five previous participants. The choices shown to participants referred to the allocation preferred by the majority of the group. Lastly, in the Computer condition, participants were told that the agent was a computer, while computer choices were in fact the same as those of the group in the Group condition (a discussion of the information about agents that was undisclosed to participants is present in Supplementary Material 6.1.2).

Participants played 63 trials of the manipulation phase in all conditions except Baseline. Correct predictions were incentivised to ensure that participants paid attention to the task. Participants received immediate feedback after each prediction, so that they could correctly learn about the agent's attitude.

The antisocial or prosocial attitude of the observed agent (α_{obs}) was controlled experimentally unbeknownst to participants. We calibrated α_{obs} separately for each participant to make her attitude, estimated from the first part of the resource-allocation game, more socially extreme in case they conformed with the behaviour

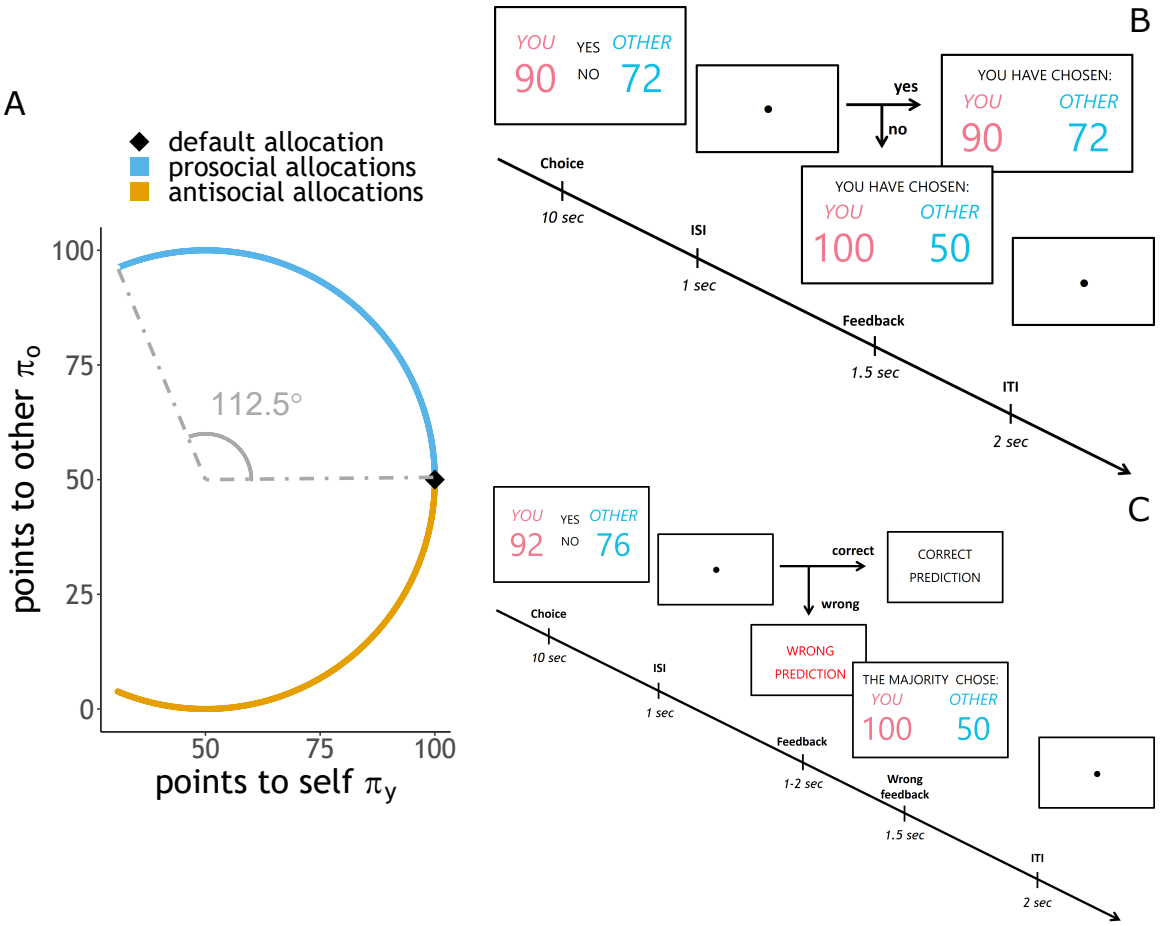

Figure 2.1: Trials in the experiment, the complete list of trials is available at osf.io/th6wp. **A**: Participants chose between a default allocation (black rhombus) and an alternative allocation, which could be either prosocial (light blue) or antisocial (orange). Allocations were limited to a certain arc of the circumference and to a certain range. **B**: Resource-Allocation Game. Participants observed the current alternative allocation and had a maximum of 10 seconds to respond. Decision cues ('yes'/'no') indicated which button to press for each decision (up/down arrows). Points for self and for the other were colour-coded (points for self: red; points for the other: blue) and were presented on the left and on the right of the decision cues. Both cues and points (self/other) switched position randomly across trials. If participants did not answer within 10 seconds, the trial ended, and they were automatically assigned the default allocation. Unanswered trials were considered missing data. After the decision, an inter-stimulus interval of 1 second divided the decision and the feedback. Feedback lasted for 1.5 seconds and displayed the the participant preferred by the participant. The trial ended with an inter-trial interval of 2 seconds. **C**: Manipulation phase. Participants were presented with an alternative allocation, and indicated whether they believed the agent preferred the alternative over the default allocation ('yes') or not ('no'). The choice could be made within 10 seconds, after which it would no longer be valid. After an inter-stimulus interval of 1 second, participants received feedback about their answer. If the prediction was correct, the feedback message 'correct prediction' appeared on the screen for approximately 1.5 seconds (minimum 1, maximum 2). If the prediction was wrong or not given in time, a similar feedback message ('wrong prediction' or 'no answer') appeared on the screen for about the same time, followed by the actual choice of the agent, lasting 1.5 seconds. The name of the agent varied between conditions (Group condition: 'majority'; Individual condition: 'participant'; Computer condition: 'computer'). After the feedback, the trial ended with an inter-trial interval of 2 seconds.

of the agent. We have chosen to polarise social attitude instead of dampening it in order to distinguish attitude conformity from a gradual increase in selfishness that was observed by Nax et al. (2015). Hence, if attitude converges towards that of the observed agent it cannot be attributed to an increase in selfishness. In addition, if participants did learn about decision makers who are less extreme than themselves, the attitude change would push them in the same direction as the regression to the mean. Any movement of the attitude away from the mean cannot then be obfuscated by this effect.

Participants' own attitude was categorised right before the manipulation phase. We determined whether participants had an α greater or less than zero (prosocial or antisocial) using the following formula (11)[7]:

$$\sum_t \mathbb{1}_{t,A} \cdot \operatorname{atan}\left(\frac{\pi_{ot} - 50}{\sqrt{(\pi_{ot} - 50)^2 + (\pi_{yt} - 50)^2} + \pi_{yt} - 50}\right), \tag{11}$$

where π_{yt} and π_{ot} are the points in the alternative allocation for self (you) and the other in trial t of the resource-allocation game, and $\mathbb{1}_{t,A}$ is an indicator variable that equals 1 when the participant preferred the alternative allocation in trial t and 0 when the participant preferred the default allocation in trial t.

If participants displayed a prosocial attitude in the game ($\alpha_{\text{before}} > 0$), they observed an agent with an extremely prosocial attitude ($\alpha_{\text{obs}} \approx 45°$, one point for the other equals one point for the self); and vice versa: if participants displayed an antisocial attitude ($\alpha_{\text{before}} < 0$), they observed an agent with an extremely antisocial attitude ($\alpha_{\text{obs}} \approx -45°$, one point for the other equals negative one point for the self). In order for participants to easily predict the agent's attitude, agents

[7]Our participant categorisation method is agnostic to model fitting (Box 2.2) since we could not tell *a priori* which model would fit participants' data best. This not withstanding, when using the winning cognitive model, 320 participants out of 335 (95.5%) fall into the same categories. Mismatch affects only participants with a moderate social attitude (mean $|\alpha_{\text{before}}| = 2.72°$, max$= 12.83°$, $N_{Baseline} = 8$, $N_{Computer} = 4$, $N_{Individual} = 1$, $N_{Group} = 2$).

were based on Baseline participants that made very consistent choices (e.g., did not choose both antisocial and prosocial allocations).

The dependent variable that we use to measure conformity is *attitude convergence*, denoted by δ_{diff}. To compute its value, we use equation (12):

$$\delta_{\text{diff}} = \delta_{\text{before}} - \delta_{\text{after}} = |\alpha_{\text{before}} - \alpha_{\text{obs}}| - |\alpha_{\text{after}} - \alpha_{\text{obs}}|, \tag{12}$$

where δ_{before} and δ_{after} are the distances between the attitude of the observed agent α_{obs} and participant's attitude estimated respectively before and after the manipulation phase. In order to have comparable results with the other conditions, we decided to use this measure also for Baseline participants as if they were predicting choices of an agent from the Group or the Computer condition.

We have chosen δ_{diff} as a measure of attitude conformity because it has two critical advantages over previous measures used in the literature (Garvert et al., 2015; Apps and Ramnani, 2017). First, δ_{diff} depends on the original distance from the model: if participant's starting attitude is very close to that of the observed agent, then δ_{diff} can only be small. As this is a conservative measure, it prevents close participants from biasing the estimate at sample level. Second, by taking into account the attitude distances from α_{obs} of both α_{before} and α_{after}, δ_{diff} differentiates between participants who shift attitude closer to the agent, and those who overshoot and become more extreme than the agent. It is indispensable to distinguish between these two types of attitude change, as the hypotheses that we test–with the exception of time-dependence–are concerned only with the former kind (moving closer to the agent).

2.2.3 Main Predictions

Using attitude convergence to measure attitude conformity, we predict the following: i) if the time-dependence hypothesis is true, we should observe a negative δ_{diff} in all conditions; ii) and iii) if the contagion or compliance hypothesis are true, average δ_{diff} should be significant and positive in all conditions excluding Baseline; iv) if the preference learning hypothesis is true, δ_{diff} should be significant and positive in the Individual and Group conditions, but not in the Baseline or Computer conditions; v) if the norm learning hypothesis is true, δ_{diff} should be significant and positive in the Group condition, but not in the Baseline or Computer conditions.

Predictions regarding attitude convergence for each of the five hypotheses are summarised in the left part of Table 2.1. Notice that the predictions of the time-dependence hypothesis have the direction opposite to all the other hypotheses. Moreover, the two versions of the norm learning hypothesis (norm uncertainty and norm salience) make no specific prediction about attitude change in the Individual condition.

2.2.4 Other Measures

While attitude convergence is the main measure that we use to distinguish among our hypotheses, we also need to test ancillary predictions that these hypotheses make to distinguish between the contagion and compliance hypotheses, and between the taste learning and norm learning hypotheses. For this purpose, we adopt a series of additional measures. The right-hand side of Table 2.1 summarizes the related predictions.

Compliance. The contagion and compliance hypotheses make identical predictions in terms of attitude change. To distinguish between them, we assess the compliance tendency of each participant. Specifically, we take advantage of a phe-

Hypotheses	Conditions				Other Measures
	Baseline	Computer	Individual	Group	
Time-dependence	↓	↓	↓	↓	
Contagion	–	↑	↑	↑	
Compliance	–	↑	↑	↑	Compliance Index (only $\geq 25\%$)
Preference learning	–	–	↑	↑	Consistency increase (in human conditions)
Norm uncertainty	–	–	– /↑	↑	Different norms (human vs. computer)
Norm salience	–	–	– /↑	↑	Same norms (human and computer)

Table 2.1: The predictions of the five hypotheses. "↑" refers to increasing extremeness of the attitudes. "–" means no change predicted. "↓" refers to the shift towards selfishness.

nomenon observed in the literature: when presented with conflicting choices during a task, such as being helpful or being spiteful towards others, complying participants think they should demonstrate both types of behaviour in order to meet the experimenter's expectations, even if these choices yield paradoxical outcomes (Zizzo and Fleming, 2011; Zizzo, 2013; Fleming and Zizzo, 2014). If we observe such behaviour in the resource-allocation game, then it is plausible that authority compliance—rather than conformity explanations—leads to attitude change.

We thus consider separately the proportion of prosocial alternatives and the proportion of antisocial alternatives chosen over the default allocation: we define our index of compliance as the smallest of these two numbers in percentage terms. To distinguish between compliant and non-compliant participants, we use a preregistered threshold set to 25% (osf.io/th6wp). In other words, a participant is said to be compliant if she chose both prosocial and antisocial alternatives at least once out of every four choices made. This index allows us to test whether conformity—and in particular conformity with a non-human agent (Computer condition)—originates from contagion or is rather linked to the tendency to comply with the

experimenter. In the latter case, we explore attitude change only in non-compliant participants.

Consistency Increase. The preference learning hypothesis predicts that participants change their attitude because they learn their own social preferences from others. Since learning in this case should reduce participants' uncertainty about how they want to behave, we should observe a corresponding increase in choice consistency after the manipulation phase in human (Individual, Group) relative to non-human (Baseline, Computer) conditions. We can test for changes in consistency by looking at the cognitive model used to understand participants choices, and in particular at the parameter σ (Box 2.2). σ represents variability in participants' choices: a small σ corresponds to very consistent choices and vice versa. Hence, we test whether participants' σ decreases after the manipulation phase (13):

$$\sigma_{\text{diff}} = \sigma_{\text{before}} - \sigma_{\text{after}} > 0, \tag{13}$$

where σ_{before} and σ_{after} are estimated from choices before and after the manipulation phase.

Norm following. The norm learning hypothesis assumes that participants' behaviour is influenced by beliefs about what constitutes a socially appropriate or inappropriate action. Accordingly, we should expect that prosocial and antisocial participants have different beliefs about what choices are considered appropriate in the resource-allocation game. To measure appropriateness perception, participants in the Computer, Individual, and Group conditions completed the norm elicitation task (Krupka and Weber, 2013) at the end of the experiment[8]. In this

[8]The decision to include the norm elicitation task followed the data collection of the Baseline experimental condition, as we initially conceived this condition as a benchmark for decisions in the resource-allocation game only.

task, participants rate on a 4-point Likert scale the degree of social appropriateness of choosing the alternative allocation over the default option in a selection of choices from the resource-allocation game. If one of these ratings, randomly chosen, matches that of the majority of other participants in the experimental session, then the participant is rewarded with €3. This procedure ensures that participants report their true beliefs about what the majority thinks is socially appropriate, or what constitutes a social norm[9].

Using the norm elicitation task, we test whether prosocial and antisocial participants have different perceptions of the social norms in the game. This is done in the Computer condition where no social information can be acquired from the agent. We expect that any difference in appropriateness ratings is due to participants' original beliefs before the task. If prosocial and antisocial participants do indeed report different normative beliefs, this could explain their differences in social attitudes, thus supporting the norm learning hypothesis.

Norm uncertainty vs. norm salience. The norm elicitation task is also used to discriminate between the two versions of the norm learning hypothesis. If appropriateness ratings differ between the Computer and the human conditions (in the direction of observed agents), then what has changed is the representation of what the norm is (norm uncertainty); if instead the appropriateness ratings are the same among antisocial (prosocial) participants across all conditions, then we assume that it is the perception of strength of the norm that has changed (norm salience). The observed agents indeed choose very consistently, which should induce participants to think that the extreme behaviour they learn is in fact well-established.

[9]Participants complete the norm elicitation task at the end of the experiment, therefore, it is possible that instead of reporting their normative beliefs they could coordinate on the behaviour of the observed agent. This however is unlikely as a recent study (d'Adda et al., 2016) shows that mere exposure to new information does not suffice to influence norm elicitation.

Box 2.2: Cognitive Modelling of Choices

Bias Parameter κ. We associate participants' choices to their social attitude via Equation 10 in the main text. Yet this estimate or that of other parameters could be biased by the participant's compliance tendency (see *Other Measures*). To control for compliance, we allow for the possibility that participants prefer an alternative allocation even when it should be on a par with the default allocation. To represent this added subjective value of the alternative allocation, we define for each participant a bias parameter κ (14):

$$V(D) = V(100, 50) = 100 + \tan\alpha \cdot 50 - \kappa, \tag{14}$$

where κ is equivalent to an amount of penalty points for the default allocation: The higher κ is, the higher the propensity to choose the alternative over the default allocation. The association between the compliance index and κ is confirmed in the Results' Box 2.3.

Variability Parameters τ vs. σ. During the allocation game, participants might show variability in the way they choose, such as being more or less prosocial (or antisocial) from choice to choice. Not accounting for this variability *within* each part of the game (before or after prediction) could bias the estimation of variability *between* parts, namely the change in attitude due to agent prediction. To estimate choice variability, we compare two types of cognitive models that also give different interpretations about the nature of social attitude.

The first model type, Stable Attitude, assumes that attitudes are a stable personal trait, and that any variability in participants' choices is due to cognitive mistakes when comparing different options. If for instance a person occasionally shows a more prosocial (or more antisocial) attitude than usual, this fluctuation is interpreted by the model as a miscalculation on how to behave. Comparisons errors are modelled through the parameter τ: the smaller (larger) τ is, the higher (lower) the probability of choosing consistently with one's own attitude. Stable Attitude models compare alternatives using a softmax function (Sutton and Barto, 1998, 15):

$$\Lambda(\Pr(D = 1)) = \frac{V_D - V_A}{\tau}, \tag{15}$$

where Λ is the logit link function, $\Pr(D = 1)$ is the probability of choosing the default allocation, V_D and V_A are the estimated values for the default and alternative allocations, as in (10) and (14).

The second model type, Variable Attitude, assumes instead that attitude is a variable mental state. If for instance people behave more or less nicely, this is interpreted as a natural fluctuation of attitude. Participants choices are modelled using random preference (Koppen, 2001; Regenwetter et al., 2010, 2011): every time the participant has to make a decision, her social attitude α is sampled from a normal distribution with centre μ and standard deviation σ. The parameter σ represents variability in the way participants behave: the smaller (larger) σ is, the more (less) consistent the participant will be across her choices. The model is defined as (16):

$$\Phi(\Pr(D=1)) = \begin{cases} \frac{T_\alpha - \mu}{\sigma}, \text{ if } \pi_o > 50 \\ \frac{\mu - T_\alpha}{\sigma}, \text{ if } \pi_o < 50 \end{cases}, \tag{16}$$

Where Φ is the probit link function and threshold T_α is the value of α for which the default and alternative allocations have equal subjective value ($V_D = V_A$, 17):

$$T_\alpha = \text{atan}\left(\frac{\pi_y - 100 + \kappa}{50 - \pi_o}\right), \tag{17}$$

If the sampled $\alpha > T_\alpha$, an allocation is preferred and consequently taken, otherwise the other option is chosen.

Error Parameter ε. The error parameter ε defines the probability with which participants make a mistake in implementing their choice (e.g., mistyping or inattention). The probability of choosing the default allocation is expressed as (18):

$$\Pr(D=1) = (1-\varepsilon)\cdot \Pr_{\text{model}} + \varepsilon \cdot \frac{1}{2}, \tag{18}$$

where Pr_{model} represents the probability of choosing the default allocation according to the model under consideration (15 for Stable Attitude or 16 for Variable Attitude). The error parameter thus allows to assume that participants' answers are a mixture between model-based choices and random errors.

Model Estimation. We estimate three versions of each model type. In the full version of a model, all parameters are estimated twice, before and after the manipulation phase. A second, simpler version of the models assumes that social attitude α is fixed for the whole task, as if it could not change with the manipulation; α is thus estimated only once across all choices. In the third version of the models instead, it is the variability parameter (σ for Variable Attitude or τ for Stable Attitude) to be estimated once for the whole task, as if participants could not get more or less self-consistent in their choices after the manipulation phase. Models thus vary based

on two factors: 2 (Stable Attitude / Variable Attitude) × 3 (fixed attitude / fixed variability / both vary). Consequently, we estimate and compare 6 unique models. Models are estimated in JAGS (Plummer, 2003) using the rjags (Plummer, 2019) and R2jags (Su and Yajima, 2015) packages. Parameters are fitted using Hierarchical Bayesian Analysis (HBA, Shiffrin et al., 2008) on two levels: a sample level (by subject) and a subject level (by time: before/after prediction; Supplementary Figures 6.1 and 6.2). For fitting we consider only trials with a response time greater than 200 milliseconds, as psychophysical limits in encoding stimuli suggest that beyond this threshold participants cannot give an answer based on the on-screen information. For each model, we ran 4 Markov chains for 100,000 iterations, with a burn-in period of 5,000 iterations and a thinning rate of 10. The model with the lowest Deviance Information Criterion (DIC) is selected and used for the statistical analyses. We use the maximum a posteriori (MAP) estimate to derive the most likely value for each parameter, including the two composite indices attitude convergence δ_{diff} and consistency increase σ_{diff}.

Model Comparison. The cognitive model that describes participants' behaviour best is the full version of the Variable Attitude model in which both α and σ vary before/after the manipulation phase ($DIC_{VA} = 26,000.66$, Figure 2.2). Model comparison thus suggests that both attitude and attitude variability change after the manipulation phase, and that α varies across trials rather than being stable. This latter finding is further supported by the generally lower DIC values of Variable Attitude models as compared to all Stable Attitude models. We use the full Variable Attitude model for all the analyses in the main text.

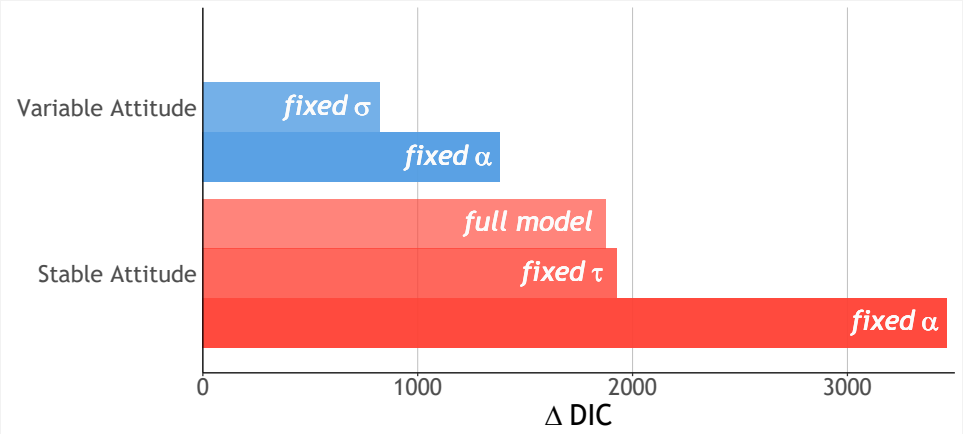

Figure 2.2: The difference ΔDIC between the Deviance Information Criterion (DIC) of a model and the full version of the winning Variable Attribute model. The Variable Attitude models are in blue, and the Stable Attitude models are in red.

2.3 Results

Based on choices before the manipulation phase, 77% (23%) of participants were categorised as having a prosocial (antisocial) attitude. When presented with prosocial alternatives, prosocial participants chose them over the default allocation 40% of the time, while antisocial participants did the same with antisocial alternatives around 54% of the time. According to our model estimates, mean social attitude before prediction α_{before} was 20° for prosocial and -22° for antisocial participants. However, for a significant portion of participants (22% of the sample), 1 point for the other player was worth less than one tenth of a point for oneself ($-5^\circ < \alpha_{\text{before}} < 5^\circ$), meaning that many participants showed a moderate, if not selfish, social attitude.

Accuracy in the manipulation phase was relatively high: even if we consider participants who were excluded for low prediction accuracy (see Methods), the average number of correct predictions in the last 20 trials was 18.6 ($M_{Computer} = 19.1$, $M_{Individual} = 17.9$, $M_{Group} = 18.7$). This result suggests that participants have successfully learned the attitude of the observed agent. Predicting choices of a prosocial (antisocial) agent increased the proportion of prosocial (antisocial) alternatives chosen over the default allocation (Baseline (no agent): -4%, Computer: +0.5%, Individual: +7%, Group: +8%). These results are mirrored by changes in social attitude, where α_{after} is on average more extreme than α_{before} (Baseline: $+0^\circ$, Computer: $+3^\circ$, Individual: $+7^\circ$, Group: $+6^\circ$). Given that after observing agent's choices participants' attitude becomes more extreme rather than more self-oriented, we reject the time-dependence hypothesis, which predicts convergence towards selfish choices.

Attitude Convergence. In each condition we measure whether attitude convergence δ_{diff} is significantly greater than zero, that is if participants' attitude moves

towards the learned information. Given that the normality assumption does not hold (Shapiro-Wilk test, $p < .001$), we adopt the Wilcoxon signed-rank test for the analyses. Consistent with both the contagion and compliance hypotheses, attitude convergence δ_{diff} is significantly greater than zero in all conditions except Baseline (Baseline: $\log(V) = 8.46$, $p = .216$, $\delta_{\text{diff}} = 0°[-2°, 1°]$, $r = .07[-.09, .23]$; Computer: $\log(V) = 7.28$, $p = .008$, $\delta_{\text{diff}} = 3°[1°, 6°]$, $r = .31[.10, .58]$; Individual: $\log(V) = 6.99$, $p < .001$, $\delta_{\text{diff}} = 6°[2°, 9°]$, $r = .50[.32, .71]$; Group: $\log(V) = 8.02$, $p < .001$, $\delta_{\text{diff}} = 6°[3°, 8°]$, $r = .54[.37, .71]$), meaning that participants in these conditions shifted attitude towards that of the observed agent.

We also measure the difference in convergence across conditions. A Kruskal-Wallis test reveals that there is a significant effect of condition on attitude convergence ($\chi^2(3) = 22.87$, $p < .001$, $\varepsilon^2 = .07[.03, .14]$). Post-hoc pairwise comparisons reveal that attitude convergence differs between the Baseline and Group conditions ($W = 6.11$, $p < .001$) and between the Baseline and Individual conditions ($W = 4.26$, $p = .042$), whereas there seems to be no statistical difference between the other conditions.

Compliance. To study the influence of the experimenter on attitude convergence, we categorise participants using the compliance index. We find that 46 participants (Baseline: 26, Computer: 9, Individual: 8, Group: 6), around 14% of the sample, are above the 25% threshold (Figure 2.7A; Supplementary Figure 6.5).

We first test whether attitude convergence δ_{diff} is significantly greater than 0 in participants above threshold. Despite a small sample size, the test is significant in the Computer condition, but not in the other conditions (Wilcoxon rank sum test, $\log(V) = 3.76$, $p = .023$, $\delta_{\text{diff}} = 14°[5°, 22°]$, $r = .81[.6, 1]$, $n_{\text{obs}} = 9$). Attitude convergence is also significantly different between conditions (Kruskal-Wallis rank-sum test, $\chi^2(3) = 9.54$, $p = .023$, $\varepsilon^2 = .21[.05, .54]$, $n_{\text{obs}} = 46$).

We divide compliant participants based on whether they completed the manipulation phase (i.e., Computer, Individual, and Group conditions) or not (Baseline) – i.e. we assume that participants are susceptible to authority compliance in the same way in all conditions with an agent of any kind. Consistent with the compliance hypothesis, attitude convergence is significantly greater than 0 in participants who predicted an agent but not in the Baseline condition (Figure 2.3; agent conditions: $\log(V) = 5.55$, $p < .001$, $\delta_{\text{diff}} = 14°[8°, 21°]$, $r = .75[.59, .90]$, $n_{\text{obs}} = 23$; Baseline: $\log(V) = 4.99$, $p = .40$, $\delta_{\text{diff}} = -2°[-10°, 6°]$, $r = .06[-.36, .48]$, $n_{\text{obs}} = 23$) and the two categories are also significantly different ($\log(V) = 4.83$, $p = .002$, $r = .45[.22, .68]$).

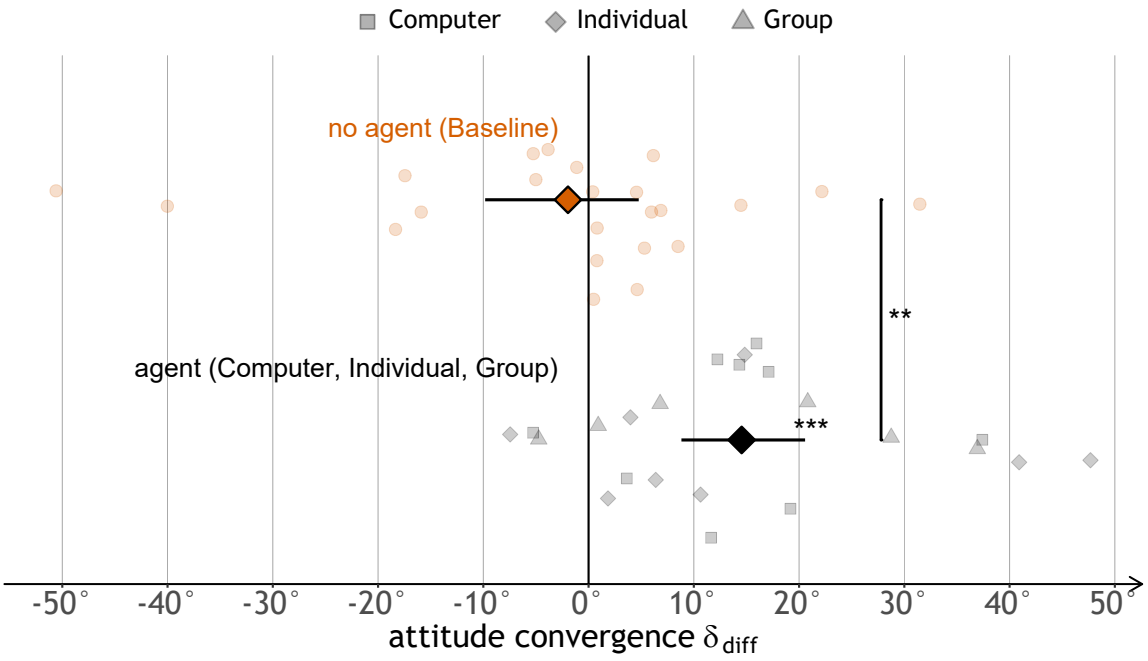

Figure 2.3: Mean attitude convergence comparing "agent" conditions to Baseline, participants above threshold. Error bars indicate t-adjusted, 95% Gaussian confidence intervals. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

We then move to participants below threshold, where we observe a different pattern compared to the full sample (Figure 2.4): δ_{diff} is still significantly greater than 0 in the Group and Individual conditions, while it is no more significant in

the Computer condition (Baseline: $\log(V) = 8.08$, $p = .254$, $\delta_{\text{diff}} = 0°[-1°, 1°]$, $r = .06[-.12, .23]$, $n_{\text{obs}} = 109$; Computer: $\log(V) = 6.87$, $p = .120$, $\delta_{\text{diff}} = 2°[-1°, 4°]$, $r = .18[-.11, .44]$, $n_{\text{obs}} = 56$; Individual: $\log(V) = 6.62$, $p = .003$, $\delta_{\text{diff}} = 4°[1°, 7°]$, $r = .44[.17, .73]$, $n_{\text{obs}} = 44$; Group: $\log(V) = 7.86$, $p < .001$, $\delta_{\text{diff}} = 5°[2°, 7°]$, $r = .53[.37, .69]$, $n_{\text{obs}} = 80$).

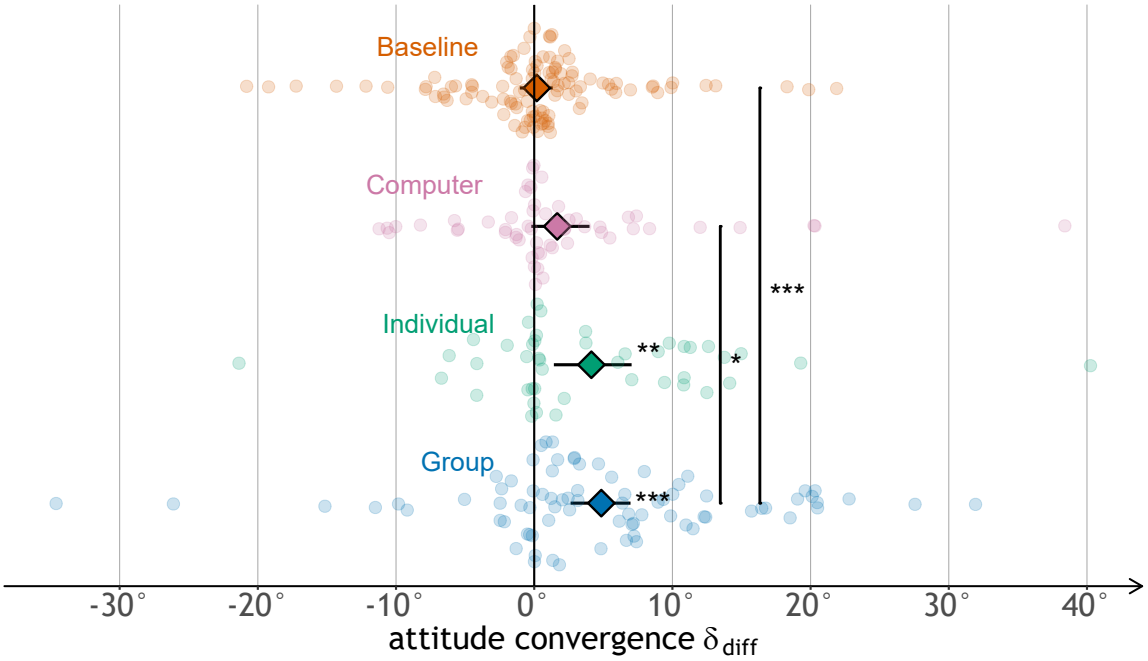

Figure 2.4: Mean attitude convergence by condition for participants below compliance threshold (25%). Error bars indicate t-adjusted, 95% Gaussian confidence intervals. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

Attitude convergence is also significantly different between conditions ($\chi^2(3) = 21.22$, $p < .001$, $\varepsilon^2 = .07[.03, .18]$, $n_{\text{obs}} = 289$, Figure 2.4). Post-hoc comparisons find a difference between the Baseline and Group conditions ($W = 6.11$, $p < .001$), and between the Group and Computer conditions ($W = 4.26$, $p = .042$).

Given that the procedure to categorise participants based on compliance was pre-registered, one may wonder whether using different thresholds could alter significantly the results. The table below shows however that these results remain largely consistent even after varying the threshold.

Figure 2.5: Results of main tests (left; attitude convergence greater than zero, attitude convergence comparison across conditions) and sample size (right) using different thresholds. Colours and asterisks code significance: *: $p < .05$; **: $p < .01$; ***: $p < .001$.

We also consider whether compliance could be a continuous, rather than an all-or-nothing, trait. To this end, we run a robust linear regression (Koller and Stahel, 2017) with attitude convergence as dependent variable, and with experimental condition and the interaction between compliance and experimental condition as predictor variables (Figure 2.6A). We compare the deviance of the regression against a simpler model using only experimental condition as an independent variable: the full model has a better fit on the data ($\chi^2(4) = 20.78$, $p < .001$). Again, the results presented above are nicely replicated, with the main effect of Group and Individual conditions being significant, as well as the interaction between the Computer condition and compliance; no other effect is significant (Figure 2.6B).

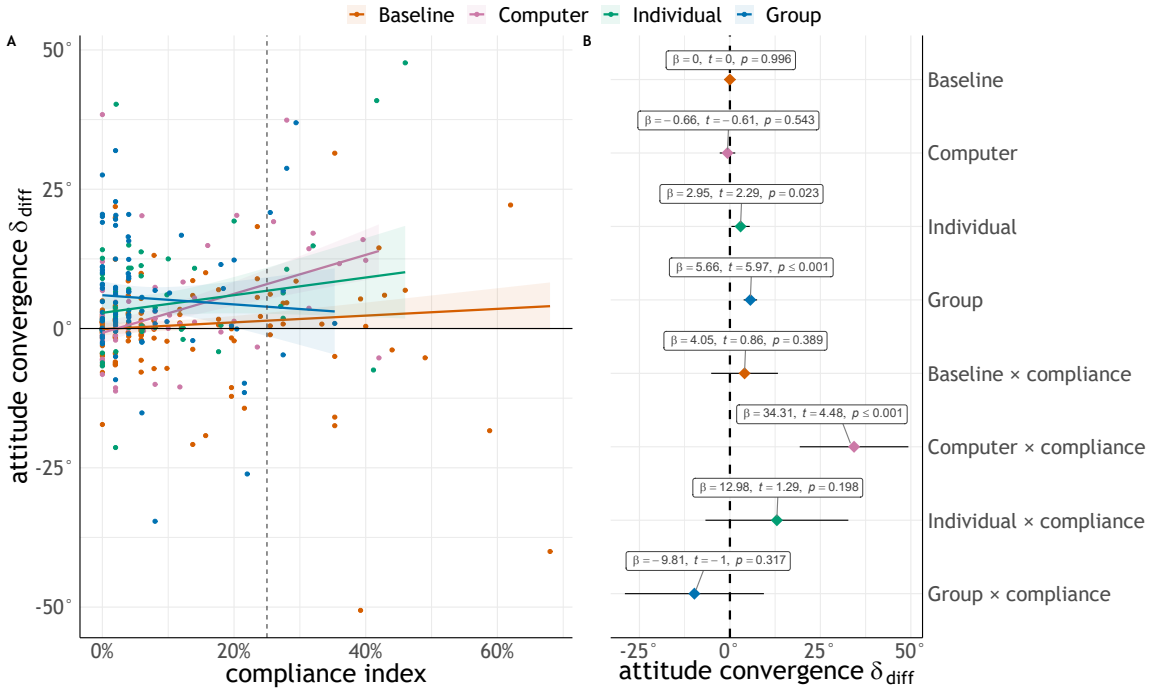

Figure 2.6: A: Robust regression on attitude convergence with experimental condition and the interaction between compliance and experimental condition as predictor variables. Shaded areas indicate 95% confidence intervals. B: Coefficients of the regression. Labels report unstandardised effect size, t-value, and p-value. Error bars indicate t-adjusted, 95% Gaussian confidence intervals.

Preference Learning. Our findings for participants not susceptible to compliance are consistent with either the norm learning or the preference learning hypotheses. We first test the second prediction of the preference learning hypothesis, namely that learning about others' attitude should significantly increase participants' consistency. We thus test whether there is a consistency increase σ_{diff}, and if this increase is higher in human conditions (Individual, Group) than after predicting a computer's choices or nothing at all (Computer and Baseline). Shapiro-Wilk test for normality is significant in all conditions ($p < .001$), therefore we adopt non-parametric tests. Wilcoxon signed-rank tests show that consistency increase σ_{diff} is significantly greater than zero in all conditions ($p < .01$, Figure 2.7B).

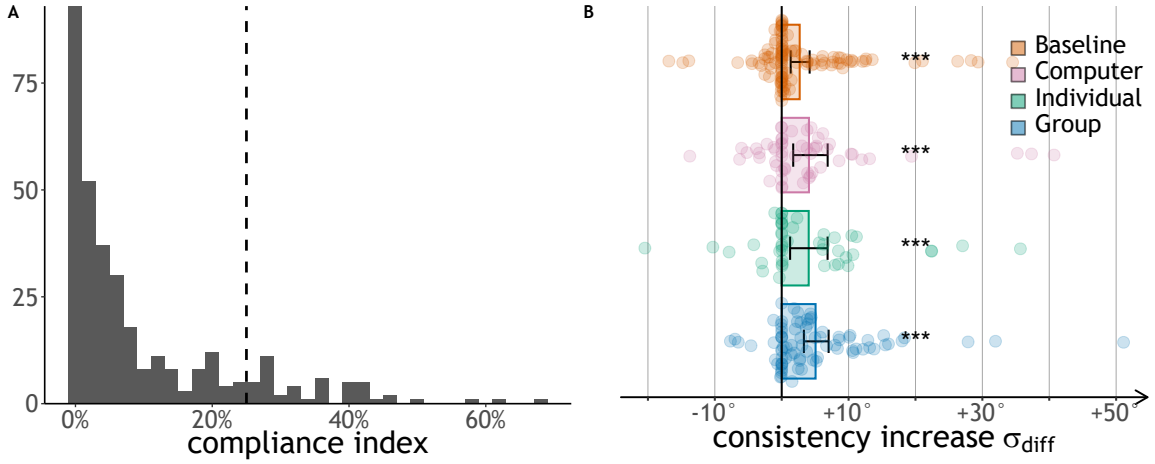

Figure 2.7: A: Distribution of the compliance index across all participants. The vertical line shows the threshold value (25%) beyond which a participant is considered to be susceptible to authority compliance. B: Consistency increase across conditions for participants below threshold. Participants become more consistent after the manipulation phase ($\sigma_{\text{diff}} > 0$), but the increase is not significantly different across conditions. Error bars indicate t-adjusted, 95% Gaussian confidence intervals. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

Box 2.3: Compliance and the κ Parameter

We test whether the bias parameter κ_{before} estimated before the manipulation phase correlates with the compliance index. If correct, the bias parameter could be then used as an improved measure of authority compliance, in that it could integrate compliance effects directly within the computation of the decision process, and improve the estimates of other parameters. Splitting participants below and above the compliance threshold, we observe that the two groups have significantly different estimated values of κ_{before} (Wilcoxon rank-sum test with continuity correction, $\log(V) = 6.39$, $p < .001$, $r = .54[.47, .62]$), with participants below threshold with average $\kappa_{\text{before}} = 1.70[1.33, 2.07]$ and participants above threshold with average $\kappa_{\text{before}} = 17.45[13.61, 21.29]$. We then measure the association between the two measures using a Spearman's rank correlation, and find a significant association ($\rho = .70[.64, .75]$, $p < .001$). These results support the hypothesis that the bias parameter κ in our model also captures a significant portion of the effect of authority on participants.

However, when we test whether σ_{diff} differs across conditions, the test fails to reject the null hypothesis that conditions do not differ ($\chi^2(3) = 7.39$, $p = .060$, $\varepsilon^2 = .03[0, .09]$). Given the near significance of the statistic, we tested post-hoc

pairwise comparisons to look for significant differences. According to a Dwass-Steel-Critchlow-Fligner test however, none of the differences were significant. Although we cannot claim that consistency increase is equivalent across conditions, these results suggest that the effect size of the differences is marginal at best. Thus, we conclude then that the preference learning hypothesis does not adequately explain our data.

Norm Learning. We use the data from the norm elicitation task to test the plausibility of the norm learning hypothesis and to distinguish between norm uncertainty and norm salience.

We first compare appropriateness ratings between prosocial and antisocial participants in the Computer condition using a series of Kruskal-Wallis tests (one test for each rating; Figure 2.8 top). Appropriateness ratings are statistically different for every rating, even after correcting for multiple comparisons (all $p < .045$). These ratings link norm perception to social attitude: prosocial participants seem to consider it very appropriate to give money to the other and very inappropriate to take money, while the opposite is true for antisocial participants.

To distinguish between norm uncertainty and norm salience, we test whether the distribution of appropriateness ratings differs across conditions, by participant type. Two Kruskal-Wallis tests out of twenty-four are statistically significant, but do not survive the correction for multiple comparisons (all $p > .065$). We cannot reject any hypotheses that the norms for prosocial (antisocial) participants are the same in all three conditions. Similar ratings in human (Group/Individual) and Computer conditions thus support the norm salience over the norm uncertainty hypothesis.

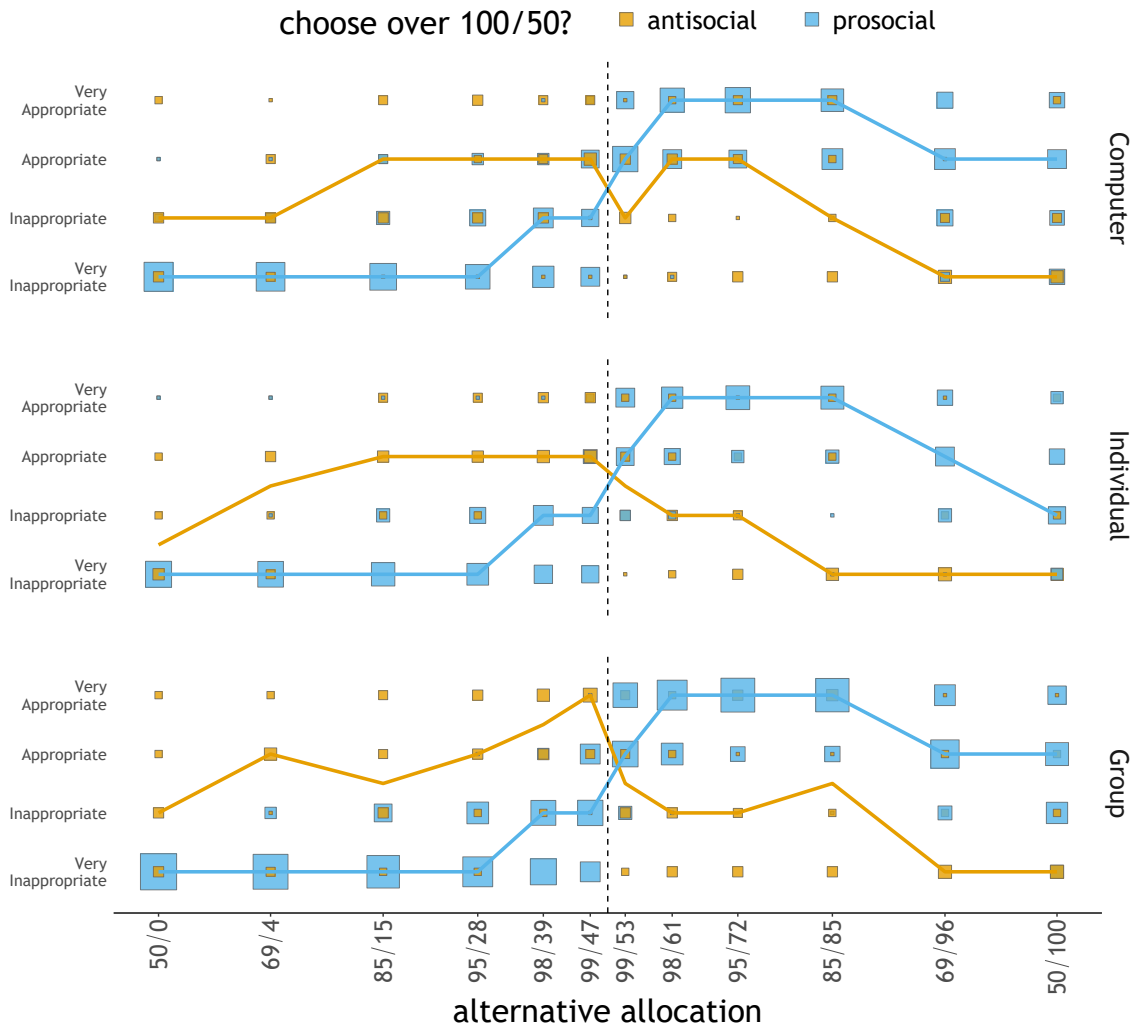

Figure 2.8: Appropriateness ratings for prosocial and antisocial participants in the Computer (top), Individual (centre), and Group (bottom) conditions. Only participants below threshold are plotted. Square size is proportional to the number of participants, whereas the lines connect the median ratings for each alternative allocation.

2.4 Discussion

In this study, we identified and estimated the contributions of several competing explanations to attitude conformity in social decision making. Attitude conformity was assessed using a series of cognitive models coupled with several experimental conditions and complementary indices, which helped to test the predictions of each hypothesis. Participants' attitude became more prosocial or antisocial when

they learned about the choices of an extremely prosocial or antisocial agent, regardless of whether the agent was a group of people, one person, or a computer. We found however that attitude conformity in the Computer condition was mainly driven by the participants who are prone to comply to authority demands, rather than reacting to the observed agent. Once we had accounted for this propensity to comply, computational modelling helped us to disentangle the surviving hypotheses—norm learning and preference learning—for the remaining participants. Specifically, we tested the preference learning prediction that participants should have become more self-consistent in their choices after learning from human agents. Since we failed to confirm this prediction, we identified social norm learning as the most plausible explanation of attitude conformity. Based on the norm elicitation task (Krupka and Weber, 2013) we further suggest that the main source of prosocial and antisocial conformity is uncertainty about *how salient* following the norm is and *not* uncertainty about the norm *per se*. We cannot exclude however that norm uncertainty can still represent a source of conformity in circumstances where social norms are less well-known (e.g., observational learning, Kenward et al., 2010).

A number of findings support the idea that norms and the beliefs related to them are at the basis of social attitudes. Social appropriateness has been shown to play a role in decisions in various economic games (López-Pérez, 2008; Krupka and Weber, 2009, 2013; Kimbrough and Vostroknutov, 2015, 2016, 2018b). More relevantly to our study, it was found that anonymity, and therefore reduced accountability, appears to have clear effect on allocation choices. Experiments with increased anonymity—also with respect to the experimenter, i.e. double blind paradigms—show plummeting contributions in economic games such as the Dictator game (Hoffman et al., 1996; Franzen and Pointner, 2012a). At the same time, even subtle cues of being observed seem to increase contributions (Haley and

Fessler, 2005). The impact of reputation can also account for attitude change driven by compliance. Participants who have a strong tendency to choose the alternative option, regardless of whether it is beneficial or detrimental for the other and regardless of the identity of the observed agent, may think that this is what authority wants, and that this is the norm in the experiment (Bardsley, 2007; Guala and Mittone, 2010; Kimbrough and Vostroknutov, 2016). The complementary result, that participants who are not influenced by authority only change their attitude when learning about other humans' behaviour, works in a similar fashion. By learning to predict the agent's behaviour, participants deduce how salient following the norm is for others, and change their behaviour to be more consistent with them. Therefore, we can conclude that the two effects that we observe—authority compliance and attitude conformity in human conditions—are both in line with the general social norms explanation.

The results of this study prompt some additional thoughts about the process of learning social norms. First, we observe that information about norms can spread through indirect transmission (Ivaturi and Chua, 2019). During the experiment participants cannot interact in any way with the observed human agent—who is not physically present—but participants can nevertheless extract some information about the norm from observing the behaviour of the agent. Indirect transmission thus highlights how adherence to social norms can be pervasive in dispersed and loosely regulated groups such as online communities. Second, the fact that participants conform by learning how salient a norm is implicates that if a norm is already salient among a group of individuals then they should be more resilient to conformity influences. If future studies do confirm that norm perception prior to observation does predict conformity, this could suggest new measures to countervail polarisation in social discourse.

Our contribution not only fosters and provides better characterisation of the norm learning hypothesis, but also systematically devalues the several competing explanations that we tested, that to our knowledge were not yet properly compared in one framework. These *non-social* hypotheses include time-dependence, contagion, and preference learning. The social/non-social distinction is crucial here as it gives an insight into how to interpret conformity dynamics in interpersonal relations: if a person changes her attitude we suggest that this change has to be primarily social in nature, and linked to the changes in social context in which the decision maker is placed. This idea can have profound implications for studying any social learning mechanisms and social decision making in general. Specifically, many non-social explanations of the change in behaviour can be ruled out.

Our study does not come without some limitations. The experiment design is between-subjects, and it is thus not possible to compare directly the effect of the various manipulations, nor does it allow to exclude the possibility that multiple mechanisms are at work simultaneously. While this weakness does not fundamentally challenge the reported findings, implementing an intermixed design such as the ones proposed in Chung et al. (2015) or Suzuki et al. (2016) could yield more powerful predictions and interpretations. A second constraint of our experiment design is that in some conditions we could not reach the pre-determined sample size necessary to achieve the power $1 - \beta = .95$. We note however that our findings seem robust, even when applying design changes such as different ways to account for the influence of compliance, suggesting that this problem might be not too concerning.

Another limitation of the design is that, given that the attitude of the observed agent was fixed, social distance from the agent and attitude change are correlated (see also Box 2.5). This means that we may be missing a connection between how close one's initial attitude is to the observed agent's and how much she will con-

Box 2.4: Contributions of Cognitive Modelling

Cognitive modelling does not only play a fundamental role in testing the predictions of the different hypotheses, but is also inherently connected to two additional contributions of this paper. First, we add to the series of studies challenging the conceptualisation of preference as a stable trait of people, and thus the use of the softmax function as the privileged method to model value-based choices. Studies on both risk (Loomes et al., 2002; Gul and Pesendorfer, 2006) and inter-temporal preferences (Moutoussis et al., 2016; Lu and Saito, 2018) have in fact highlighted how choice variability can be better explained by fluctuations of subjective preferences rather than "errors" in comparing different alternatives. This is in line with our finding that the Variable Attitude model explains behavioural data better than the Stable Attitude model. While we do not claim that computational distortions are absent during the estimation of value, we nonetheless support the idea that this mechanism cannot be the only one, nor can it be the main cause for choice inconsistencies in value-based decision making.

This interpretation finds additional support in recent perspectives on brain architecture, which hold that value representation is less specifically defined and is more distributed than current thinking suggests (Hunt and Hayden, 2017; Meder et al., 2017; Yoo and Hayden, 2018; Polania et al., 2019). Assuming that preferences vary across contexts and across time requires a network of resources that not only keeps track of the current internal state, but that takes also into account the situational factors and the different scopes within which the choice is considered. For instance, a decision to act prosocially would require the integration of the tendency of an individual to help others, considerations related to the nature of the interaction and of the other person, the general goals of the decision maker, as well as the history of choices preceding that particular choice. Considering the complexity of a choice and of the neural substrates that make it possible, it seems hard to postulate the stability of subjective value as a justifiable premise for studying personal preferences and attitudes.

While we stand by the current findings, future research could improve the Variable Attitude model by accounting for some of its limitations. One way to do this could be to integrate both types of choice variability (errors in comparison and variability in attitude) under a common cognitive model to test whether these mechanisms coexist and what are their individual contributions (see for example He et al., 2019; Regenwetter et al., 2018; Regenwetter and Robinson, 2017; Bhatia and Loomes, 2017). Such a model, however, requires either a prohibitive number of trials per participant, or the integration of some other type of information. This problem could

possibly be overcome by integrating temporal information to simple choice data: several studies have successfully analysed subjective choices with this method before using so-called sequential sampling models (SSM, see for instance Mormann et al., 2010). While this approach would require challenging improvements, such as disentangling variability both within and between trials, it could also promote the analysis of other decision components, such as the trade-off between fidelity with one's preferences and speed in making a decision.
A second contribution of cognitive modelling is the use of a computational parameter to directly measure the impact of authority compliance on the decision process (Box 2.2 and 2.3). This parameter correlates with the compliance index that we used in the present study to categorise participants. We propose that this parameter can be used independently to measure compliance to authority demands. Directly including the effect of compliance in the computational model has the advantage that other estimates, such as the person's attitude or its choice consistency, are corrected for the presence of this effect. We also consider this estimation procedure as more reliable than alternatives in the literature: while other methods indeed exist, they are based on ad hoc tasks to quantify authority demand (e.g., Fleming and Zizzo, 2014; De Quidt et al., 2018), whereas the measures we use work within the main task of the experiment, thus reducing the risk that results in one task do not extend to another. As a limitation of our approach, it could be argued that using a default option might seem too unequivocal; we argue however that this feature of the task design actually simplifies the expression of attitude by participants as it makes value comparison less challenging also from a computational point of view (see for instance Rangel and Clithero, 2012; Stewart et al., 2015). We thus think that our computational parameter could be of value to researchers who need to control for the influence of the experimenter when fitting decision models.

form after learning. To solve this problem, in future experiments we propose to dynamically adjust the attitude of the agent depending on participants' own attitude. This design can also help to understand what happens when prosocial participants observe an antisocial agent and vice versa. We have deliberately excluded this question from consideration in our experiment because we were not sure *ex ante* if we would manage to separate the effect of learning about a very socially distant agent from the drift of attitudes towards selfishness (though, *ex post*

Box 2.5: Attitude Distance from the Observed Agent

Given that attitude convergence δ_{diff} depends on the initial attitude distance from the agent, we test whether there are any differences in distance across conditions by means of a Kruskall-Wallis rank-sum test (Figure 2.9). The test is significant both with and without the participants above the 25% threshold (all $p < .001$). Post-hoc Dwass-Steel-Critchlow-Fligner tests reveals that participants in the Group and Individual conditions are on average significantly less distant from their agents than participants in the Baseline condition ($p < .05$), whereas there is no difference across all other comparisons ($p > .05$). If participants in the Group and in the Individual conditions had an average initial attitude closer to the agent than participants in the Baseline condition, then their maximum possible attitude change must have been smaller than that of Baseline participants. Since participants' attitudes in the Individual and Group conditions change *more* than those in the Baseline condition, this evidence–if anything–supports the idea that differences in attitude convergence across conditions cannot be explained by differing initial distances.

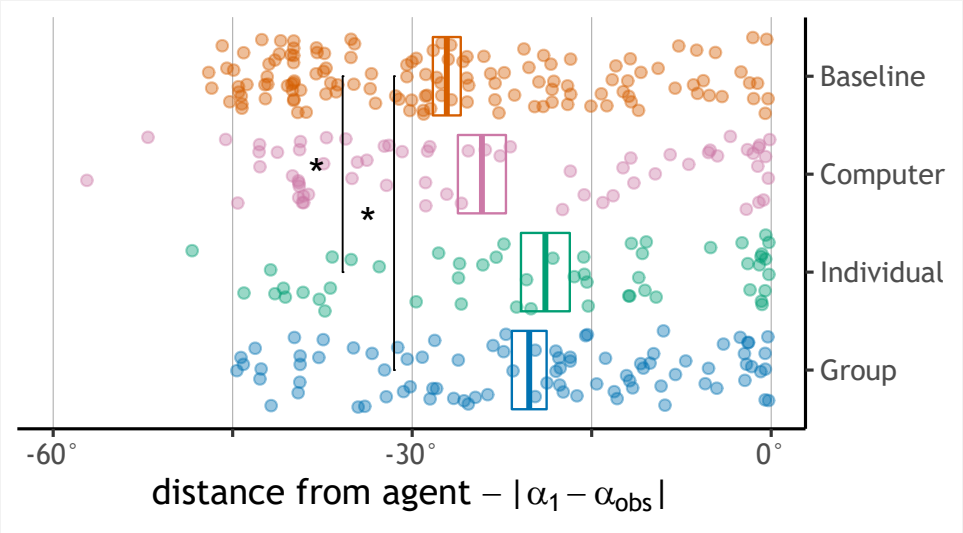

Figure 2.9: Crossbar plot of participants' attitude distance from the agent's before the manipulation phase. The graph includes all participants. Error bars indicate *t*-adjusted, 95% Gaussian confidence intervals.

we know that such time-dependence is not there, which should make it straightforward to test hypotheses about observing others with very different attitudes).

We would like to note that, contrary to the predictions of the norm learning hypothesis, attitude change in the Individual condition was not significantly smaller than attitude change in the Group condition. This unexpected result could be linked to the fact that participants were not informed about the size of the group, which in turn could have influenced their representation of the agent. A key direction for the future research will be to explore the relationship between group size and attitude change (e.g., Park et al., 2017). Another possible explanation for the lack of the difference between the Individual and Group conditions could be that participants in the Group condition were not connected in any way with the group of people whose behaviour they observed, and that they would have conformed more on average had they identified more with the group. This scenario could be compatible with recent findings suggesting that norms are stronger when there is a stronger group identification (Pryor et al., 2019). Testing this idea would require more rigorous control of the perception of the group by participants.

Finally, we would like to comment on the implicit assumption that we make when distinguishing between norm uncertainty and norm salience. Specifically, we assume that the norms elicited in the Computer condition were not influenced by the predicting of computer's behaviour in the second task, and thus these norms are those that participants had in mind while choosing in the first part of the resource-allocation game. It can be argued that learning about the computer's "attitude" can change the perception of norms and that our assumption is therefore incorrect. We disagree with this opinion on the following grounds. The computer is not a social agent, so whatever it is doing should not, by definition, change the perception of the *social* environment that participants are in. This is evident from the fact that attitude conformity is not significant in the Computer condition after

controlling for compliance to authority. Moreover, the fact that we do not find any differences in the elicited norms across the three conditions is much more likely to reflect the stability of normative beliefs across conditions rather than beliefs that change in all three conditions in exactly the same way. It can nonetheless still be argued that the mere experience of the task influences norm perception. This idea has been directly tested by d'Adda et al. (2016) who found no evidence to support it. Overall, we believe therefore that our treatment of norms in the Computer condition is legitimate.

In conclusion, we find that compliance to authority and learning how consistently others follow social norms are the most likely explanations behind prosocial and antisocial conformity. We hope that these findings will shed some light on the polarisation and viral diffusion of information online, that it will push towards a similar systematic exploration of preferences across other domains, and to a renewed interest in the cognitive and brain processes underlying these changes.

3 Meta-Context and Choice-Set Effects in Mini-Dictator Games

3.1 Introduction

The history of thinking about the nature of social behaviour in games can be seen in terms of the scope in which players in a game operate. In early days of game theory it was common to make the simplest assumption that players are maximizing their own self-interest (Morgenstern and Von Neumann, 1953). However, with the rise of experimental economics, a lot of evidence has emerged that challenged this hypothesis. Frameworks like Kahneman, Knetsch, and Thaler's Dictator game (Kahneman et al., 1986c) have contributed to the formulation of alternative models collectively known as social preferences (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). These models introduced the idea that, to explain deviations from the self-interest, the utility of an outcome should also include the payoffs of other individuals obtained in this outcome. Thus, the 'scope of operation' has been enlarged to include 'universal' considerations about others players' payoffs like inequity aversion. This process, however, did not stop there. Later experimental studies have shown the susceptibility of participants to experimental cues or context. Dictator game experiments that manipulated anonymity (Hoffman et al., 1996; Franzen and Pointner, 2012a) or choice set (List, 2007; Bardsley, 2007; Cappelen et al., 2013) have shaped the notion that the utility, while still seemingly modulated by other players' payoffs, also depends on the context of a game. For example, List (2007) and Bardsley (2007) show that providing dictators with the possibility to take money from the recipients makes them move away from the half-half split towards more selfish options, as if the complete set of possible outcomes played a role in participants' decisions. These results enlarged

the scope even further by showing that universal social preferences could not account for context effects and that the characteristics of a game at hand needed to be taken into account. The challenge of incorporating context was taken by the social norms literature where players maximize norm-dependent utility (Cappelen et al., 2007; López-Pérez, 2008; Krupka and Weber, 2009, 2013; Kessler and Leider, 2012; Kimbrough and Vostroknutov, 2016). Norms represent the rules of social conduct that prescribe the appropriateness of each action in a given setting. To the extent that the cues about an environment also inform about the norms, norm-dependent models have been able to explain the differences in behaviour due to context effects.[10]

The focus of some studies in this line of research is on the context meant as a frame in which the game is presented in the instructions (e.g., Liberman et al., 2004; Lightner et al., 2017; Chang et al., 2018; Kimbrough et al., 2018). Other studies manipulated the set of options in the Dictator game (DG), as in the above mentioned experiments, or the dictator's and recipient's initial endowments (e.g., Cox et al., 2016). While the meaning of context varies widely across all these studies, it is generally limited to the game played at the moment: participants are presented with one situation, and the influence of context is measured between-subjects.

What is missing in this conceptualisation, and what we want to address in our experiment, is that a game can be perceived as being part of a larger game that includes its outcomes as a proper subset. Knowing that some action is possible in principle, *even if there is no possibility to take this action*, could affect behaviour. In other words, a game can be perceived in terms of a 'meta-context' or a supergame that includes the current one, and, thus, can influence decisions. We conjecture that what is getting transferred from the meta-context are the norms, which define

[10]Social norms literature does not claim that people do not have social preferences, but rather that social preferences *emerge* as a consequence of following norms (Kimbrough and Vostroknutov, 2016).

the appropriateness of actions in the game that is actually being played.[11] Some support for this hypothesis comes from Thomsson and Vostroknutov (2017), where a 'constrained' DG is studied in which dictators cannot give more than half of their endowment to the recipient. The results show that dictators believe that equal split is the most appropriate action, as if they are considering the whole dictator game without constraints, which is also reflected in their behaviour.[12] Chlaß and Moffatt (2012) study choices in a sequence of Dictator games with varying choice sets. The authors find that, if in the first game participants are given an additional option to take from the recipient, the amount given in subsequent games is less likely to change (even when this take option is absent) than in a treatment in which this option is only introduced in later games. This suggests that participants use the early option to take as a reference for the subsequent games, thus creating a meta-context.

Despite our thinking that meta-context influences the behaviour in social situations, we do not believe that it is the only important factor. In particular, we propose that, in addition to meta-context, the specific choice set of a single game may also affect decisions. In fact, in order to detect any choice-set effects, we need meta-context as a benchmark, since otherwise it would be impossible to tell what is a choice-set effect and what is not.[13] Therefore, the second goal of this study

[11] We do not consider the possible effects of social preferences, since these models are based on outcomes rather than contextual information. This means that meta-context is irrelevant for the calculation of social utility (see Section 3.4 for discussion).

[12] Note that the subjects could have thought about the constrained DG with dictator's endowment X (where they cannot give more than $X/2$) as a DG with endowment $X/2$ plus a 'gift' of $X/2$ to the dictator. In this case it is reasonable that they should consider sending $X/4$ as a socially appropriate choice. However, a considerable fraction of subjects still chooses to send the amounts close to $X/2$, as Thomsson and Vostroknutov (2017) report. Interestingly, List (2007) studies a DG where subjects are told that they receive $X/2$ for themselves and another $X/2$ that they can share with the recipient. In this case few dictators send more than $X/4$. This shows that meta-context reasoning can be induced by simply manipulating instructions.

[13] We could have used social preferences models such as inequity aversion in place of meta-context. However, as we discussed above, choice-based or outcome-based preferences cannot account for any context effects, thus, what we could achieve at best is a result that, in some specific

is to explore whether and in what way a restricted choice set (compared to the supergame) can influence dictator giving. We are, of course, not the first to think about this problem. The studies mentioned above are, to an extent, concerned with the comparisons of Dictator games with varying sets of options. However, unlike these studies where two or three treatments are contrasted against each other, we aim at uncovering the systematic relationships between the allocations available in the restricted choice sets and the observed behavioural changes that could be generalised beyond Dictator games.

In our experiment we employ an extensive within-subject design in which participants choose in a series of two-alternative mini-Dictator games (mini-DGs) with all allocations of payoffs for a dictator and a recipient drawn from the 'supergame' set $\{(\pi_d, \pi_r) \mid \pi_d, \pi_r \geq 0 \text{ and } \pi_d + \pi_r = 60\}$. In other words, dictators always choose between two alternatives that lead to some divisions of 60 tokens (without the possibility of taking money from the recipient). Next, we use the task proposed by Krupka and Weber (2013) to elicit norms from the same participants in a subset of these mini-DGs. Finally, we estimate the propensity to follow rules by means of a rule-following task (Kimbrough and Vostroknutov, 2018b). In order to test the meta-context hypothesis, that participants derive the appropriateness of actions in mini-DGs to some extent from the standard DG (the meta-context), we check if the norms, elicited in mini-DGs, reflect the normative values generally obtained from the same task in the standard DG (e.g., Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016, 2018b). If the meta-context hypothesis holds, we should observe some consistency between the norms elicited in mini-DGs and the norms elicited in the standard DG. In order to disentangle choice-set effects from the meta-context we should consider two possibilities: 1) the elicited norms

supergame, there are some choice-set effects. This, though interesting, would not be generalisable to other environments.

do reflect choice-set dependence, in which case we can use the elicited norms to detect choice-set-dependent *and* choice-set-independent components; 2) the elicited norms do not reflect choice-set dependence, in which case we use the variance in the choice data, unexplained by the norm-dependent utility, to estimate choice-set effects. In both cases, we use the subject-specific normative evaluations to fit the parameters of a norm-dependent utility to the choice data and check how well the elicited norms can explain the behaviour.

The data from the norm-elicitation task confirm the meta-context hypothesis: participants do perceive the norms in binary mini-DGs as derived from a supergame. Interestingly, the normative ratings are so consistent with those in a standard DG, that we detect no choice-set dependence at all. Thus, participants express clear meta-context norms, which are not influenced by the choice set of the mini-DGs. The estimation of the norm-dependent utility shows that these normative ratings explain a sizeable portion of variance in choices. This being said, we also find that for a half mini-DGs the norm-dependent utility fails to completely explain the behaviour. The discrepancy in predictions is the largest for the mini-DGs in which the dictator receives very high payoffs in both allocations and recipient receives very low payoffs. Notably, the participants are more generous in these situations than the norm-dependent utility predicts. We show that, in order to account for this behaviour, we need to assume that participants care about the cost of choosing a non-selfish option, measured in percentages of their wealth, relative to the gain of the recipient, measured in percentages of her wealth. Put simply, participants are more generous when they lose relatively little in comparison with a large increase in recipients' wealth. Thus, we identify a very specific type of choice-set effect that works on top of the meta-context.[14] Our last finding concerns the na-

[14]To make sure that we are dealing with the choice-set effects and not with a misspecification of an outcome-based utility model, we estimate several other outcome-based models. None of them reconciles the observed behaviour.

ture of the choice-set dependency. We find that its strength is correlated with the individual measures of rule-following propensity, thus, supporting the hypothesis that choice-set effects are normative and do not operate through a separate channel, despite not being detected by the norm-elicitation task.

Our findings demonstrate that context can influence social decisions in a rather complex way. In particular, there are two types of normative reasoning that contribute to pro-social choices in a given situation. One is the 'absolute' normative component that comes from the meta-context and is independent from the choice set of a specific game. It determines how the appropriateness of an allocation is viewed on a larger scale irrespective of the specifics of choices involved. Another is the 'relative' normative component that elaborates on the first one by taking the available alternatives into account. Both components contribute substantially to the incentives that drive pro-social behaviour.

3.2 Experimental Design

Participants were recruited from the subject pool of the Cognitive and Experimental Economics Laboratory at the University of Trento and invited via e-mail. 166 subjects (93 female, mean age = 22) completed the experimental task. The study was approved by the University Ethical Committee. Experiment was programmed in z-Tree (Fischbacher, 2007). Experimental sessions were run in May 2017, February and March 2018. There were no pilots, and no data were discarded. The experiment consisted of three decision-making stages: the Rule Following (RF) task, the series of mini-Dictator games, and the norm-elicitation task. Participants earned on average €9.32 based on their choices, no show-up fee was included.

3.2.1 The Rule-Following Task

In the rule-following task (Kimbrough and Vostroknutov, 2018b), participants decided how to allocate 50 balls between a blue and a yellow bucket. The position of the two buckets was randomised across individuals. Participants earned €0.05 for each ball they dropped in the blue bucket, and €0.10 for each ball they dropped in the yellow bucket. The instructions explicitly stated that *'the rule is to put the balls into the blue bucket.'* This payment scheme and the rule were the only information provided to participants. The total payoff in this stage was the sum of earnings from both buckets. Therefore, the amount of money earned could vary from €2.50 (always following the rule) to €5.00 (never following the rule). The RF task creates a situation in which participants are asked to follow an arbitrary rule that decreases their payoff and yet entails no cost of breaking it. This allows us to measure the propensity of participants to stick to a non-social rule that, as was demonstrated by Kimbrough and Vostroknutov (2016), also predicts pro-social behaviour.

3.2.2 The Mini-Dictator Games

In Stage 2, participants played 182 mini-Dictator games with constant sum of payoffs. In each game each participant chose between two divisions of 60 tokens (1 token = €0.10) split between her and an unknown other. 27 different allocations were used in all mini-DGs, including an equal split. Out of them, 13 allocations were *advantageous* to the dictator (she received more than the recipient), and 13 allocations were *disadvantageous* (the dictator received less than the recipient). The allocations were combined to form a mini-DG according to the two criteria: 1) one of the two allocations has to be equal split or advantageous to the dictator; 2) if one allocation is disadvantageous, it has to be less unequal than the other (advantageous) allocation. These criteria yielded a total of 182 mini-DGs. The list of all mini-DGs can be found in Table 6.2, Appendix 6.3. Mini-DGs were presented to

each participant in an individually generated random order. After the task was completed, participants were randomly paired and one of them was selected as a dictator. Consequently, one of the choices of this individual was randomly implemented[15]. Participants were fully informed about all these procedures.

3.2.3 Norm Elicitation

In order to elicit participants' beliefs about norms we used the norm-elicitation task proposed by Krupka and Weber (2013). Participants were presented with a selection of 18 mini-DGs that they encountered in the previous task. They rated on a 4-point Likert scale the degree of social appropriateness of picking each alternative in each game. Each mini-DG was presented on a separate screen, and the order of their appearance was randomised. To detect choice-set effects, half of the allocations were repeatedly presented in combination with different alternatives. The list of all mini-DGs used for norm elicitation can be found in Appendix 6.3 Table 6.3. To incentivise precise answers, participants were rewarded by means of a coordination game. One option in one of the mini-DGs was randomly selected. If a participant's rating matched that of the majority in the session, she received €5.00.

3.2.4 Meta-context and choice-set predictions

We summarise here the main hypotheses and predictions concerning meta-context and choice-set dependence.

Norm Elicitation. We test two complementary hypotheses regarding normative ratings in the norm elicitation task, namely that A) elicited norms reflect the Meta-

[15]Given the large number of mini Dictator Games presented and the payment scheme, one could argue that some of the participants' choices might be driven by boredom or distraction. Although we cannot exclude that some choices might be driven by these factors, a data inspection suggests that most participants paid constant attention during the task (Supplementary Figure 6.7)

context, therefore appropriateness ratings of an allocation should remain the same regardless of the alternative; and B) elicited norms reflect choice-set dependence, therefore appropriateness ratings of an allocation should vary depending on the alternative. Based on previous findings Kimbrough and Vostroknutov (2018b), we also make the additional hypothesis that C) rule-followers and rule-breakers have a different perception of appropriateness of the actions in the mini-DG, and therefore should rate allocations differently in the elicitation task.

mini-Dictator Games. The main prediction that we make is that in the regression predicting participants' choices, D) coefficients capturing D1) meta-context and D2) choice-set dependence should be positive and significantly greater than zero. Secondly, we predict that even if they are not represented in the norm elicitation ratings, E1) Meta-context and E2) Choice-set dependence are normative in nature, and as a consequence the related coefficients in the regression should correlate with the participants' rule-following or rule-breaking propensity.

3.3 Results

3.3.1 Summary Results

We start with reporting some summary results in order to show consistency between the measures of rule-following and behaviour in mini-DGs. Figure 3.1A shows the histogram of rule-following propensity estimated by the RF task. Consistent with findings in different countries (Kimbrough and Vostroknutov, 2018b), there are participants who are strong rule-followers, strong rule-breakers, or have intermediate level of rule-following propensity. The proportion of strong rule-followers, who put all the balls in the blue bucket, is 20%, whereas the proportion of strong rule-breakers (all balls in the yellow bucket) is 13%.

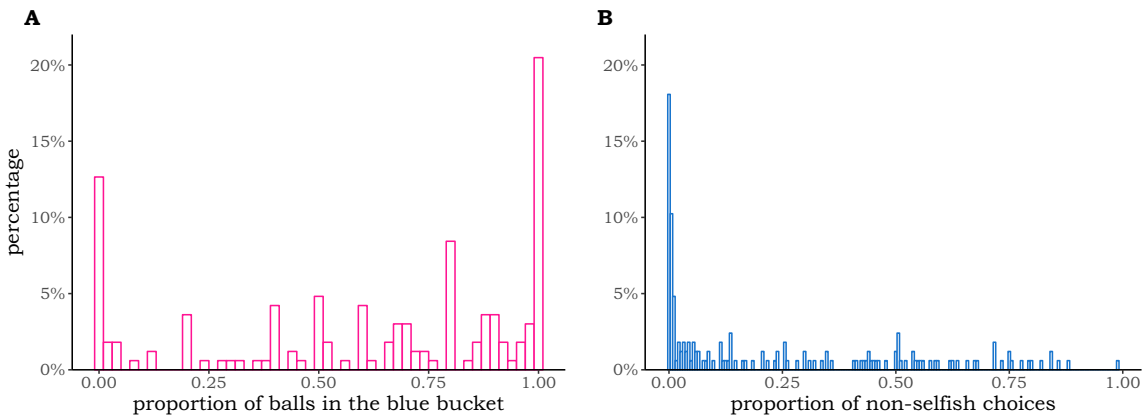

Figure 3.1: A. Histogram of the propensity to follow rules as measured by the proportion of balls in the blue bucket in the RF task. B. Histogram of the proportion of non-selfish choices in the mini-DGs.

Next, we look at the aggregate behaviour of participants in the mini-DGs. We calculate for each participant the proportion of non-selfish choices (participant chooses the allocation that gives her less payoff than the alternative). Figure 3.1B shows the histogram. These results cannot be directly compared to the standard Dictator game, as the choices in our task are between two allocations, and we employ multiple games rather than a single one. Hutcherson et al. (2015) and Kuss et al. (2015) use mini-DG tasks that are similar to ours. The two papers report the proportions of 21% (sd=18%) and 13.2% of non-selfish choices respectively. Our results are in line with these findings if not slightly higher, with an average of 23.9% (sd=27.3%) of non-selfish choices. Notice that more than half of the participants made at least 10% of non-selfish choices, and these choices are more likely when the cost of taking them is small (logistic regression of absolute difference in own payoffs on the probability of choosing a non-selfish option, $Z = 4.7$, $p < .001$). This shows that participants do respond to standard economic incentives.

We observe that the behaviour is very heterogeneous in both the mini-DGs and the rule-following task. Thus, we explore the relationship between individual rule-following propensity and the proportion of non-selfish choices (Figure 6.6 in Ap-

pendix 6.3 shows the scatter plot). Notice that rule-breakers on average choose selfish allocations 97% of the time, whereas for rule-followers this rate is only 58%. Spearman's rank correlation shows that this relation is significant ($\rho = 0.42$, $p < .001$). This association between rule following and dictator giving confirms the findings of previous experiments (Kimbrough and Vostroknutov, 2016, 2018b) and provides a good indication that the behaviour in mini-DGs is related to participants' sensibility to norms.

3.3.2 Meta-Context

While the RF task is a simple measure of the propensity to follow non-social rules, behaviour in the Dictator game setting is arguably driven by social inputs. In order to understand, from the normative perspective, whether mini-DGs are considered as standalone games or as a part of a meta-context, we elicit the beliefs about socially appropriate behaviour using the norm-elicitation task. Figure 3.2 plots the appropriateness ratings of the 19 allocations presented during the task averaged across subjects.[16] Mean, standard error, and median ratings for each allocation in each mini-DG are presented in Table 6.3 in Appendix 6.3. It is clear that participants' ratings show the same pattern that emerges when the norms are elicited in standard DG, namely, the half-half allocation is considered the most appropriate, while the more unequal allocations are increasingly less appropriate. This demonstrates that, normatively, participants treat mini-DGs as a part of a meta-context, in this case standard DG, in which all actions are available. In support of this finding, notice that average ratings show little variability, which means that participants mostly agree on the appropriateness evaluations.

[16]Some allocations were presented multiple times in pair with different allocations. In this plot we average across multiple presentations.

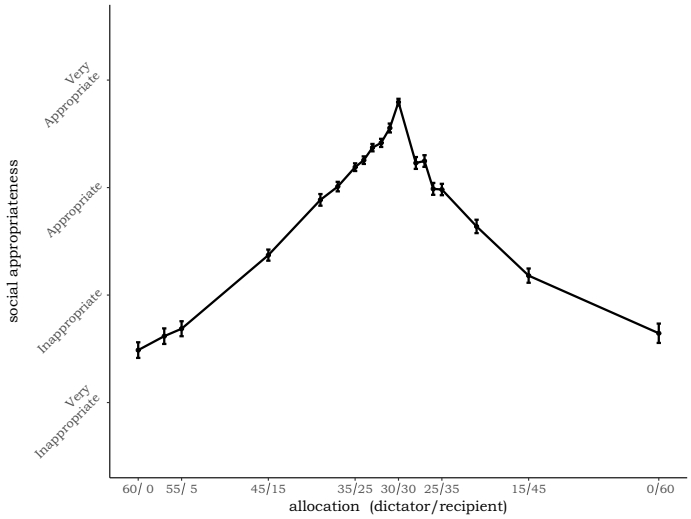

Figure 3.2: Average appropriateness ratings across allocations. On the x-axis: Dictator's payoff/Recipient's payoff.

Despite the similarities with norm elicitations in standard Dictator game, we observe that the shape of the curve is slightly different. We expected to observe an asymmetry in ratings between advantageous (more money to oneself) and disadvantageous (more money to the other) splits, with the former being rated as less appropriate than the latter (Kimbrough and Vostroknutov, 2016; Krupka and Weber, 2013; Chang et al., 2018). This, however, is not the case, as the two types of splits are almost perfectly symmetric. Another aspect that differs from previous experiments is that the increase in appropriateness from less to more appropriate allocations is much less convex than in other studies. We do not have a precise explanation for these differences, but speculate that restricted choice set could have played a role.

We also check whether perception of appropriateness differs between rule-followers and rule-breakers. We divide the sample by median of rule-following propensity and run one Wilcoxon rank-sum test for each allocation presented, with participant's type as a grouping variable. Unlike Kimbrough and Vostroknutov (2018b), we find that appropriateness ratings are not significantly different (all

$p > .05$). This suggests that, while choices in the mini-DGs range from completely selfish to very generous, participants are all similarly aware of what the normative expectations are. Finally, we test for the presence of the choice-set effects, namely, that allocations, which were repeatedly presented in different mini-DGs, are rated differently depending on the alternative. None of the rating differences are significant even at the 10% level (Kruskal-Wallis test).

All these findings taken together provide a strong evidence that participants derive normative values of allocations in mini-DGs from a single meta-context, which is independent of the allocations in specific games. We further discuss the implications of this result in Section 3.4.

3.3.3 Norm-Dependent Utility Estimation

The next step in our analysis is to estimate how well norm-dependent utility, which uses the information about elicited norms, fits the observed choices. Given that the appropriateness beliefs do not depend on the rule-following propensity or choice set, we use the ratings from the allocations in the norm-elicitation task to interpolate the appropriateness of all other dictator allocations in the mini-DGs. Having an appropriateness rating for all allocations would allow us to use them in a utility function. We estimate a 'norm function' as a piece-wise power curve for advantageous and disadvantageous allocations, fitted separately for each participant. The parameters $a_{adv}, b_{adv}, c_{adv}, a_{dis}, b_{dis}, c_{dis}$ of the power curves are estimated using non-linear fitting of the form:

$$N(\pi_d, \pi_r) = \begin{cases} a_{adv} + b_{adv} \cdot E^{c_{adv}} + \varepsilon_{adv} & \text{if } \pi_r \geq 30 \\ a_{dis} + b_{dis} \cdot E^{c_{dis}} + \varepsilon_{dis} & \text{if } \pi_r \leq 30 \end{cases}$$

The dependent variable $N(\pi_d, \pi_r)$ represents the social appropriateness ratings of the 19 allocations. The values of $N(\pi_d, \pi_r)$ are in the set $\{-1, -\frac{1}{3}, \frac{1}{3}, 1\}$, where -1 stands for 'very inappropriate' and 1 for 'very appropriate.' The independent variable E is a standardised measure that orders allocations by their equality, ranging from 0 (highest inequality, e.g., 60/0 or 0/60) to 1 (complete equality, e.g., 30/30).[17] Errors are assumed to be normally distributed. A parameter summary for the power curves (Table 6.4) together with a fitting example (Figure 6.8) are presented in Appendix 6.3.

Next, we compare advantageous and disadvantageous curves at group level. Analysis of the power parameters c_{adv} and c_{dis} reveals that ratings for disadvantageous allocations are more convex (i.e., differences in appropriateness are more accentuated) than for advantageous allocations (median $c_{dis} = 2.29$, median $c_{adv} = 1.49$; Wilcoxon signed rank test with continuity correction, $V = 2927$, $p < .001$). This difference means that as the disadvantageous options become more unequal, they decrease in appropriateness faster than advantageous options. This is in line with observations we made in regard to Figure 3.2. Neither a_{adv}, a_{dis} nor b_{adv}, b_{dis} differ significantly between the two allocation types.

With an appropriateness rating for each allocation at hand, we estimate a utility function for participants' choices. We model the utility of participant i using the specification by Kessler and Leider (2012): $U_i(\pi_d, \pi_r) = \beta_i \pi_d + \gamma_i \cdot N_i(\pi_d, \pi_r)$. This model assumes that participants derive utility from their own gains (first term) and from the norm compliance (second term). The normative component of the utility is a norm function weighted by γ_i, which defines the propensity of the participant

[17] Since we fitted a power curve to both advantageous and disadvantageous allocations, we decided to use the ratings for 30/30 allocation for both curves. Given that we need only one estimate for each allocation, we decided to use the one from the advantageous curve, since the 30/30 split was only paired with other advantageous allocations. We, nonetheless, tested whether estimates from the two curves differed in their predictions using a paired t-test, and found no significant difference ($t(165) = -1.54$, $p > .1$).

to follow the social norm. We estimate parameters β_i and γ_i by means of a logistic regression assuming random utility, namely that the probability of choosing the option with the highest utility is proportional to the utility difference between the two options (McFadden, 1976). We estimate the following regression:

$$\Pr(\text{non-selfish choice}) = \alpha_i + \beta_i \cdot (\pi_d^{nonself} - \pi_d^{self}) + \gamma_i \cdot (N_i^{nonself} - N_i^{self}) + \varepsilon$$

$\pi_d^{nonself} - \pi_d^{self}$ is the difference in dictator payoffs between the non-selfish and the selfish allocation. $N_i^{nonself} - N_i^{self}$ is the difference in social appropriateness ratings between the two options, computed as explained above. We also estimate a constant α which captures the general tendency to choose non-selfish option.

	Quartiles				
Coefficient	First	Median	Third	Rank-sum test	Pr(non-selfish)
α_i	–4.04	–1.38	–0.21	$p < 0.001$	-3.026^{***}
β_i	0.00	0.02	0.13	$p < 0.001$	0.021^{***}
γ_i	0.00	1.27	3.45	$p < 0.001$	1.265^{***}
N observations					30,212
N groups/subjects					166

Table 3.1: Summary of the coefficients estimates from individual regressions and a random effects logit regression on all data (errors are clustered by subject). In individual regressions, the participants who only made selfish choices are excluded. *** stands for $p < 0.001$.

Table 3.1 shows the summary of the estimates of the coefficients across participants. Notice that the distribution of α_i lies mostly in the negative domain with zero being in the 4th quartile, and the distributions of β_i and γ_i are in the positive domain with zero in the 1st quartile. Rank-sum tests show that all three coefficients at sample level are significantly different from zero. The same conclusion can be made from a random effects logit regression that includes all data (last column of Table 3.1): the coefficients on personal utility, norm utility, and intercept are highly significant and have the predicted signs. We conclude that the choices

of the participants are driven by the norm-dependent utility maximization and, thus, by meta-context since the elicited norms used in the estimation conform with those in the standard DG.

As an additional test of consistency of the choices with norm-following, we test our hypothesis that γ_i represent participants' rule-following propensities and, thus, should be correlated with the estimates obtained in the RF task. Indeed, Spearman's rank correlation between γ_i and the number of balls put in the blue bucket is $\rho = .29$ ($p < .001$). We also find that γ_i correlates with the proportion of non-selfish choices (Spearman's $\rho = .44$, $p < .001$), which provides additional support to the hypothesis that participants choose non-selfish options because they follow norms.

Our analysis up to this point has provided support for the norm-dependent utility specification, with norms coming from meta-context, and has shown that the rule-following propensity can explain the non-selfish choices in mini-DGs. However, the utility function that we estimated is essentially outcome-based since the normative term assigns one value to each allocation based on the meta-context and does not take into account the possible influence of the unchosen option.[18] Therefore, this analysis does not capture any choice-set dependence that could still be an important factor that drives participants' choices. In the following sections we analyse regression residuals to see how well the norm-dependent utility, specified above, describes the data and check whether choice-set dependent utility is needed to improve the fit.

[18] In the logistic regression the two utilities are compared. However, this dependency models the choice process and not the changes in utility due to the presence of another option.

3.3.4 Anomalies in Non-Selfish Choices

In order to understand how well the elicited norms explain behaviour, we look at the norm differences in the subsets of mini-DGs in which the change in personal payoff between the two alternatives is fixed. According to the norm-dependent utility, in such cases only differences in norms should influence the probability of non-selfish choices. Figure 3.3A plots $N_i^{nonself} - N_i^{self}$, the difference in norms between the non-selfish and selfish options, as dependent on the lower payoff in the *advantageous mini-DGs* with payoff difference equal to 5.[19] In these trials the model predicts that the more equal the payoffs are in each allocation, the more appropriate it is to choose the non-selfish option. For instance, according to the normative ratings, choosing 30/30 over 35/25 is considered more socially appropriate (average norm difference $\approx$ 0.38) than choosing 55/5 over 60/0 (average norm difference $\approx$ 0.18).

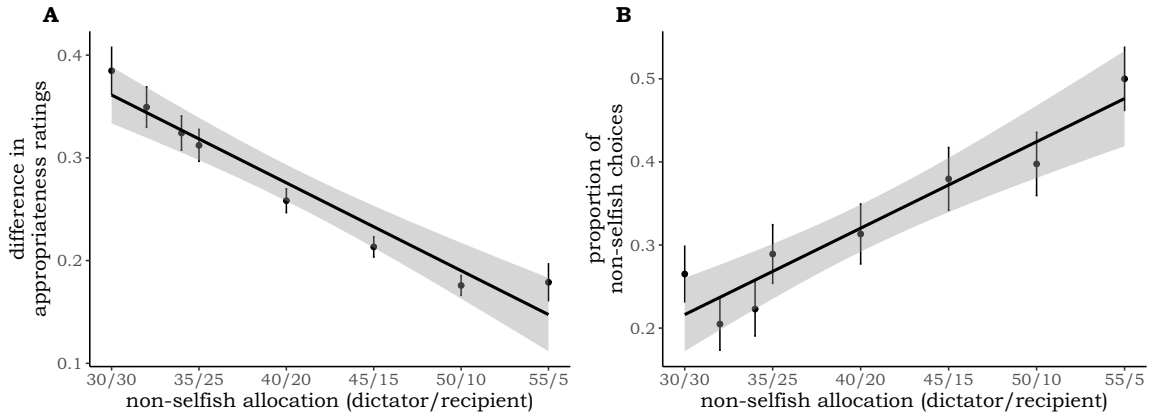

Figure 3.3: Advantageous mini-DGs with payoff differences equal to 5: A. Norm differences as dependent on the non-selfish dictator's payoff; B. Proportion of non-selfish choices as dependent on the non-selfish dictator's payoff.

If we assume that participants do care about complying with norms—and this is supported by the positive coefficients on the normative term in the regressions—

[19] We will distinguish between advantageous mini-DGs in which both allocations are advantageous and disadvantageous mini-DGs in which one of the allocations is disadvantageous.

participants should be less selfish in mini-DGs with more equal options as compared to mini-DGs with less equal options. What we observe instead is that participants display the exact opposite behaviour. Figure 3.3B shows that they are *less* selfish when alternatives are *less* equal. The same is true when we analyse the mini-DGs with other fixed payoff differences (Figure 6.9 in Appendix 6.3).[20]

This discrepancy should also be reflected in the regressions. We check whether the residuals show an anomalous trend (i.e., depend monotonically on payoff when they should not) for advantageous mini-DGs.[21] We analyse the residuals using locally weighted scatter-plot smoothing, which plots residuals against the payoffs in the non-selfish option ($\pi_d^{nonself}$ and $\pi_r^{nonself}$). As it can be seen in Figure 3.4, the residuals show a positive trend: a linear regression confirms that the tendency is significant (residuals as dependent on non-selfish payoff, $p < .001$, see Column 1 in Table 6.5, Appendix 6.3 for the regression).[22]

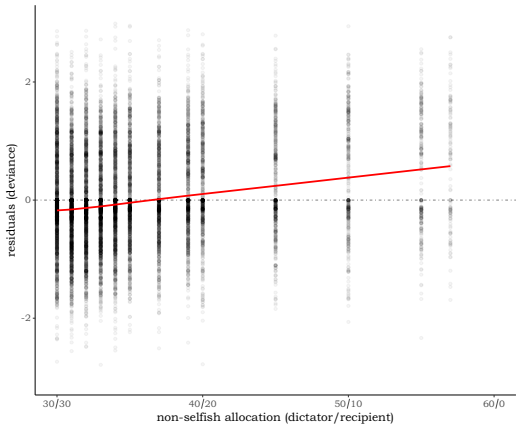

Figure 3.4: Locally weighted scatter-plot smoothing of deviance residuals across all individual regressions.

[20]This relationship is not present among the disadvantageous mini-DGs, where meta-context norm predictions match participants' behaviour.

[21]We drop participants who never chose a non-selfish option from this analysis because they are unaffected by any characteristics of the task.

[22]We also check the residuals for disadvantageous mini-DGs but do not find any significant anomaly for the current nor for the other regressions (see below). A plot for the disadvantageous mini-DGs residuals is presented in Appendix 6.3, Figure 6.10. See Column 7 in Table 6.5, Appendix 6.3, for the regression.

There are two reasons why we may observe the trend in residuals. One is the misspecification of the norm-dependent utility. It is possible that the shape of the norm function that we get from the data is influenced by the design and does not reflect correctly the perceived appropriateness of allocations. For example, in Figure 3.3 the increase in the proportion of non-selfish choices would be qualitatively consistent with the change in differences of appropriateness ratings if the norm function were concave instead of convex. The second reason is that the misspecification could originate from the *choice-set dependence* that outcome-based utility functions are not able to capture. In what follows we examine the first possibility (misspecified norm function) by fitting other outcome-based utility functions to see if the trend in the residuals disappears. Specifically, we test a polynomial utility and a concave norm utility.

The regression with polynomial utility tests the idea that *linear* norm-dependent utility has not enough degrees of freedom to explain participants' behaviour. We allow personal payoffs and the norm function to interact in a non-linear fashion. We test a set of such models which contain polynomial terms up to degree 4. According to the Bayes Information Criterion (BIC), the best model is that of degree 3 (BIC: 18153), which shows that the increase in degrees of freedom does not improve the fit of the utility after some point. Figure 6.11 in Appendix 6.3 shows the residuals of the winning model: the trend is still statistically significant ($p < .001$, see Column 2 in Table 6.5, Appendix 6.3, for the regression). This analysis demonstrates that the complexity of the model does not help to explain away the trend in residuals.

The regression with the concave norm utility is based on the assumption that the trend in the residuals is the result of a distortion in our measurement of social appropriateness. It might be the case that the procedure we used to elicit appropriateness ratings provides a biased estimate of actual beliefs. For instance, the

complete equality allocation (30/30) might be particularly salient to participants, so much that it is reported as being much more appropriate than even a slight deviation from it (e.g., 35/25), which in turn is rated as significantly less appropriate.[23] As a result, the norm function is 'forced' to be convex, while it may actually be the case that participants do not really consider the allocations around equal split as being much normatively distant from it. Along the same lines, participants assign almost the same ratings to giving little (e.g., 55/5) and giving nothing (e.g., 60/0) to the recipient, whereas in reality these two options may be perceived as very different. Such bias, created by the norm-elicitation task, can, in principle, lead to the utility misspecification. To test whether we can eliminate the trend in the residuals, we replace the norm function estimates with a fixed concave norm function.

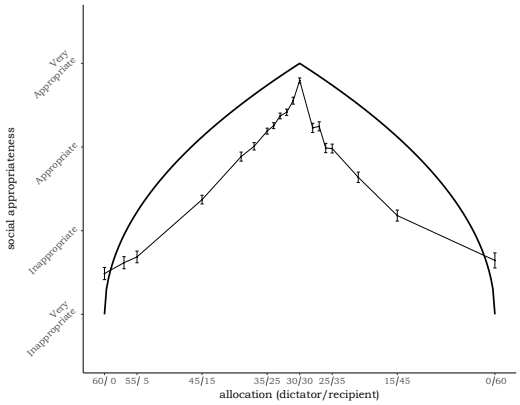

Figure 3.5: Concave norm function overlapping the data from norm-elicitation task.

We model the concave norm as a power function and fix the three sets of parameters: $a = -1$, $b = 2$, $c \in \{0.1, 0.5, 0.8\}$. Figure 3.5 illustrates the concave norm function with $c = 0.5$. Unlike in our data, here the decrease in appropriateness is less steep around the 30/30 allocation and more steep around very unequal allocations (60/0, 0/60). For simplicity, the same values are used for both advantageous and disadvantageous allocations. For all three values of c the trend in residuals

[23] This might be the case because the norm elicitation task is essentially a coordination game in which allocation 30/30 is focal (Sugden, 1995).

remains (for $c = 0.5$ Figure 6.12 in Appendix 6.3 shows the plot). The coefficients from the linear regressions of residuals on non-selfish payoffs are significantly different from zero for all three models ($p < .005$, see Columns 3-5 in Table 6.5, Appendix 6.3, for the regressions). Interestingly, the extremely concave norm function with $c = 0.1$ does explain well the choices between very unequal allocations, but misses out on the ones close to the equal split, while the least concave function with $c = 0.8$ does the opposite.

Overall, this analysis shows that the outcome-based utilities that we considered are not very good at explaining the observed behaviour.

3.3.5 Choice-Set Dependence

In the previous section we tried to correct for the trend in residuals with two outcome-based utility specifications. However the problem persisted. Therefore, we try a new approach that takes into account the possible dependence of the utility of each option on the alternative, in other words, choice-set dependence. We hypothesise that the decision to choose a non-selfish option is modulated by its *relative cost*, which depends on how much the dictator loses (in percent) and how much the recipient gains (in percent) if the non-selfish option is chosen. Note that relative cost violates an important tenet of the outcome-based utility models where it is assumed that decision makers first attribute a value independently to each alternative and then compare these values to select the optimal option. Relative cost, instead, implies that the value of each allocation depends on the other available alternative.

To represent relative cost, we first notice that the dictator's direct cost of choosing the non-selfish option is $d = \pi_d^{self} - \pi_d^{nonself} > 0$, which is also the gain of the recipient $d = \pi_r^{nonself} - \pi_r^{self}$, since the sum of payoffs for each option is constant.

We define the percent losses of the dictator and gains of the recipient as

$$C_d = \frac{d}{\pi_d^{self}} \text{ and } C_r = \frac{d}{\pi_r^{nonself}}.$$

For the dictator, C_d is the loss from choosing the non-selfish option relative to the highest payoff. For the recipient, C_r is the gain measured relatively to the recipient's highest payoff (when the non-selfish option is chosen by the dictator). We hypothesise that dictator's decision depends on $C_r - C_d$[24]. For example, for the choice between 60/0 and 55/5, by choosing the non-selfish option the dictator loses only around 8% of his maximal payoff, while the recipient gains 100%. In such situation the dictator might be willing to be more generous than when the difference in percentages is smaller. We are not the first to notice that relative cost might play a role in dictator games. For example, Smeets et al. (2015) analyse sharing decisions by millionaires and find that they are extremely generous when they are paired with a low-income recipient: around half of them give the entire €100 that they are endowed with. This effect might also manifest itself as *maximin preferences*, when participants sacrifice payoff efficiency for the sake of higher payoff for the 'poorest' participant (Engelmann and Strobel, 2004; Baader and Vostroknutov, 2017). Moreover, our formulation is akin to context-sensitive models in psychology, such as the one by González-Vallejo (2002); indeed, relative cost could be seen as a special case of González-Vallejo's model, where the two attributes participants consider are the gains for themselves and the gains for the recipient.

We estimate the following regression:

$$\Pr(\text{non-selfish choice}) = \alpha_i + \beta_i \cdot (\pi_d^{nonself} - \pi_d^{self}) + \gamma_i \cdot (N_i^{nonself} - N_i^{self}) + \delta_i \cdot (C_r - C_d) + \varepsilon.$$

[24]Notice that both C_r and C_d are computed using the payoffs from *both* allocations. Thus, the dependence of dictator's preference on $C_r - C_d$ cannot be represented by any outcome-based social utility.

The results are presented in Table 3.2. We see that the distributions of α_i, β_i, and γ_i are very similar to those from the individual regressions without relative cost in Table 3.1. The same holds for the logit regression on all data shown in the last column. The individual values of δ_i are mostly positive, which is supported by the rank-sum test and the significance of the coefficient in the all-data logit regression. Thus, the probability of non-selfish choice increases when $C_r - C_d$ is large or, in other words, when both allocations in a mini-DG give most of the payoff to the dictator.

	Quartiles				
	First	Median	Third	Rank-sum test	Pr(non-selfish)
α_i	–5.06	–1.91	–0.45	$p < 0.001$	–3.509***
β_i	0.00	0.04	0.16	$p < 0.001$	0.034***
γ_i	–0.00	0.66	3.14	$p < 0.001$	0.609***
δ_i	0.00	2.69	6.50	$p < 0.001$	4.073***
N observations					30,212
N groups/subjects					166

Table 3.2: Summary of the coefficients estimates from individual regressions and a random effects logit regression on all data (errors are clustered by subject). In individual regressions, the participants who only made selfish choices are excluded. *** stands for $p < 0.001$.

Introducing the relative cost term to the regression eliminates the trend in residuals ($p = .97$, linear regression in Column 6, Table 6.5 Appendix 6.3), which suggests that the model successfully captures the features of participants' behaviour (Figure 3.6). Even though some of participants' choices are explained by the normative term in the utility, there is a significant portion of variance not captured by it, but which is explained by the context-dependent term.

As a control measure, we check whether introducing the relative cost term affects the correlation of the norm coefficients γ_i with the rule-following propensity and the proportion of non-selfish choices. The correlation with the number of balls in the blue bucket is weakened, but still significant (Spearman's $\rho = .17$, $p = .031$),

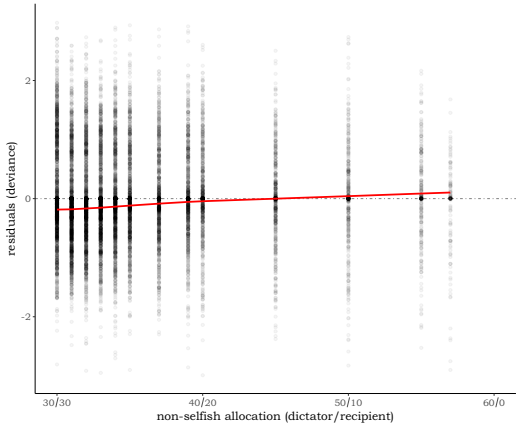

Figure 3.6: Locally weighted scatter-plot smoothing of deviance residuals across all individual regressions with relative cost.

as is the correlation with the proportion of non-selfish choices (Spearman's $\rho = .17$, $p < .033$).

These results demonstrate that the choices in mini-DGs are significantly affected by the context. In particular, the dictators become more generous when their 'wealth' is much higher than that of the recipients.

3.3.6 The Nature of Choice-Set Dependence

The results of the previous section show the presence of the choice-set effect on participants' generosity. However, it remains unclear whether this effect is part of norms-driven behaviour or is a separate phenomenon. More precisely, there are two possibilities: 1) the choice-set effect is working independently from norms, which are correctly represented by the ratings obtained in the norm-elicitation task; 2) the choice-set effect is related to norms but is not captured by the normative ratings. The latter possibility might, in turn, stem from two sources. It can be that our norm-elicitation task in not flexible enough to capture the choice-set effect or that participants, when evaluating the appropriateness of different allocations, do not consciously perceive it. Unfortunately, with our design we cannot tell these

two apart (see Section 3.4 for discussion); however, we can definitely test whether choice-set effect is related to norm-following or not.

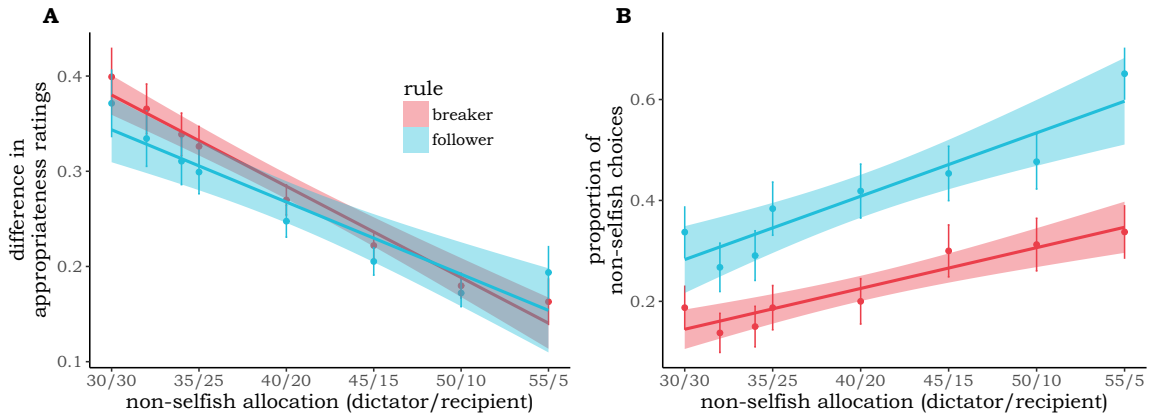

Figure 3.7: Advantageous mini-DGs with payoff differences equal to 5: A. Norm differences as dependent on the non-selfish dictator's payoff; B. Proportion of non-selfish choices as dependent on the non-selfish dictator's payoff.

Our previous analysis has shown that the rule-following propensity, estimated in the RF task, is a good proxy for the coefficient γ in the norm-dependent utility function: it correlates with the proportion of non-selfish choices and with the individual γ_i coefficients obtained from the norm-dependent utility regressions. Thus, in order to test whether choice-set effect is norm-related or not, we can use these estimates to see if there is a difference in the strength of the effect between rule-followers and rule-breakers.

Figure 3.7 shows the data presented in Figure 3.3 separately for rule-breakers and rule-followers, the two groups of participants defined by the median split of the number of balls in the blue bucket in the RF task. We see that the normative differences are the same for both groups: this is not surprising given that normative ratings, as we have seen, are not statistically different between the two groups. However, rule-followers are more generous than rule-breakers overall and show stronger reaction to the relative cost $C_r - C_d$, which increases in dictator's payoff (the slope for rule-followers is higher). This suggests that rule-followers are

more influenced by the choice set than rule-breakers, which supports the hypothesis that choice-set dependence is norm-related. This finding is also supported by the correlation between the rule-following propensity and the coefficients δ_i from the individual regressions with relative costs discussed in the previous section (Spearman's $\rho = 0.20$, $p = 0.009$). Positive correlation means that the more rule-following a participant is, the more she reacts to the relative cost by increasing the probability of choosing the non-selfish allocation.

	residuals	SE
intercept	–0.630***	(0.056)
follower	–0.285***	(0.080)
$\pi_d^{nonself}$	0.018***	(0.002)
follower $\times \pi_d^{nonself}$	0.011***	(0.002)
N observations	30,212	
N groups/subjects	166	

Table 3.3: OLS regression of the residuals from the logit regression in Table 3.1. *** stands for $p < 0.001$.

To support these results even further, we take the residuals (deviance) from the random effects logit regression reported in Table 3.1 and regress them on the non-selfish payoffs of the dictator, the dummy for rule-followers, and their interaction. Table 3.3 shows that the interaction term is significant and large, in comparison with the coefficient on the payoff. Thus, the trend in residuals is stronger for rule-followers, which means that the choice-set effect is more pronounced for them. Figure 6.13 in Appendix 6.3 shows this result graphically.

We can conclude that the choice-set effect that we observe in our data is definitely related to the rule-following propensity, and, thus, to norm-following in general, even though our norm-elicitation task does not capture it. In the next section we discuss the possible implications of this finding.

3.4 Discussion

Summary of the results. The goal of the research project reported in this paper was to study the influence of meta-context and choice set on pro-social behaviour. Our original intuition was that by measuring the normative perceptions of actions in different mini-Dictator games we would have been able to observe the influence of the choice-set context and, consequently, explain the context-dependent behaviour. However, what we uncovered is that the normative ratings of allocations in mini-DGs, instead of being linked to the specific choice set of each game, reflect *meta-context*, which derives from a superset of hypothetical alternatives (a standard DG). The ease and unanimity of participants in expressing the normative values as coming from a meta-context suggest that it is natural for them to think in such terms. More importantly, we demonstrate that these meta-context considerations do explain a substantial portion of variance in the choices and, therefore, should be taken into account in normative models of social behaviour (e.g., Kimbrough and Vostroknutov, 2018a).

This being said, we also find that in some circumstances the observed choices do not conform with the meta-context norms that our participants agree on. In particular, when the payoffs of a dictator are much higher than those of a recipient in both allocations of a mini-DG, there is a discrepancy between the normative evaluations and the proportion of generous choices: participants behave much more pro-socially than what meta-context predicts. We demonstrate that this bias can be explained by assuming that there is an influence of choice set on participants' decisions, which is expressed as a *relative cost* of being selfish. To strengthen this result, we show that several outcome-based utility specifications, which have many degrees of freedom but do not consider the choice-set effects, fail to explain this behaviour. Finally, we find strong association between this choice-set effect and

the individual propensity to follow rules, which suggests that the influence of the choice set is *normative* in nature.

Overall, we found strong support for our original hypothesis that choice-set dependence in mini-DGs is related to normative considerations. However, the situation is more complex than we thought. There are *two normative components* that shape pro-social behaviour. One, derived from meta-context, represents an 'absolute' normative value of each allocation, irrespective of other available options. Another, based on choice set, is a 'relative' normative value determined by the comparison of an allocation with the alternatives available in the specific situation.

Meta-context and outcome-based social preferences. The idea that some social decisions are explained by the presence of an absolute normative value, which is a function of an outcome and not of the choice context, may sound similar to certain definitions of outcome-based social preferences (e.g., inequity aversion). However, we would like to point out that this resemblance is superficial. The meta-context of a mini-DG that we talk about comprises all hypothetically conceivable allocations *of a specific type*. Since the participants only choose among divisions of 60 tokens, the meta-context is a standard Dictator game with possible allocations $\{(\pi_d, \pi_r) \mid \pi_d, \pi_r \geq 0 \text{ and } \pi_d + \pi_r = 60\}$. Notice that this would be different if participants were sometimes also presented with an option to take money from the other player. In this case, meta-context would include additional allocations where dictator can earn more than 60 tokens. As is well-known from the studies by List (2007), Bardsley (2007), and Cappelen et al. (2013), in such circumstances the choices become much less generous than in the standard DG setting, and this is exactly what we would expect if participants were presented with taking options.[25] Thus, meta-context represents all possible allocations similar to those ex-

[25]Similar line of reasoning about changing meta-context can be applied to an ongoing debate regarding the availability of punishment options in experiments. Some researchers claim that the

perienced in the experiment and not all allocations that can happen in principle. In this sense, it is still a *context* that does not apply in all situations. Following this mode of thinking, the immutable universal social preferences may be seen as coming from some grand meta-context that covers all possible kinds of social interactions. It may well be that this is indeed the case, and that such considerations do play a role in social decisions (see Kimbrough and Vostroknutov, 2019, for discussion). However, given their universality, it is very plausible that they will be overshadowed by more specific contextual effects.

Norm elicitations and choice-set dependence. As we mentioned in Section 3.3.6, there is an unclarified issue related to the data obtained from the norm-elicitation task. Why is it that participants, when asked to guess the prevalent social norms in mini-DGs, express normative values for each allocation as if they were coming from a meta-context? Why do not they combine the absolute and relative normative components in their evaluations? Unfortunately, our design does not allow us to answer these questions directly. Nevertheless, we can discuss the possible reasons behind this phenomenon.

The simplest possibility is that the norm-elicitation procedure itself precludes participants from correctly expressing their normative views. The fact that the 18 mini-DGs used for norm elicitation are presented one after another might be conducive to giving the same rating to an allocation, regardless of the alternative option. This effect might be exacerbated by the coordination game structure of the task: it might be tempting to think that everyone in a session will give the same rating to each allocation to simplify the coordination problem.

ubiquity of punishment observed in, say, repeated Public Goods games is due to an experimenter demand effect created by the fact that only punishment is available. According to this argument, rewards should also be possible in order to eliminate the demand effect. In our opinion it is not that simple. The introduction of rewards would change the nature of the meta-context, which can lead to completely different social attitudes than in case when only punishment is available.

On a deeper level, this problem might have nothing to do with the norm-elicitation task per se, but rather with how the normative evaluation is performed by the brain. It is possible that people, when asked to give appropriateness ratings, process information differently from the situations when they actually have to make normative decisions. In the former case, they might tend to provide general judgements related to the meta-context of a given situation without thinking about specifics, while in the latter they start taking choice set into account since, after all, they must make a choice. This distinction may be similar to attentional mechanisms responsible for the visual exploration, which operates on two levels: a general first impression is quickly constructed, and then, if necessary (e.g., when a choice should be made), a more detailed analysis is performed in order to elaborate on the first impression (Duchowski, 2007). We hope that future experiments will clarify the situation around the normative perceptions of decisions affected by a choice-set context.

Theoretical and experimental implications. Finally, we discuss some implications of our study for theoretical modelling and design of experiments. If, as we suspect, the influence of meta-context and the choice set spreads beyond mini-DGs, then it becomes important to appropriately model their effects.

From the theoretical perspective, it would be interesting to understand what exactly constitutes a meta-context for each given environment. Another natural application is in dynamic games: from the perspective of a player who acts in a subgame, the outcomes of the entire game can be viewed as meta-context (see Kimbrough and Vostroknutov, 2018a). Experimentally, sensitivity to meta-context implies that care should be taken in within-subject designs where participants are presented with a series of similar tasks. In such cases the knowledge of the past possibilities can affect how the actions in a game at hand are viewed normatively (e.g., Chlaß and Moffatt, 2012).

The main challenge for choice-set dependence is to develop a method of incorporating the alternatives into the utility function. Kimbrough and Vostroknutov (2018a) propose one such theoretical approach. They consider the same norm-dependent utility specification that we used in this study, but calculate the norm associated with each allocation as dependent on all other allocations that can be realised in a game. Thus, the utility of each allocation stops being solely outcome-based, as now it depends on the alternatives through the norm function. It should be mentioned that in their model meta-context can also be easily accounted for, though the authors do not suggest how it should be chosen.

Another important line of future investigation is norm-elicitation tasks. As we mentioned earlier, it remains unclear whether the original task proposed by Krupka and Weber (2013) creates some biases in elicited norms or not. For example, it might be the case that the convexity of the norm functions in DG is an artefact of the elicitation method (see Section 3.3.2), or that seeming context-independence of the elicited norms is a consequence of participants' attempts at coordination. We are currently working on a new norm-elicitation task that will rectify these and other potential issues.

3.5 Conclusion

After three decades of experiments with the Dictator game (DG) it may seem that the motives of experimental subjects are pretty well understood. Nevertheless, we uncover two effects that, to our knowledge, were not studied together in the previous literature. In our experiment each participant plays 182 mini-Dictator games with two actions and fixed payoff efficiency. After that we ask participants to express their normative evaluations of the actions in 19 of these mini-DGs.

The first effect that we discover is what we call meta-context. Specifically, the probability of the non-selfish choice in a mini-DG is modulated by the normative

reasoning related to the supergame that contains it, in our case, a standard DG. This means that participants take into account the appropriateness of the available actions as if these actions were part of the standard DG. We conclude this from the observation that the normative ratings, which participants express in the norm-elicitation task, do not depend on the choice set and conform with similar elicitations obtained in a standard DG. Moreover, these normative evaluations explain a sizeable portion of variance in choices.

The second effect is choice-set dependence. While meta-context accounts for some pro-social choices, we find a systematic deviation from its predictions in mini-DGs that give very high payoff to the dictator in both available allocations. In such games participants behave more pro-socially than they would if only meta-context mattered. We show that this is explained by assuming that the utility of each choice depends on the alternative. This dependency can be expressed as a *relative cost* of non-selfish choice. In particular, participants seem to take into account how important the difference in payoffs between the two allocations is for them and for the recipient in terms of their respective wealths.

Our final finding is that the choice-set effect is not independent from the propensity to follow rules. Thus, the two effects that we observe represent two distinct components of the normative value that participants take into account. This demonstrates that pro-sociality is a complex phenomenon that in each particular case depends on the specifics of the situation and a general context to which this situation belongs.

4 Fairness Bias in Memory Representation of Social Decisions[26]

4.1 Introduction

The study of memory suppression has a long history in psychology, and the search for its bases dates back to the 19th Century (Weiner, 1968). People tend to remember less correctly information that is threatening to the self (Sedikides et al., 2004), and negative information is processed in a more shallow manner and recalled more poorly than positive experiences (Sedikides and Green, 2000). Numerous theories have been proposed to explain suppression of unpleasant memories: that negative information conflicts with a person's expectations (Mischel et al., 1976) or with a positive self-image (Swann Jr et al., 2003), or that mis-remembering follows from a failure in emotion regulation (Mather, 2006). Such explanations, however, failed to hold in a series of experiments (Sedikides and Green, 2009), that instead point to the idea that people forget harmful memories to protect their self-image.

This line of research has mainly explored how memory biases originate from outside threats (e.g., negative feedback from others). More recently however, research has shown that this phenomenon can also arise from internal conflicts: people can indulge in dishonest behaviour (e.g., cheating, stealing) that does not fit with their personal beliefs of fairness and, as a result, they maintain less vivid memories of ethically relevant material (Shu et al., 2011; Shu and Gino, 2012). Biased remembering also occurs when participants are asked to recall their fairness-violating choices (Kouchaki and Gino, 2016), and even when remembering is incentivised (Carlson et al., 2018; Saucet and Villeval, 2019).

[26]Joint work with Nadège Bault and Joshua Zonca.

Internally driven memory biases may be related to *moral disengagement*, namely the inhibition of psychological mechanisms controlling ethical reasoning (Bandura, 1986; Detert et al., 2008). Moral disengagement arises after an individual behaves against her own moral principles, and can take different forms: *post-hoc* rationalisation of one's own actions, offload of responsibility to others or dehumanisation of the victim. When disengagement is not possible, the conflict between actions and thoughts causes cognitive dissonance (Festinger, 1962). The person can then try to alleviate this discrepancy by either adjusting her behaviour, or by modifying her beliefs. For instance, after they are given the chance to cheat in a game, participants can adjust their concept of fairness to consider low rates of cheating as morally acceptable (Mazar et al., 2008). Memory biases of own choices could thus act in the opposite direction by distorting not the concept of fairness, but rather the memory of fairness violations, engaging a similar self-protective mechanism to the one in memory biases following external judgements.

Although previous research has shed some light on why people engage in biased forgetting, several questions remain unanswered; in this paper, we address two of them. First, due to discordant results, it remains unsettled whether forgetting affects objectively[27] unfair decisions or whether forgetting is mediated by personal fairness beliefs. In other words, where is the moral touchstone? Do people evaluate their own choices based on external shared norms, or rather on internalised moral beliefs? Second, while research has focused on the effect of memory biases on the immoral action itself, it could actually be possible that the bias also influences other aspects of a memory. For instance, does the bias affect relevant or even irrelevant information about the context in which the immoral decision was made? As research on eyewitness testimony suggests, this is not a trivial ques-

[27] We borrow the improper term 'objective' from Carlson et al. (2018) to refer to a fairness evaluation that can be quantified numerically compared to a standard.

BOX 4.1: MEMORY BIASES IN THE BRAIN

Neuroscientific studies have explored the mechanisms of memory suppression employing psychological paradigms such as directed forgetting, where participants are explicitly asked not to remember particular stimuli they were previously presented with (Anderson and Hanslmayr, 2014). Several relevant findings come from this literature. Inhibitory mechanisms related to motivated forgetting share similar features to motor suppression (Anderson, 2003), suggesting that such mechanisms are physiologically ingrained. Motivated forgetting is also an effortful act that involves several frontal areas of the brain generally associated with active action control. Activity in the dorsolateral prefrontal cortex (DLPFC), for instance, has been shown to exert regulatory control on hippocampus activity, that in turn correlates with successful encoding of new information (Rizio and Dennis, 2013; Paller and Wagner, 2002; Hanslmayr and Staudigl, 2014). A similar inhibitory mechanism seems to occur also at memory retrieval (i.e. when information is sourced from memory) and has been associated with both dorsal and ventral areas of the lateral prefrontal cortex (Depue et al., 2007; Benoit and Anderson, 2012; Anderson and Levy, 2009). While this evidence pictures a structured brain network for intentional forgetting, none of the paradigms used is able to capture the underlying reasons why participants suppress unwanted memories (Anderson and Hanslmayr, 2014).

Additional insights on the mechanisms underlying memory bias could come from brain studies on cognitive dissonance. The anterior cingular cortex (ACC, Kitayama et al., 2013) seems to be associated with detection of internal conflicts, while the DLPFC appears to be involved with the adjustment of how choices and beliefs are represented (e.g., Izuma et al., 2015). In addition, after taking a conflicting choice, the preference for the options seems to change–along the related neural activity–arguably to justify the decision made (Voigt et al., 2019). Crucially, this change in mental representation occurs only when the choice is actually remembered, suggesting that forgetting can be an alternative mechanism to deal with decisions that risk of creating moral dissonance.

tion. Consider for instance the weapon focus effect, namely the reduced accuracy in the recollection of an event caused by the presence of an emotionally-charged stimulus, such as a weapon in a crime, drawing the bystander's attention (Loftus et al., 1987; Kensinger and Ford, 2020). It is possible, then, that a personal moral violation could have a similar attenuating effect on the perception of contextual information that would otherwise be encoded in memory.

To respond to these questions, we adopt a between-subjects experimental design. Participants first play a resource-allocation game in which they can trade off personal gains to either increase or decrease earnings of another unknown participant. Later, after a distractor task, participants are asked to recall some information related to their choices: either what their choices were (choice), the alternatives between which they chose (relevant context), or an unrelated feature of the decision (irrelevant context). We denote fairness violations based on two opposing definitions, subjective and objective (Carlson et al., 2018). The subjective measure derives from a self-report questionnaire at the beginning of the experiment that elicits participants' personal beliefs about what is considered a fair choice. The objective measure is based on the notion of equity: namely, that a fair choice is one that reduces differences in payoffs between people. We then test if participants' choices in the game are more in line with their reported fairness principles or with the more general equity principle. We count fairness "violations" as choices that go against either of these definitions and then test if a memory bias is associated with either subjective violations or objective violations.

Our results suggest that participants' subjective principles are mostly in line with the general notion of equity, and that there thus is no substantial difference between the two measures. We find evidence of a memory bias, which leads to less-accurate memory of choices associated with fairness violations, but only for information related to the choice. While we find no support for a memory bias in the context conditions, we suspect that this lack of evidence is due to the difficulty of the memory tasks used in these conditions compared to the one in the Choice condition. Hence, conclusive evidence is missing on whether memory of contextual information could be biased or not. Lastly, based on mixed experimental evidence, we propose an additional memory bias mechanism unrelated to moral reasoning but rather associated with probabilistic reasoning, pushing participants to over-

represent in their memory the particular type of choice that they made most often during the game. Further evidence however is required to support this hypothesis.

4.2 Methods

4.2.1 Participants

Participants were recruited through the online platform Prolific Academia, with no restrictions set for participation. The experiment was conducted between the end of April and the beginning of June, 2019. 181 participants (age 28 ± 9 (S.D.), 68 females) took part in the experiment. We excluded participants' data if they: reported in a post-experimental questionnaire that they took notes during the experiment; reported different demographics than those stated on the recruitment website (e.g., gender); took more than 60 minutes to complete the experiment; chose the same alternative 80% or more of the time in the resource-allocation game (choosing the alternative on the left/right side of the screen) and memory task (giving the same response, e.g. "strong yes"). 11 participants were excluded for these reasons, so experiment analyses were conducted on 170 participants. The experiment lasted 35 minutes on average. All procedures were approved by the University of Trento Ethical Committee. All participants gave their informed consent for participating in the experiment. The study was pre-registered on the open science framework where supplementary materials can also be found (osf.io/g5u73/); a summary of the original hypotheses and amendments to the pre-registered protocol are presented in Supplementary Material 6.4.1.

4.2.2 Experimental Procedure

The experiment had a between-subject design, therefore participants were randomly assigned to one of three experimental conditions: Choice ($N = 59$), Rel-

evant Context ($N = 57$), and Irrelevant Context ($N = 54$). Participants first responded to a self-report evaluation of fairness perception, then completed a resource-allocation game with another participant and last, after a distractor task, performed a memory task specific to the experimental condition. The independent variable differing between conditions was the type of information from the resource-allocation game to be recalled in the memory task: either what choices the participants made (Choice condition, forced-choice recognition task), the alternatives between which they had to choose (Relevant Context, old-new recognition task), or the unrelated symbols that were displayed right before each choice (Irrelevant Context, old-new recognition task). With the exception of the initial fairness self-report, participants were paid for their choices in all tasks (average earnings: £3.39 in Choice, £2.78 in Relevant Context, and £2.90 in Irrelevant Context conditions), as well for their participation (£2.50/hour). Participants were informed of the incentive schemes in the instructions before each task. All tasks were coded in JavaScript using lab.js (Henninger et al., 2019).

Fairness Self-report. Our first research question asks which behaviour, if any, could be perceived as a threat to self-image, either not behaving equitably *per se* (objective violation), or perception of own behaviour mediated by some internal moral compass (subjective violation). To measure subjective violations, at the beginning of the experiment we elicited participants' personal beliefs about what does or does not constitute a fair distribution of resources. The elicitation procedure was adapted from the circle test (Sonnemans et al., 2006, Figure 4.1). Participants considered a hypothetical situation similar to the one in the ensuing resource-allocation game: they were in charge of distributing some money between them and another person. Participants chose between 29 allocations of money, ordered from "selfish" (higher earnings for oneself, £2.00 versus £0.50) to increasingly "prosocial" (higher earnings for the other participant, £1.00 ver-

sus £1.50, Supplementary Table 6.6). Participants were then asked to select, from selfish to prosocial, the first allocation that "the other person would consider an acceptable offer". Crucially, the allocations from which they chose were some of those employed in the resource-allocation game, scaled down into pounds[28]. We were thus able to compare participants' subjective beliefs of fairness to their actual choices in the resource-allocation game by transforming participants' answer in the self-report to the equivalent in the game (4.2.3).We call the chosen allocation the "fairness reference point," or subjective moral standard of the participant.

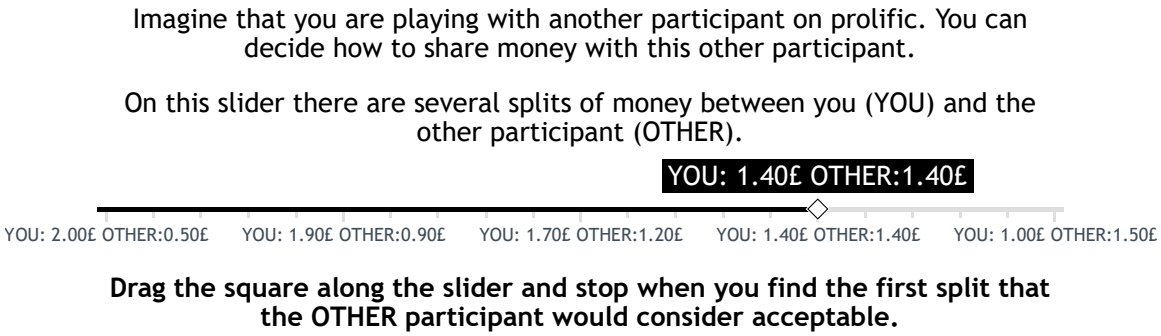

Figure 4.1: Fairness self-report measure. Participants chose their reference point by dragging the square along the slider. In this example the participant selected the half half split as her fairness reference point.

Resource-Allocation Game. After responding to the self-report, participants played the resource-allocation game (Figure 4.2). The game consisted of 50 two-alternative forced choices in which they were asked how they would prefer to allocate points (1 point = £0.02) between themselves and an unknown other (Supplementary Material 6.4.2 and Supplementary Spreadsheet at osf.io/g5u73/). 46 out of 50 choices offered a trade-off in gains between the participant (selfish option) and the other person (prosocial option), with the more prosocial allocation being also more equitable in terms of point distribution. The remaining 4 choices contrasted the proso-

[28]Allocation correspondence was not disclosed to participants as it could have influenced their subsequent responses in the resource-allocation game. We could not find any other design solution that did not involve omission of this particular information.

cial allocation with an antisocial option that left both participants worse off than the alternative. We included these trials to detect any potential antisocial participants (those who prefer to decrease the other person's earnings even if at a cost to themselves) that would have otherwise been mistaken as selfish. At the end of the experiment, participants were randomly paired, one participant in each pair was randomly selected, then one of her decisions was randomly selected and implemented for payment. Points for self and for the other ranged from 2 to 160 points, with payments ranging from £0.04 to £3.20.

Shape-detection Task. Presentation of the allocations in the resource-allocation game was anticipated by the appearance of a symbol on the screen (Figure 4.2). In the Irrelevant Context condition, participants were further instructed to press the space bar if and only if any feature of the symbol included at least a closed shape. Participants earned £0.50 at the end of the experiment if their button decisions to press (or not to press) were correct in at least 90% of the trials. The shape-detection task served to prompt participants to pay attention to the symbols–which was the information to be recalled in the Irrelevant Context condition–and thus to recall them correctly in the subsequent memory task.

Distractor Task. Between their decisions and the moment of recollection, participants played an unrelated distractor task that served as a time gap necessary for the encoding of memory. The distractor task was a two-dimensional mental rotation task (Cooper, 1975). In each trial, participants chose which of two objects was a rotated version of a target object. Participants earned £0.50 if their answers were correct in 40 or more comparisons within a 5-minute window.

Memory Task. After the time gap, participants played a memory task specific to their experimental condition. Participants in the Choice condition played a forced-choice recognition task: they had to retrace their 50 choices in the resource-

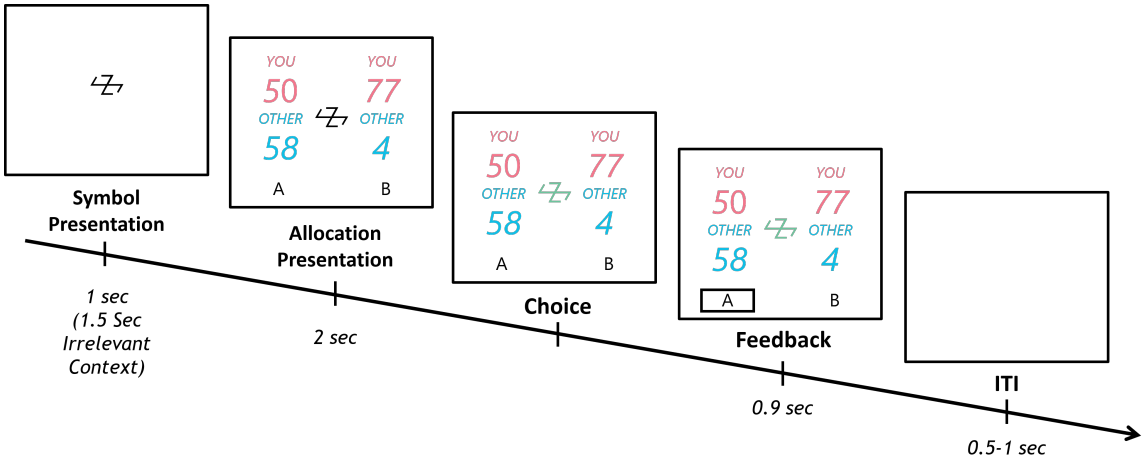

Figure 4.2: Trial structure in the resource-allocation game. Each decision was preceded by a black symbol appearing at the centre of the screen (Supplementary Spreadsheet at osf.io/g5u73/ for a list of symbols). After 1 second (1.5 seconds in the Irrelevant Context condition), the two allocations appeared next to the symbol. After 2 more seconds, the symbol turned green to signal participants that they could make their decision. The symbol had two purposes: it served as a fixation point before the presentation of the allocations, and represented the cue that signalled participants when they could make their choice. Participants were thus forced to observe the allocations for 2 seconds, a duration we assumed was sufficient for participants to memorise the allocations. Participants had no time restraints to give their answer. The trial ended with a randomly jittered inter trial interval (ITI) of 0.5 to 1 seconds. In the Irrelevant Context condition, participants were asked to press a button if and only if the symbol preceding the allocations featured at least a closed shape. Participants could respond during the 1.5 seconds in which the symbol appeared alone before the allocations.

allocation game by recalling which of the two alternatives (left or right) they had preferred in each trial. Participants in the Relevant Context condition played an old-new recognition task: they observed 30 allocation pairs from the game and 20 new pairs (Supplementary Material 6.4.6), answering for each pair whether they had or had not seen it during the game. Similarly, participants in the Irrelevant Context condition also played an old-new recognition task: observing 30 symbols from the resource-allocation game and 20 new symbols (Supplementary Material 6.4.5), and answering whether they had or had not seen the symbol during the game. Participants' responses were not time-constrained. They could respond with either of two buttons for each possible answer, indicating strong (weak) con-

fidence in their memory. Confidence served to test whether participants' memory mistakes were simply the result of guessing, due to low confidence, rather than of the memory bias associated with fairness violations that we aimed to observe.

4.2.3 Statistical Analyses

Fairness measure. To measure subjective violations, we employ participants' answer in the fairness self-report. We compute the distance between the allocation chosen in the self report and the two allocations seen in each trial of the game (Supplementary Method 6.4.4). If the participant chose the more distant (i.e. less subjectively fair) allocation, we counted this choice as a subjective fairness violation. To measure objective violations of fairness, by contrast, we considered choices in the resource-allocation game in which the participant preferred the allocation with the greater difference in payoffs between herself and the other(19):

$$\max\left(|\pi_o^1 - \pi_y^1|, |\pi_o^2 - \pi_y^2|\right), \tag{19}$$

Where π_y and π_o are points for the dictator (you) and for the other, respectively, in the two allocations, indicated by superscripts 1 and 2. The objective measure is based on the concept of equity[29], namely the idea that fairness is based on an even distribution of resources among players (Fehr and Schmidt, 1999).

Statistical tests. We use both measures of fairness violation to detect the presence of a memory bias. We conduct analyses both within and between subjects. To test whether choices within the same subject leading to a fairness violation are recalled worse than other choices, we run a series of mixed-effects logistic regressions with

[29]Notice that more equitable allocations in the game also present higher earnings for the other; this is because less-equitable allocations in the game are never advantageous to the other player, as we assumed that participants would never choose such an option.

memory accuracy (dummy: 1 = correct; 0 = incorrect) as dependent variable and participant ID as random intercept. We compare regressions of different predictors, including fairness violation, confidence, and choice frequency (i.e. whether the participant violated or upheld the fairness measure in most trials).

To test differences across participants (between subjects), we run two types of analyses. The first analysis divides participants in each condition based on the number of fairness violations using a median split and then tests differences in overall memory accuracy between the two sub-samples (Wilcoxon rank sum tests). The second analysis instead concerns only the Choice condition, where we test whether the proportion of recalled fairness violations is smaller than the actual proportion of fairness violations made (Wilcoxon signed-rank test). We set for all tests the significance threshold α to 0.05, two-tailed. In case of correction for multiple comparisons, we employ a Benjamini-Hochberg correction (Benjamini and Hochberg, 1995). Statistical tests are conducted using base R (R Core Team, 2018) and ggstatsplot (Patil, 2018). Non-parametric statistics are log-transformed for conciseness. Table 6.6 summarises the three dimensions along which the experiment design varies.

Memory task	choice options (Rel. Context)	symbol (Irrel. Context)
Fairness violation	objective (equity)	subjective (self-report)
Analyses	between-subjects	within-subjects

Table 4.1: Dimensions of the experimental design: participants took part in one of the three conditions; fairness violations were measured using two opposing definitions; lastly, analyses of memory bias were conducted both between and within subjects.

4.3 Results

Fairness Self-report. 39% of participants reported a fairness reference point that favoured themselves, whereas only three participants (2%) gave an answer that

favoured the other. A considerable majority (59%) reported the most equitable allocation (you: £1.40, other: £1.40) as their reference point. Strong homogeneity in answers suggests that fairness beliefs are widely shared among the sample.

Resource-Allocation Game. On average, participants chose the more equitable allocation 66% (SD: 30%) of the time. 118 participants (around 70%) chose the more equitable allocation in at least half of the trials or more, 21 of which (12%) chose the more equitable allocation in all trials. In comparison, only 9 participants (5%) always chose the most selfish allocation and only 2 (1%) always chose the most disadvantageous option for the other (antisocial). If instead of considering the proportion of equitable choices (objective measure) we measure the proportion of choices in line with their self-reported fairness reference point (subjective measure), we observe similar numbers: 138 participants (81%) chose consistently with their own moral standards in at least half of the trials, 20 of which (12%) never violated their own moral standards. The number of objective and of subjective fairness violations are actually correlated according to Spearman's rank correlation test ($\rho = .66$, $p < .001$, Figure 4.3): the less equitably a person behaves, the less consistently she will stick to her fairness beliefs. This association is partly driven by the fact that fairness beliefs are similar across participants (i.e., the equitable allocation is the fairness reference point for most). Furthermore, it turns out that results are similarly significant or not significant regardless of which of the two measures we employ. For this reason, we choose to report results based only on the objective fairness violation (i.e. based on the number of equitable choices).

Memory Task. Memory accuracy differs between conditions: while, on average, participants in the Choice condition recalled their choices correctly 87% (SD = 14%) of the time, performance was much lower in the Relevant (M = 62%, SD = 13%) and Irrelevant (M = 60%, SD = 14%) Context conditions, in which old-new recog-

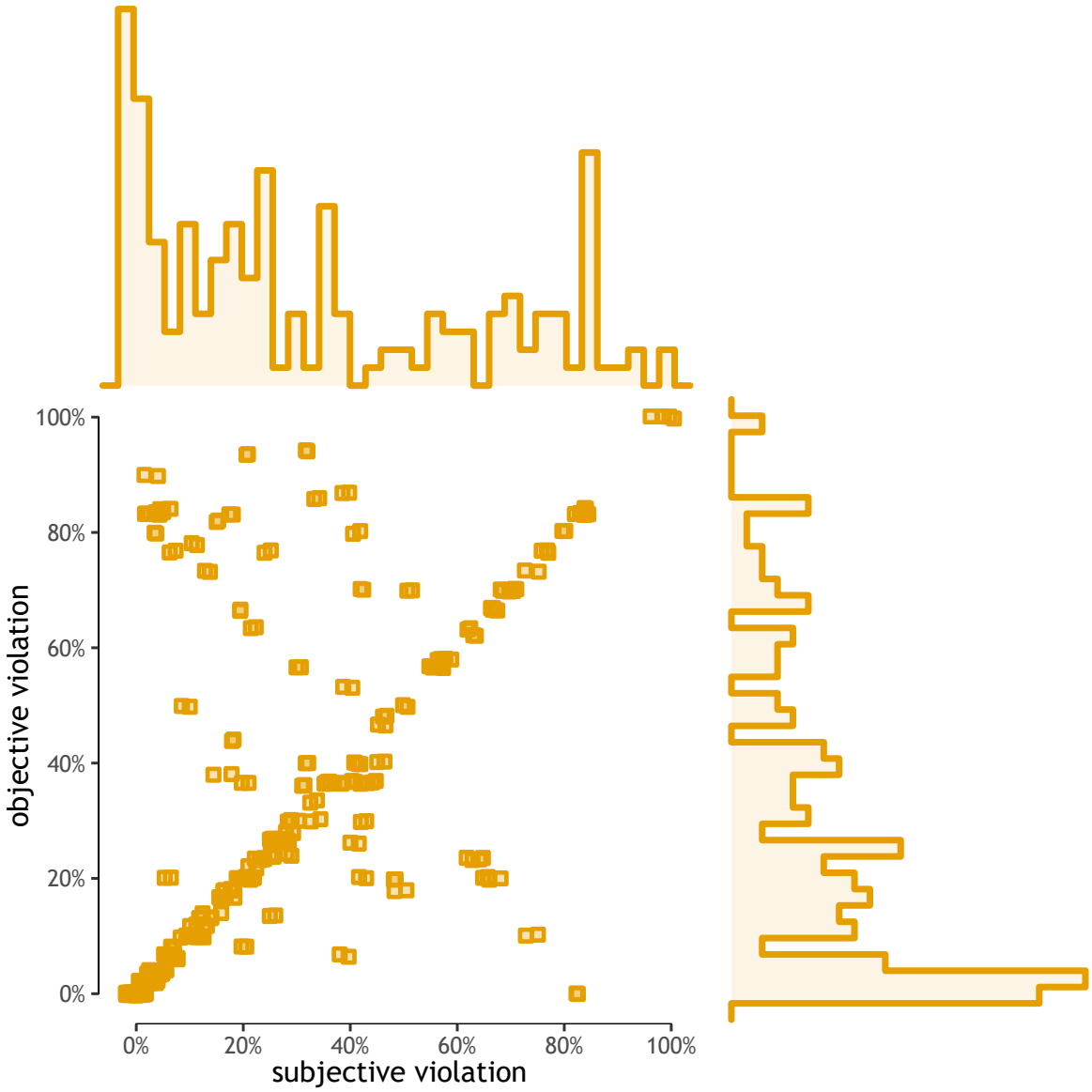

Figure 4.3: Association between subjective and objective measures of fairness violations. The marginal histograms represent the proportion per paticipant of fairness violations according to the subjective fairness reference point (top) and to the objective proportion of inequitable allocations chosen (right). The dot plot represents the joint distribution of the two measures, each square representing a participant.

nition was used (ROC curves are presented in Supplementary Result 6.4.7). These numbers are mirrored by the proportion of participants with an accuracy above chance: 95% in the Choice condition, 81% in the Relevant Context condition, 74% in the Irrelevant Context condition. We also measured the percentage of responses given with high confidence: 77% (SD = 23%) in the Choice condition, against only

36% (20%) in the Relevant Context and 44% (18%) in the Irrelevant Context conditions. Confidence and accuracy are actually strongly correlated according to a Spearman's rank correlation test ($\rho = .55$, $p < .001$, data pooled from all conditions). In the following sections, we test whether there is a fairness memory bias common to all (within subjects) or specific to some participants (between subjects) and, if so, what features of memory it affects (choice, relevant or irrelevant context).

Memory Bias Within Participants. We test whether information related to fairness violations (e.g., choosing a selfish or antisocial allocation) is recalled less correctly than information related to equitable choices. We employ a mixed-effects logistic regression with fairness violation as a predictor dummy variable (1 = equitable choice; 0 = fairness violation), participant id as random intercept, and whether the information was correctly recalled (1 = correct; 0 = incorrect) as the dependent variable. Fairness violation has a significant effect on memory accuracy in the Choice condition, but not in the context conditions (Choice: $\beta = 1.155$ (SE: .142), $z = 8.131$, $p < .001$; Relevant Context: $\beta = .031$ (.117), $z = .263$, $p = .793$; Irrelevant Context: $\beta = .123$ (.122), $z = 1.015$, $p = .310$; Table 4.2).

We test whether the memory bias is robust to confounding variables. We first test whether the memory bias in the Choice condition is driven by the considerable proportion of participants that chose equitably in all trials of the game, or whether the non-significant results in the context conditions are due to the high percentage of participants with memory performance below chance. The results however hold in both cases when excluding such participants (Choice: $p < .001$, Relevant Context: $p > .05$, Irrelevant Context: $p > .05$). Next, we introduce two other variables into the regression. The first variable that we consider is confidence. Participants in fact could have been less confident when recalling a fairness violation; this lack of confidence could, in turn, explain the lower memory accuracy in these

Condition		Fixed Effects				Random Effects			Model fit (R^2)		Sampling Units		
		β	S.E.	z	p		Var.	S.D.	marg.	cond.	$N_{obs.}$	$N_{part.}$	N_{trials}
Choice	Intercept	1.61	.200	8.021	<.001***	Participant (Intercept)	1.486	1.219	.053	.348	2950	59	50
	Fairness Violation	1.16	.142	8.131	<.001***								
Relevant Context	Int.	.469	.104	4.506	<.001***	Part. (Int.)	.153	.391	.000	.044	1710	57	30
	F.V.	.031	.117	.263	.793								
Irrelevant Context	Int.	.341	.110	3.11	.002**	Part. (Int.)	.204	.452	.001	.059	1620	54	30
	F.V.	.123	.122	1.02	.310								

Table 4.2: Results of the mixed-effects logistic regressions of fairness violation over memory accuracy, by experimental condition. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

trials. Secondly, participants could recall fairness violations less well because most participants in the sample rarely made selfish or antisocial choices. We therfore consider choice frequency as the second variable, which we define as the type of choice (equitable/violation) that the participant made more frequently during the game (dummy variable: 1 = frequent; 0 = not frequent).

Choice	LR Test	F.V.	Confidence	F.V.×Conf.	Frequency
Fairness Violation		<.001***	-	-	-
Confidence	<.001***	<.001***	<.001***	-	-
F.V.×Confidence	.011*	<.001***	<.001***	.597	-
Frequency	<.001***	.023*	<.001***	-	<.001***
Relevant Context	LR Test	F.V.	Confidence	F.V.×Conf.	Frequency
Fairness Violation		.793	-	-	-
Confidence	.012*	.877	.013*	-	-
F.V.×Confidence	.339	.538	.028*	.370	-
Frequency	.144	.609	.014*	-	.145
Irrelevant Context	LR Test	F.V.	Confidence	F.V.×Conf.	Frequency
Fairness Violation		.310	-	-	-
Confidence	<.001***	.096	<.001***	-	-
F.V.×Confidence	.300	.053	<.001***	.311	-
Frequency	.549	.120	<.001***	-	.548

Table 4.3: Comparison of the mixed-effects logistic regressions, by condition. On the left column are the different model versions tested, whereas on the top row we have the Likelihood-Ratio test ('LR Test') for model comparison and the models' parameters ('F.V.' = fairness violation; 'F.V.×Conf.' = interaction between confidence and fairness violation). *: $p < .05$; **: $p < .01$; ***: $p < .001$.

We run a series of regressions including these variables, also considering a possible interaction between fairness violation and confidence (Table 4.3). Results remain largely consistent irrespective of the regression considered: Confidence is significant in all three conditions, suggesting that guessing plays a meaningful role in explaining memory accuracy, whereas the interaction between fairness vio-

lation and confidence is not significant. In the Choice condition, choice frequency also seems to play a role in memory accuracy. Even after introducing these two variables however, the fairness violation parameter still remains significant. These results suggest that fairness violation is a significant predictor of memory accuracy even if we control for guessing (low confidence) or the frequency of a choice. We proceed to test whether there are differences between participants in memory bias, based on the overall proportion of fairness violations.

Memory Bias Between Participants. We categorise participants in each condition based on the number of fairness violations made in the game and then split the samples based on the median value. We first test whether participants with a proportion of fairness violations strictly higher than median (fairness violators) recalled information worse than participants with a proportion of fairness violations strictly lower than median (fairness upholders, Figure 4.4). Since data are not normally distributed (Shapiro-Wilk tests for normality, $p < .05$), we employ a Wilcoxon rank-sum test to compare the groups. Memory accuracy is significantly worse for fairness violators than for fairness upholders in the Choice condition ($log(V) = 4.60$, $p < .001$, $r = .64$, $n = 55$) but not in the Relevant ($log(V) = 5.83$, $p = .956$, $r = -.01$, $n = 52$) nor in the Irrelevant ($log(V) = 5.96$, $p = .683$, $r = -.06$, $n = 54$) context conditions. These results hold even after excluding either participants below chance recall rates or participants that chose only equitably (Choice: $p < .001$, Relevant Context: $p > .05$, Irrelevant Context: $p > .05$).

We also consider, for the Choice condition, whether the proportion of recalled fairness violations is lower than the proportion of fairness violations actually made. We compute the difference in proportions for each participant and test the sample distribution against zero by means of a Wilcoxon signed-rank test (Figure 4.5A). The test is not significant at sample level ($log(V) = 6.19$, $p = .555$, $r = .08$, $n = 59$); If, however, we divide the sample between fairness violators and fairness uphold-

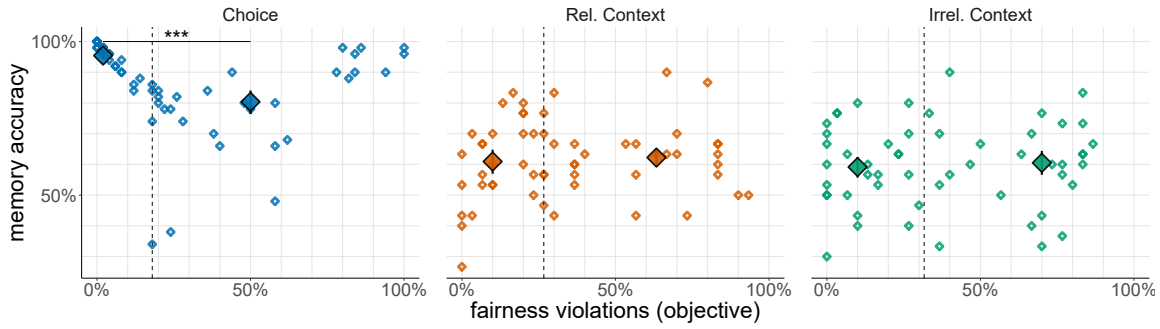

Figure 4.4: Memory accuracy, by percentage of fairness violations: participants in each condition are divided based on a median split (dashed line). Error bars indicate *t*-adjusted, 95% Gaussian confidence intervals. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

ers using the median split, fairness upholders do significantly recall having made fewer violations than they actually made ($log(V) = 3.09$, $p = .010$, $r = .49$, $n = 28$), whereas fairness violators do not ($log(V) = 5.21$, $p = .590$, $r = -.11$, $n = 27$).

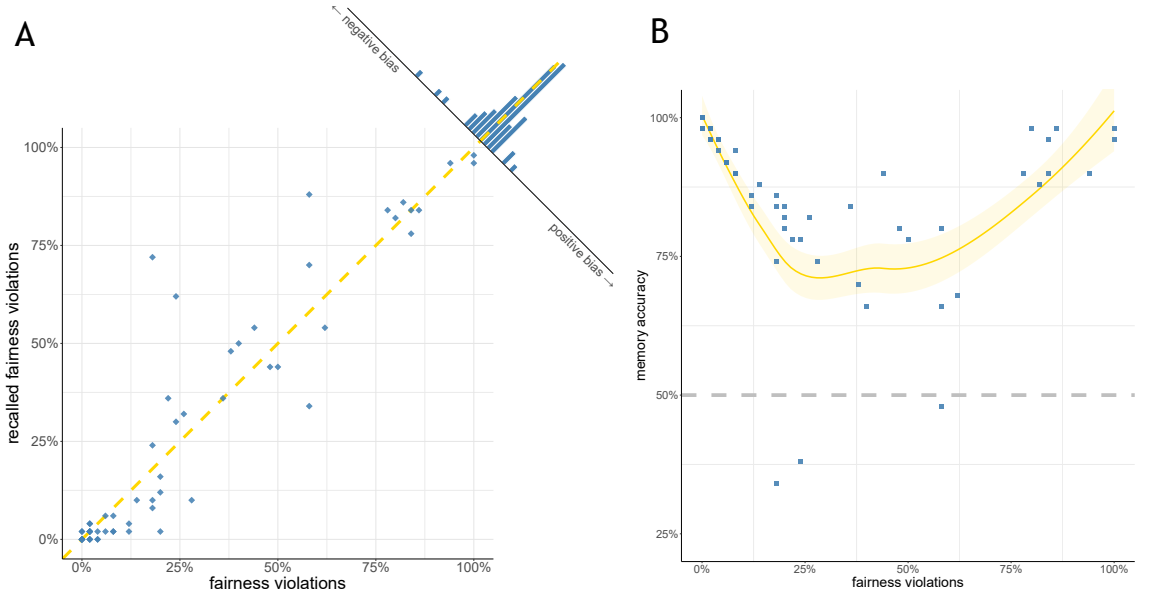

Figure 4.5: A. Relation between fairness violations and recalled fairness violations. Observations under the yellow line indicate a positive memory bias (recalling making *less* fairness violations than in actuality), whereas observations over the yellow line indicate a negative bias (recalling making more fairness violations than in actuality). B. Locally weighted regression for memory accuracy over the percentage of fairness violations. Grey dashed bar indicates chance level.

Fairness upholders thus seem to recall having made more fair choices than they actually did, a finding that could explain why, in our within-subject analysis, choice frequency predicts memory accuracy: the more frequently a participant chose fair allocations, the more she would recall having done so. Furthermore, not only fairness upholders, but also consistent fairness violators seem to have better recognition rates than the rest of the sample (Figure 4.5B). High memory accuracy in strong violators also suggests that the association between frequency of a behaviour and recall rates could be a general phenomenon: the more frequent a type of action is, the more likely it will be recalled correctly. To test this hypothesis explicitly, we check whether participants over-represent in memory the type of choice that they made more often (either selfish and antisocial for fairness violators or equitable for fairness upholders), in addition to a fairness memory bias. We divide the sample into participants that violated the fairness standard in more (or less) than 50% of trials. We then repeat the analyses in Figure 4.5A, this time considering the chosen proportion versus the recalled proportion of the most frequently chosen option. Contrary to this over-representation hypothesis, however, the results are not significant (Wilcoxon signed-rank test, $log(V) = 6.17$, $p = .515$, $r = .09$).

4.4 Discussion

While the mechanisms adopted by humans to cope with self-threatening memories have long been studied, only recently has research addressed the study of memory biases generated by an individual's own actions. In this chapter, we sought to explore two questions related to own-induced memory biases, namely (a) whether such biases could be predicted by a person's moral standards rather than objective antisocial behaviour, and (b) which memory contents (the choice itself, relevant

information used to make the choice, contingent data irrelevant to the choice) are affected by the bias.

We tried to answer the first question by using two different definitions of fairness violation, one that entailed participants' own reference for fairness (subjective self-report) and one based on participants' choices (proportion of equitable choices in a game of resource allocations). We do find evidence of a memory bias associated with fairness violations both within and between participants in the Choice condition; we are unable, however, to tell which of the two definitions of fairness is behind this bias. Indeed, a majority of participants reported the same subjective fairness reference point: the one allocating points should not get more points than the one who receives them (equitable split). This widely shared answer shows that there is broad agreement about what constitutes appropriate behaviour concerning the task at hand. Moreover, the proportion of equitable choices in the task inversely correlated with the number of subjective fairness violations, suggesting that subjective evaluations are strongly influenced by the idea of an even redistribution of resources, which also forms the basis of our objective definition of fairness violation. The overlap between the two measures was such that the statistical tests employed in the experiment yielded almost identical results, regardless of which of the two fairness definitions was used.

Our results differ from those of Carlson et al. (2018), where, in a series of experiments, the authors report a between-subject difference in memory recollection when categorising fairness violators based on their subjective fairness reference point, but not when categorising them based on overt behaviour (i.e., proportion of selfish versus fair choices). It is possible that differences in experiment design could have contributed to the fact that, in our experiment, we find a memory bias regardless of the definition of fairness. Our resource allocation task was based on 50 two-alternative choices (mini-Dictator Games) in contrast to the five choices

with 11 alternatives (standard Dictator Games) employed in the study by Carlson and colleagues. In addition, the subjective definition of fairness differs between the studies: in the present experiment the fairness reference point is the minimum acceptable offer by the recipient, whereas in the Carlson study, the reference point is framed in terms of maximum acceptable share kept by the player making the decision. Thus, contrary to previous work we cannot to tell whether a fairness memory bias depends on a subjective fairness reference point or on the objective concept of equity. To disentangle the two hypotheses, we suggest that future research could use a dedicated task in which participants' subjective fairness evaluations are less in agreement, for instance by creating an experimental context in which social norms are less certain. One such task could rely on a mechanism that hides the choices of the player from the recipient, for instance by making it impossible for the recipient to understand whether a donation has come from the player or from a random computer choice. It has, in fact, been shown that, when recipients have no direct information about the source of money they receive, there is more heterogeneity in how they judge the entity of the donation (Dana et al., 2007).

To answer the second question, we designed a between-subjects memory task in which participants had to recall information from the resource-allocation game: either their choices (Choice), or whether they had observed a particular allocation pair or not (Relevant Context), or whether they had observed an unrelated symbol or not (Irrelevant Context). We ran analyses both within and between subjects, finding results compatible with a fairness memory bias in the Choice condition, but not in the two context conditions. These results differ from a previous study that provided evidence that people do recall better allocation distributions—i.e. contextual information relevant to the choice—when their choice is prosocial rather than selfish (Saucet and Villeval, 2019). One possible explanation for this discrepancy is that the apparent absence of memory biases in the context conditions in the

present study could have been mediated by task difficulty. Indeed, due perhaps to the different nature of the tests (forced-choice recognition in the Choice condition versus old-new recognition in the context conditions), average recall accuracy was lower in the context conditions than in the Choice condition. If the task difficulty was too high, this could have attenuated performance, even in the absence of fairness violations, thus obscuring any effect of a memory bias. We therefore cannot definitively rule out an influence of memory bias on non-choice information.

If such a bias exists, easing the recognition task would likely improve the detection. Possible changes include using a different set of decoys, or perhaps even employing a different task. Old/new recognition tasks are, in fact, more biased than other memory instruments such as the forced-choice task used in the Choice condition (Rotello and Macmillan, 2007). As a possible substitute in the Relevant Context condition we propose a cued recall task similar to the one used in Saucet and Villeval (2019). Participants would then attempt to recall the points (for self or for the other) of the chosen allocation, instead of trying to recognise whether they have seen a particular allocation pair or not. The Irrelevant Context condition, on the other hand, could implement an implicit memory task to explore possible priming effects (Rossell and Nobre, 2004): symbols shown during the game would be cued before choices in an unrelated task, and then it could be tested whether symbols related to fairness violations (upholding) disrupt (facilitate) responses.

While we failed to find any evidence concerning memory bias in the context conditions, the branched results in the Choice condition hint at two complementary memory mechanisms, a fairness bias and an over-representation bias. We find evidence that fair choices are recalled better than fairness violations (i.e. choosing selfishly or outright antisocially) and that this effect is robust to competing explanations and external confounds, such as confidence or the presence of participants with extremely fair behaviour. Our results thus hint at the presence of a

memory bias across participants, in accordance with previous work (Saucet and Villeval, 2019). On top of this fairness memory bias, however, participants appear also to better recall choices that they made more frequently (either violations or upholding). Moreover, whereas we do find that fairness upholders recall better than fairness violators (similar to what was found by Carlson et al., 2018), this effect seems to be driven by a group of very consistent upholders. Indeed, consistent fairness violators also recall their choices accurately, but are under-represented in the sample.

We link the higher accuracy at the extremes of behaviour to another memory bias related to choice frequency, an over-representation bias. This hypothesis suggests that the more frequently a type of answer is given, the better its representation in memory. This in turn could explain why more consistent participants have better memory recognition rates: if a participant always chooses to be selfish or fair, there is no possibility that selfish or fair choices could be over-represented; over-representation should instead affect participants with more variable behaviour. We propose that this over-representation bias acts at retrieval: when participants try to recall their choices, they instead re-create them on the spot, based on their recalled most frequent behaviour. A selfish participant, for instance, could have completed the recall task not by remembering her individual choices, but rather by recalling that for the majority of trials, she made selfish choices, and that any choice was more likely to have been of that kind. Such an over-representation memory bias would then serve no self-protecting purpose, but rather would be a probabilistic mechanism to reconstruct one's own choices, regardless of their moral valence. This mechanism resembles the availability heuristic (Tversky and Kahneman, 1974), which states that information that is more available biases how we represent a related event. In this case, the representation of an event is the memory of a specific choice, whereas the more available information is the type of choice

that was made more frequently (either violating or upholding the fairness standards).

Evidence in support of an over-representation memory bias, however, is mixed. While we do find that fairness upholders recall having made more fair choices than they actually did, this result does not replicate for fairness violators, suggesting that not all participants "over-represent" choices of their preferred type. We do not have a precise explanation for this assortment of results; we speculate however, that the lack of significance at sample level could derive from the small proportion of consistent fairness violators. A larger sample could be used to further test this hypothesis by distinguishing between participants with consistent versus mixed behaviour. A second hypothesis is that the fairness bias and the over-representation bias could have interacted in fairness violators, an event that could, in turn, explain the mitigation of both effects in these participants (Figure 4.6). It is not obvious, however, how these biases would influence each other, nor if they act at the same stage of memory. To summarise, our results in the Choice condition hint at the contribution of a fairness bias and possibly also of an over-representation bias, such that memory of choices depends on both the fairness and consistency of a participant's behaviour . More evidence, however, is required to understand the actual impact and interconnection of these mechanisms.

Finally, we highlight some characteristics of the experimental design that might have limited the interpretability of the results. First, as we note above, and unlike previous studies employing similar tasks (Carlson et al., 2018; Saucet and Villeval, 2019), our sample consists mostly of prosocial participants. Such a sample composition could be explained by the particular set of allocations that we employed in the resource-allocation game: in most trials, participants were faced with a trade-off between a fair, more equitable split of points and an allocation in which they earned a high payoff at a similarly high cost for the other. It is possible that par-

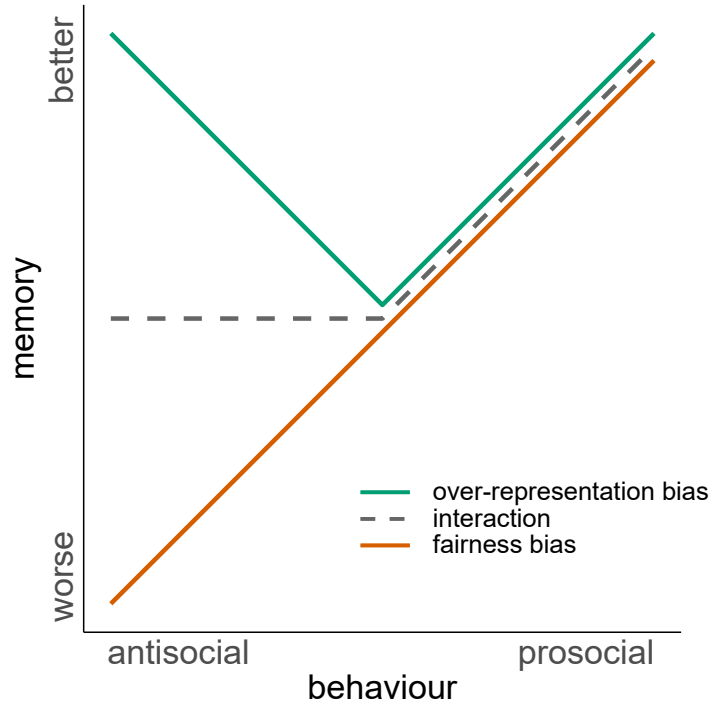

Figure 4.6: Possible model of how a fairness bias and an over-representation bias could interact to affect memory of social behaviour. Since the two biases push in different directions for fairness violators (participants behaving selfishly or antisocially) this could explain why we do not observe an over-representation bias in these participants.

ticipants disregarded selfish allocations, as they detracted too many points from the other, and as a consequence, they frequently chose the more fair allocation. Since the fairness self-report was composed of similar allocations to those of the resource-allocation game, the non-viability of selfish choices could also explain participants' largely homogeneous ratings. We propose, then, that a different set of trials could produce fewer extremely prosocial responses from participants, thus allowing the exploration of memory bias via a more heterogeneous sample.

Other characteristics of the design could have contributed to limiting the detection of a memory bias. The time distance between the resource-allocation game and the subsequent memory task was about five minutes to replicate a design similar to Carlson et al. (2018); it might be that the delay needs to be longer (e.g., 20-25 minutes) as such a short delay could mean that participants are still engaging some working memory process. A longer time could allow the studying of

biases unmistakably associated with long term memory. Another limitation of our study is that we do not collect any direct measure of memorisation abilities of participants: having a separate assessment would help accounting for individual differences in performance and would constitute a baseline upon which to compare the effects of any memory bias. Finally, we take into account the relevance of the resource-allocation game to participants' self image. In order for an event to be self-threatening it must, in fact, affect a person's central traits (Sedikides and Green, 2009). While we assert that our choices have a relevant impact on the wealth of both the participant and of the other player, we cannot affirm that participants felt that their choices meaningfully defined their moral integrity. To address this uncertainty, future research would have to measure the relevance of participants' choices to their central traits.

In conclusion, our study sought to detail the mechanisms behind fairness memory biases. We find that participants' considerations of fairness are overlapping to objective definitions of fairness (equity), and that a memory bias is present when recalling socially relevant choices, but not when recalling other types of information. Lastly, we propose that the memory bias is of two types: a self-serving mechanism in an effort to forget fairness-violating choices, and a probabilistic mechanism to recall correctly that recreates choices by over-representing the choices of the most frequent type. It is left to future research to further explore the impact of memory bias on contextual information and to investigate the existence of an over-representation bias.

5 General Discussion

5.1 Summary of findings

The goal of the present thesis was to explore open problems in the literature regarding social attitudes with the help of new theoretical models and experimental tasks. Our work focused mostly on three open questions concerning social influence, contextual information, and self-image. In Chapter 2, we tested the variability of social attitude following learning of others' behaviour. We found that participants observing others did indeed shift their attitude towards that of the observed agents, and that this change can be attributed to two normative forces: norm learning (due to horizontal influence from peers) and authority compliance (driven by the desire to fulfil the experimenter's expectations). In Chapter 3, we tested the concept of meta-context, namely the idea that information about unavailable actions influences beliefs and choices regarding the available actions. We found that participants took into account unavailable options when evaluating the social appropriateness of an allocation, and that normative beliefs in turn predicted participants' choices in the main dictator game task when using a norm-based utility function. Additionally, our results suggest that there are supposedly two types of normative influences, one related to meta-context and one related to the specific context, but that only the meta-context one is captured with norm elicitation methods. The second type of normative influence, relative cost, still influences participants decisions and correlates with participants' propensity to follow arbitrary rules as measured in the rule following task. Lastly in Chapter 4, we studied how participants process memories of morally compromising choices such as behaving selfishly or even anti-socially in a Dictator Game. Consistent with previous studies, we found that memories of prosocial choices are recalled better than selfish or anti-social choices, but also found a possible additional influence of choice frequency as

a memory bias: participants seem to over-represent the type of choice they made more frequently in the task, regardless of its moral valence. This hypothesis however needs to be investigated further. As a second exploration, we tested whether contextual information in the Dictator Game is also susceptible to memory bias; due perhaps to problems in the experiment design, however, we were unable to confirm the presence–or absence–of any bias concerning this information. Beyond the main research questions, the conclusions drawn in these chapters also dismantle some widely held theoretical beliefs concerning economic choice theory. We approach in different chapters two different axioms of economic decision making: namely that we value alternatives separately, then we compare them (Chapter 3), and that preferences for an object or an action are fixed and do not change with time (Chapter 2). We have obtained credible evidence that, when social decisions are concerned, such assumptions do not always hold true. It is important then to re-conceptualise both the use of preferences to define social attitudes and the idea that context has no influence on value estimation (see for instance Huber et al., 1982).

5.2 The Role of Norms in Social Attitude

As a whole, these studies point at the relevance of participants' normative beliefs, supporting the use of a norm-based theory to study social attitudes. Norms seem to shape the way we treat others, the way we learn from them, and even our memories of past actions. A norm-based approach draws the intuitive yet critical conclusion that social attitude is, indeed, social, in the sense that it depends more than anything else on the social situation in which it is expressed. We note however several complications of norm-based theories, starting from the potential presence of multiple norms (Chapter 3), the influence of experimenters (Chapter 2), but also the locus of norms (subjective/internal vs. shared/external, Chapter 4). These rel-

evant features of norms, if unaccounted for, could explain why normative ratings do not differ between rule followers and rule breakers in Chapter 3, whereas we find a strong relation between behaviour and norms in a less conventional task such as the Resource-Allocation Game in Chapter 2. These problems highlight the lack of adequacy of current elicitation measures, as they are unable to explain this contextual variability for apparently simple games with similar structure. A partial solution might come from methodological improvements: for instance – as we suggest in Chapter 3 – there is a need for more fine-grained indices of normative beliefs, but also of norm-following propensity. Increased quality of data could allow using normative beliefs as a predictor variable within subjects. One method we adopted for norm elicitation was point interpolation (Chapter 3), but this could more easily be addressed by increasing the range of allocations rated during norm elicitation. Continuous, individual-based measures could have multiple applications other than computations of norm-based utility. One example could be the measurement of social influence: does normative salience predict how strongly one will conform to others' social behaviour? (Chapter 2).

5.3 Drivers of Social Attitude

The development of social attitude research also comes from theoretical improvements. In particular, given the uneven development of social attitude theories, we believe that the development of a broader framework is necessary. Experimental evidence should inform such framework on what personal and contextual features drive the actions of an individual. The discussion of the results in the previous chapters indeed suggests that participants' behaviour in Dictator Game-like situations depends on the information available to them, even if seemingly irrelevant. Such information in turn influences how participants interpret, and thus how they approach, the game. Multiple motivations can stem from a deceivingly simple de-

scription of a Dictator Game, from being truthful to one's preferences or ideals, to providing an impression of oneself to others. Thus, to organise some of the phenomena which we have observed in this thesis and in the previous literature, we propose a descriptive classification of some of the information and motives driving the expression of social attitude (Table 5.1).[30]

This organisation of information and motives does not intend to be a prescriptive or exhaustive list, but is rather conceived as a tool helpful for experiments employing tasks similar to the Dictator Game. In our view, a systematic, experimentally-based approach to social attitude is conducive to understanding how information influences the game representation (i.e., the game structure) and what motivations guide participants' choices (i.e. their utility representation). This approach, we hope, could then inform the design of future experiments in the field.

State	**Locus**	
	Internal	**External**
Perception	Personal Value	Frame
Expectation	Self-image	Social Norms
Uncertainty	Memory	Noise

Table 5.1: Possible categorisation of drivers of behaviour in social decision making.

We categorise these drivers based on the physiological or psychological state to which they can be associated (*state*), as well as whether they arise inside (motives) or outside (information) the individual (*locus*). We provisionally identify three types of states: states relating to perceived characteristics of the situation (perception, or 'what is'), states based on some normative evaluation (expectation, or 'what should be'), and states deriving from lack of information (uncertainty, or 'what should be known').

[30]Note that we do not present here the drivers under an exact utility form, as we have yet limited knowledge about how they contribute or interact in the expression of a person's attitude.

Internal perception refers to the inner motivation of a person to respond to a certain need (Juechems and Summerfield, 2019). In utility terms, one could consider these motives as the representation of personal value: monetary gains and losses could be interpreted as appetitive or aversive stimuli whose value depends on the internal state of the individual, similar to physiological needs such as thirst or hunger. Whereas money is used as a universal incentive in experimental settings–arguably for its ability to satisfy any unobserved personal preferences–more internal states could have a non-negligible impact on participants' decisions, such as social acceptance or boredom.

External perception refers to salient information related to the experimental setting. External perception modulates the personal value of an option, for instance by making it seem more valuable in the absence of a better alternative. Whereas this information is generally controlled for in economic experiments, it is known to lead to some utility paradoxes (e.g., the decoy effect, Huber et al., 1982). Our experiment in Chapter 3 partly points at this problem when showing how unavailable allocations have an influence in choosing between the available options.

Internal expectations are motivations arising from an internal judgement, for instance based on an individual's self-image. Examples include the need to be self-consistent (Mullen and Monin, 2016), but also internalised norms (Jimenez-Buedo and Guala, 2016; Bicchieri, 2005), namely those moral prescriptions that are followed by the individual even in the absence of any repercussions . It remains almost impossible to account for this inner voice of the participant if not by means of an elicitation measure external to the experimental task; moreover, internal expectations can be easily confounded with other states also modulating personal value. Fairness self-report ratings in Chapter 4, for instance, leave open the question whether participants' behaviour is guided by internal but widely common expectations or rather by external judgement. We thus deem necessary to develop

new methods to differentiate internal expectations from other drivers of social behaviour.

External expectations refer to others' judgements about a possible choice of the individual. In most cases, external expectations follow a social norm violation, defined as the lack of adherence to a rule widely followed by others (Bicchieri, 2005). In this sense, external expectations are the drivers behind the norm-component in the utility function of norm-based models. While external expectations may push participants to behave differently from their preferences (e.g., to act selfishly in a Dictator Game), they can also be exploited to signal information to observers. Signalling occurs in some instances of experimenter demand where participants want to convey a positive self-image to the experimenter (Zizzo, 2009). Vice versa, when personal actions can be fogged by increasing uncertainty about one's actions (e.g., anonymity), then instead participants exploit this moral wiggle room to behave more freely (Dana et al., 2007; van Baar et al., 2019). External expectations are thus the cornerstone of normative influences, but experimenters should beware of the sources from which they originate, for instance by measuring experimenter demand as in Chapter 2; alternatively, if experimenters are more interested in internal states (e.g. tracing personal traits), then experimental settings should reduce to a minimum any chance of signalling or information about the intentions or identity of agents.

Lastly, a state of uncertainty refers to a lack of external information and to the internal need to make sense of the situation. We can represent uncertainty in utility functions in terms of noise, or of variability in the value of an option. Internal uncertainty can stem from unfamiliar problems, where participants do not know their preferences; internal uncertainty can thus motivate a participant to explore the possible choices to learn the outcomes. This type of internal uncertainty can be represented using random utility models as in Chapter 2. In addition, internal

uncertainty can also arise from a limited recollection of information (Bhatia, 2018) or even faulty memories, as is the case in Chapter 4. External uncertainty refers instead to missing or conflicting information about the choice setting, such as when the outcomes of a choice are not certain. Experimenters can reduce external uncertainty by increasing clarity in task-related information, and internal uncertainty by avoiding choices that are demanding to participants.

We could consider to include multiple other components, such as a temporal dimension (a value of a choice on the short versus long term, Urminsky and Zauberman, 2015), or a more detailed categorisation of external expectations (consider for instance the distinction between descriptive and injunctive norms, Cialdini and Goldstein, 2004). We also note that this categorisation has a distinctive focus on one-directional social behaviour: indeed, we deliberately excluded from the cases considered any implication concerning interactions between agents. There are certainly variations of the Dictator Game involving some possibility for the recipient to respond to Dictator's decisions, such as in the Impunity Game (Bolton and Zwick, 1995) where the recipient can destroy the donation from the Dictator, or the Ultimatum Game (Güth et al., 1982), where in addition to destroying their donations they also destroy the resources left to the dictator. In interactive games of this kind, however, there are multiple more features influencing choice, since decision makers need to take into account other's motivations as well (for an in-depth neuroscientific exploration on the subject, see Hill et al., 2017). Given the nearly uncountable combinations of and interactions between each agent's drivers, we thus left strategic concerns out of this classification. While we do not argue against the use of interactive games to study social behaviour, we suggest that they are not the most accessible setting to pin down the exact motivations behind people's social attitude, as compared to simplest settings such as a Dictator Game.

Precisely because of the number of motivations and the quantity of information to consider in social situations, one might wonder how people are still able to make a choice. We conjecture, based on our categorisation, that multiple drivers can influence participants' decisions in an interchangeable way, such that not all relevant information has to be actually taken into account to make a choice. For instance, dictators could fall short to consider the recipient's opinion on their choice (external expectation) but still act generously due to moral commitment (internal expectation). This idea implies that humans are able to figure out how to respond in a social situation by filtering only the salient information according to the strongest drivers. We propose then as future research on social attitude to study the dynamics between its drivers, and whether some drivers are more salient than others.

5.4 Conclusions

Over the years, the Dictator Game has been recurrently declared dead, and with it the ability to understand the social and normative influences that characterise social attitudes. The empirical findings of this thesis, along with its methodological and theoretical contributions suggest that to the contrary, there are still many paths to follow before we get an accurate understanding of human social reasoning. We thus hope that the proposals that we brought in this thesis would be of use for future researchers in economics, psychology, and neuroscience. Social Attitude remains an evolving concept and we must take all the care possible not to be drawn into traps of theories that limit our perspective on one aspect or another. Every aspect matters.

6 Appendix

6.1 Chapter 2: Supplementary Methods

6.1.1 Preregistration amendments

This study was originally preregistered at the Open Science Framework osf.io/th6wp. Although we tried to be as faithful as possible to the original project, we made some changes which we report here:

- Design: the original plan to recruit 100 participants was not reached due to a limited subject pool.
- Hypotheses:
  1. We added the time-dependence hypothesis (participants change attitude even in the absence of another agent).
  2. We added the norm salience hypothesis (participants change attitude because they learn how salient the underlying norm is).
  3. The norm learning hypothesis does not strictly exclude attitude change in the Individual condition.
  4. New prediction for Preference Learning: increase in consistency.
- Exclusion Criteria:
  1. We changed the prediction accuracy requirement given that the original formulation (90% accuracy after first 10 trials) excluded several dozens of participants.
  2. The prediction accuracy requirement has become a necessary condition.
  3. We exclude trials in which participants' responses were implausibly fast ($<$ 200ms).
- Modelling:
  1. MCMC Chains were run only for the original 100,000 iterations even when $\hat{R} > 1.05$ due to a software issue. This notwithstanding, almost all $\hat{R}$ estimates were well below a more lenient threshold of 1.1 in all models.
  2. We introduced the bias parameter κ and the error parameter ε in the models.
  3. We changed the operalisation of the dependent variable (Section 2.2.2).
- Analyses:
  1. We could not test for participants' numeracy as originally planned due to a problem with the software that inflated the number of mistakes by participants.

2. We did not use Bayes Factor given that almost all measurements were strongly violating the necessary assumptions.
3. Exploratory analyses such as linear regressions were dropped because hardly applicable due to the nature of our data and to the absence of a reasonable methodological alternative. One example which we retain is the Robust Regression in Figure 2.6.

- Names:

1. Epistemic Uncertainty $\rightarrow$ Preference Learning
2. Preference Temperature $\rightarrow$ Stable Attitude
3. Preference Uncertainty $\rightarrow$ Variable Attitude

6.1.2 Undisclosed Information

We omitted to inform participants about several features of the agents, such as the nature of computer choices, the representativeness of the choices of human (Individual and Group) agents, and the size of the group in the Group condition. Choices of the computer agent were the same as the ones of the group to keep a consistent manipulation across conditions. Informing participants about the social nature of computer choices would have influenced the perception of an agent that was supposed to be inherently non-social. Human agents instead were not representative of participants in the Baseline condition as they were carefully selected to display a highly consistent and pre-specified social attitude. Participants were not informed about the agent selection procedure to reduce the risk of experimenter demand effects: telling participants that agents were arbitrarily chosen by the experimenter would have biased the following conformity to the agent. In addition, no clear information about norms can be learned from an unrepresentative sample, thus providing this information would have prevented any test of the Norm Learning hypothesis. Similarly, we did not disclose the size of the group, as the actual number of participants (five) could have similarly biased the idea of representativeness of the observed group. We could not find any alternative solution to omitting this information to participants in order to achieve the same goals.

6.1.3 Cognitive Model Priors

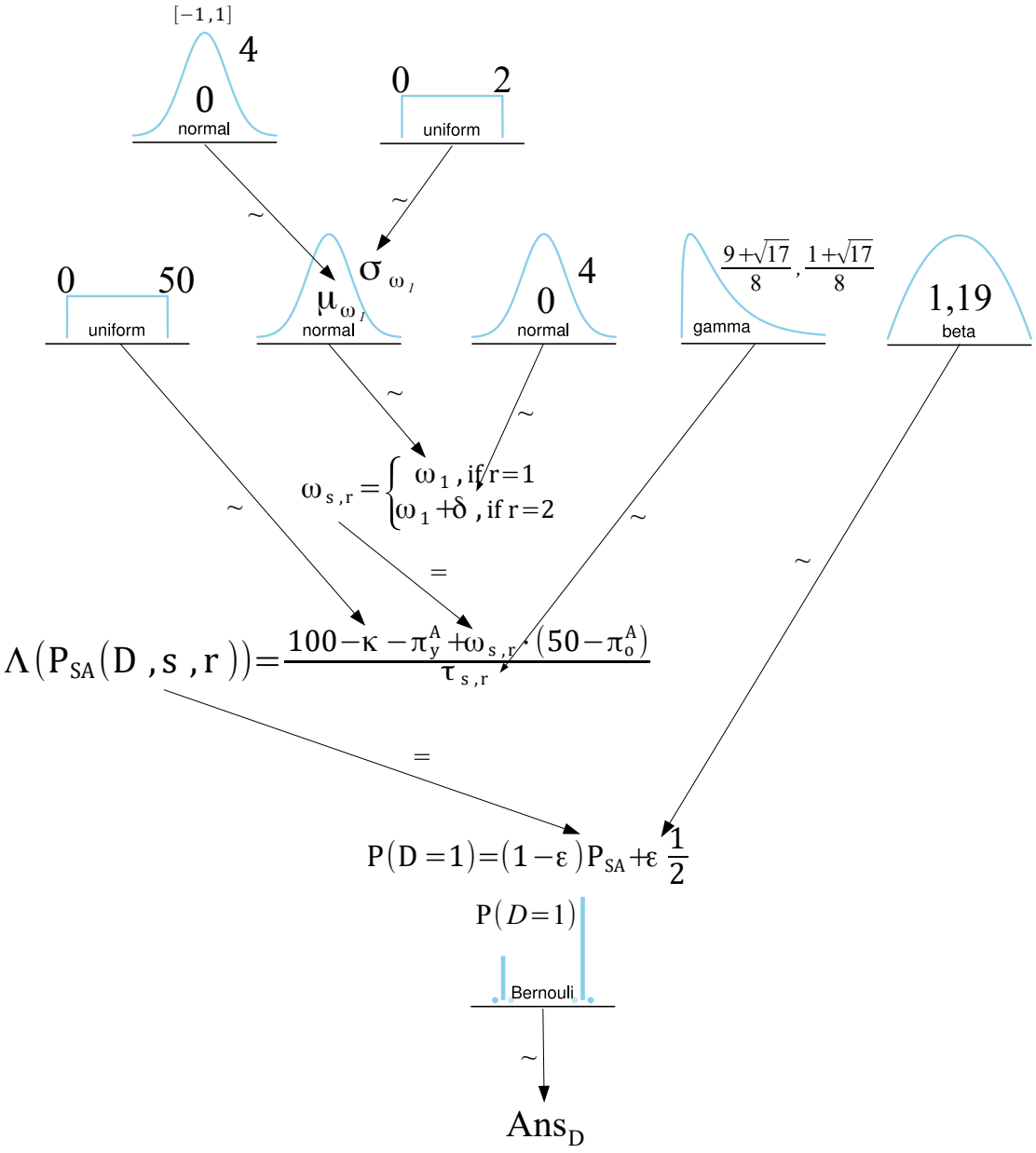

Figure 6.1: **Stable Attitude Model and Priors.** Kruschke-style diagram of the Stable Attitude full model. Subscripts: *A* indicates the points in the alternative allocation, *r* indicates whether the trial was presented before (1) or after (2) the manipulation phase, and *s* indicates the participant that gave the answer. Angles are expressed and estimated in radians. Courtesy of Rasmus Bååth, 2016 (http://www.sumsar.net/blog/2013/10/diy-kruschke-style-diagrams/). Note: gamma distribution values indicate shape and rate.

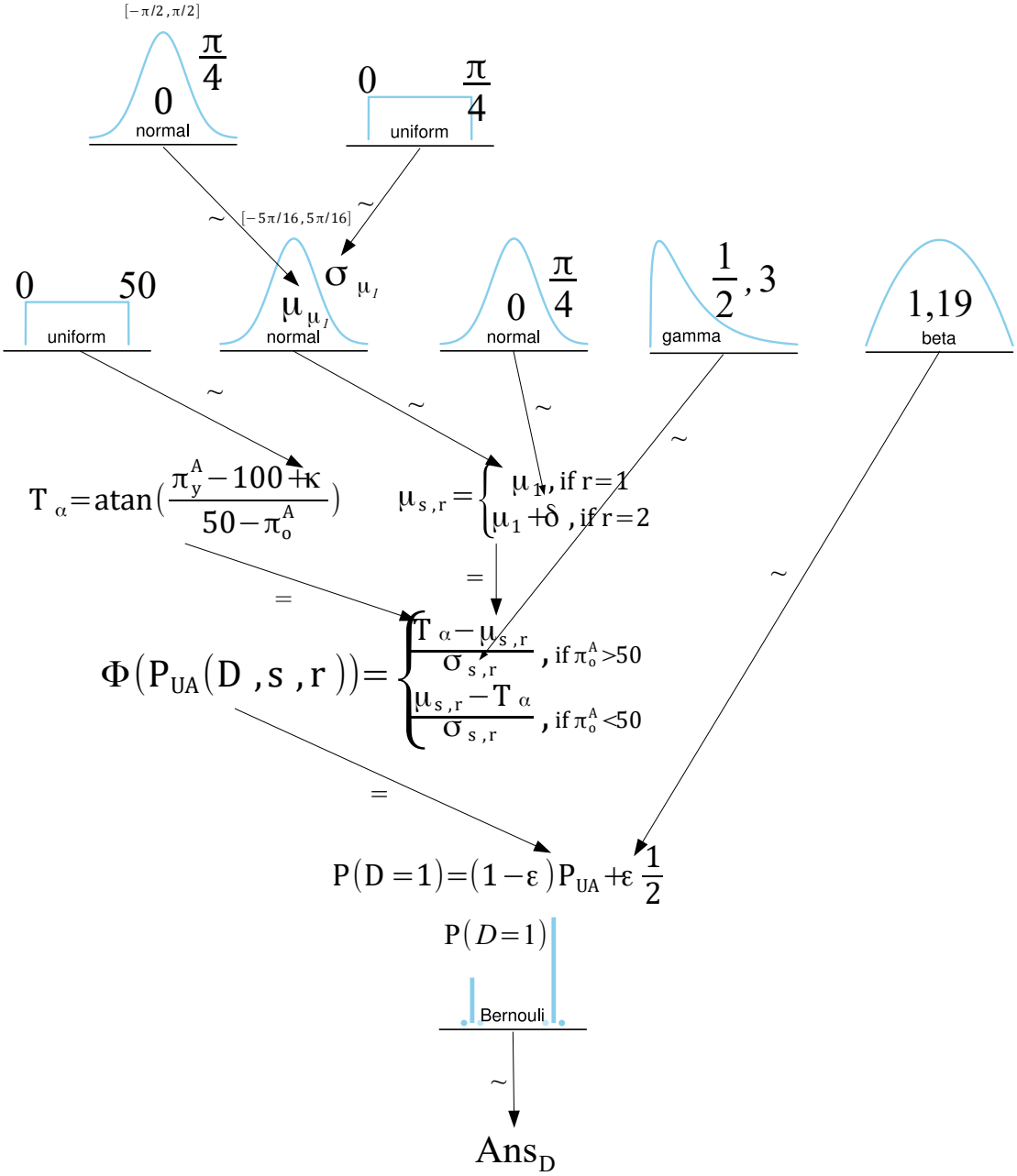

Figure 6.2: **Variable Attitude Model and Priors.** Kruschke-style diagram of the Variable Attitude full model. Angles are expressed and estimated in radians. Note: gamma distribution values indicate shape and rate.

6.2 Chapter 2: Supplementary Results and Discussion

6.2.1 Main analyses with excluded participants

We repeat the main tests on attitude conformity including all excluded participants, with the exception of four participants in the Baseline condition for whom we lost response data ($N = 372$).

Similar to what we observe when excluding participants, attitude convergence δ_{diff} is significantly greater than zero in all conditions except Baseline (Baseline: $\log(V) = 8.47$, $p = .186$, $r = .08[-.10, .25]$, $n_{\text{obs}} = 132$; Computer: $\log(V) = 7.59$, $p = .001$, $r = .37[.17, .56]$, $n_{\text{obs}} = 74$; Individual: $\log(V) = 7.32$, $p = .006$, $r = .32[.10, .55]$, $n_{\text{obs}} = 66$; Group: $\log(V) = 8.30$, $p < .001$, $r = .51[.36, .66]$, $n_{\text{obs}} = 100$).

There is also a significant effect of condition on attitude convergence (Kruskal-Wallis test, $\chi^2(3) = 18.46$, $p < .001$, $\varepsilon^2 = .05[.02, .11]$). Post-hoc pairwise comparisons reveal that attitude convergence differs between the Baseline and Group conditions ($W = 6.13$, $p < .001$), whereas there seems to be no statistical difference between the other conditions.

When we control for compliance, δ_{diff} is still significantly greater than 0 in all conditions excluding Baseline, with the Computer condition being still marginally significant (Baseline: $\log(V) = 8.09$, $p = .217$, $r = .08[-.11, .26]$, $n_{\text{obs}} = 109$; Computer: $\log(V) = 7.07$, $p = .036$, $r = .25[.01, .49]$, $n_{\text{obs}} = 60$; Individual: $\log(V) = 6.86$, $p = .002$, $r = .44[.22, .65]$, $n_{\text{obs}} = 50$; Group: $\log(V) = 8.03$, $p < .001$, $r = .47[.30, .64]$, $n_{\text{obs}} = 89$).

Attitude convergence is significantly different between conditions ($\chi^2(3) = 15.77$, $p = .001$, $\varepsilon^2 = .05[.02, .12]$, $n_{\text{obs}} = 308$, Supplementary Figure 6.3). Post-hoc comparisons find a difference between the Baseline and Group conditions ($W = 5.34$, $p = .005$).

The significance of attitude convergence in the Computer condition after controlling for compliance does not drastically change the results compared to the original finding, especially when considering that the hypothesis tested is unidirectional (if the test is run bidirectionally, $p = .053$). In addition, when considering compliance as a continuous measure, we observe that in the Computer condition attitude convergence is significant only for high values of compliance, suggesting that any effect that we detect in this condition is in fact related to the presence of the experimenter (Supplementary Figure **??**). Notably, this relation is only present in the Computer condition, whereas in the Individual condition this trend is even negative: the higher the compliance index, the less participants conform on average. All remaining findings concerning the robust regression are preserved with respect to the original test.

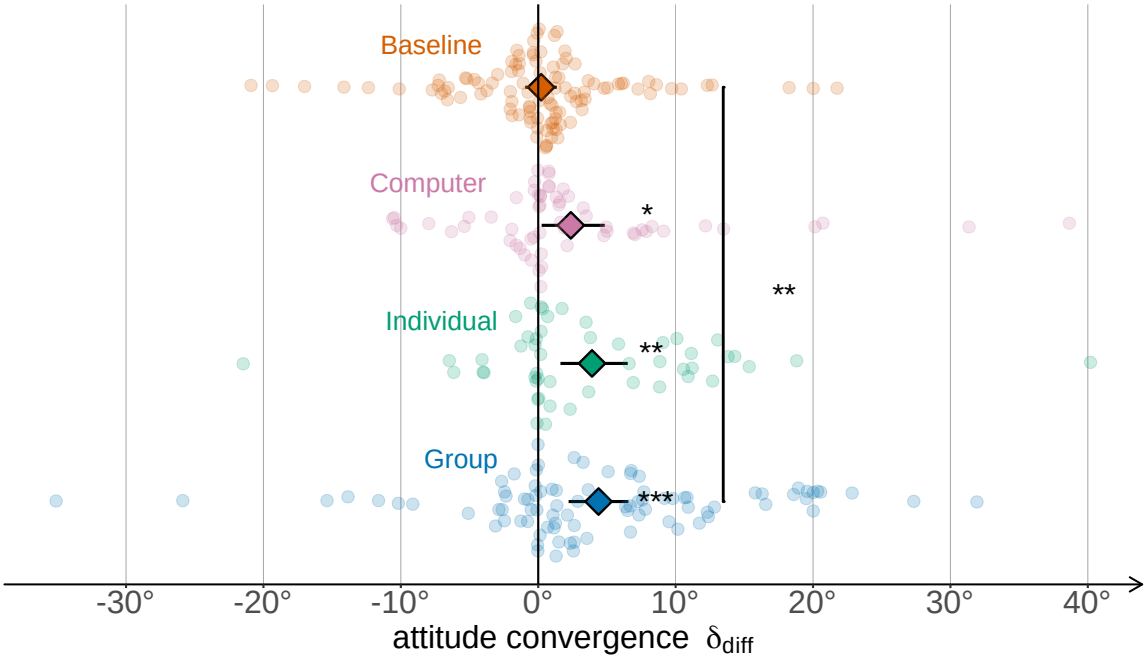

Figure 6.3: Mean attitude convergence by condition for participants below compliance threshold (25%). Error bars indicate t-adjusted, 95% Gaussian confidence intervals. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

6.2.2 κ and ε Parameter Summary

Table 6.1 summarises the estimates of the bias parameter κ and of the error parameter ε across all participants, before and after the manipulation.

Parameters	**Before Manipulation Phase**		**After Manipulation Phase**	
	Median (IQR)	**Mean (SD)**	**Median (IQR)**	**Mean (SD)**
ε (% of trials)	1.5 (4.1)	1.4 (4.6)	0.4 (1.3)	1.4 (2.2)
κ (penalty points)	0.8 (2.7)	3.9 (7.8)	0.8 (1.3)	2.3 (6.0)

Table 6.1: Parameter summary. ε indictates the percentage of trials in which there was likely a mistake by the participant; κ indicates the penalty points of the default allocation with respect to the alternative allocations.

6.2.3 Compliance Index by Condition

6.2.4 Attitude Convergence and Consistency Increase

As an exploratory question, we asked whether participants would become more consistent the more they conform. We tested whether attitude convergence and

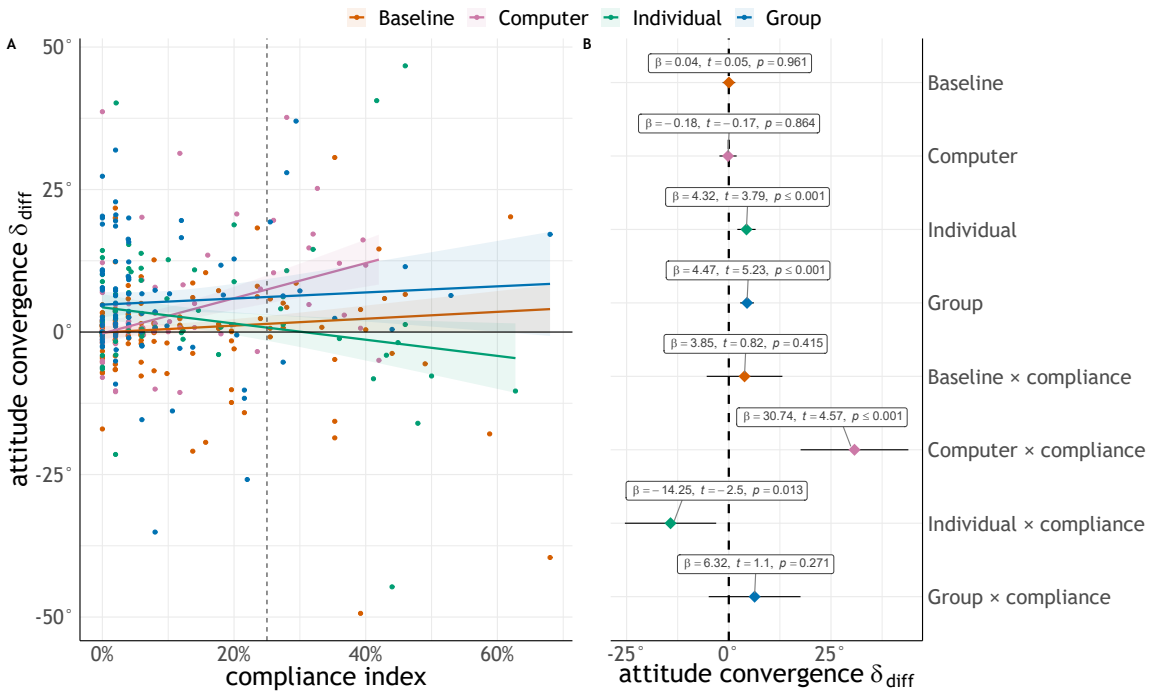

Figure 6.4: A: Robust regression on attitude convergence with experimental condition and the interaction between compliance and experimental condition as predictor variables. Shaded areas indicate 95% confidence intervals. B: Coefficients of the regression. Labels report unstandardised effect size, t-value, and p-value. Error bars indicate t-adjusted, 95% Gaussian confidence intervals.

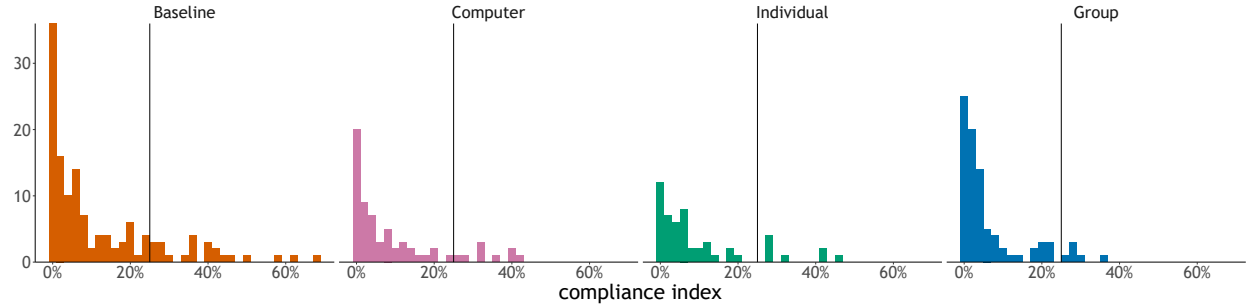

Figure 6.5: Distribution of the compliance index by condition.

consistency increase were positively correlated by means of a directional Spearman's rank correlation. Correlation was not significant in any condition after correcting for multiple comparisons (all $p > .067$).

6.3 Chapter 3: Supplementary Materials and Analyses

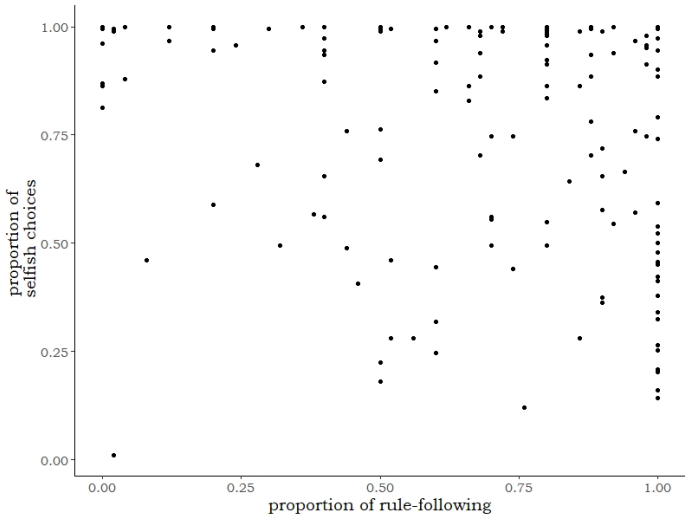

Figure 6.6: The relation between the rule-following propensity and the proportion of selfish choices.

Allocation							
A	B	A	B	A	B	A	B
3/57	60/0	30/30	33/27	37/23	26/34	50/10	31/29
5/55	55/5	30/30	35/25	37/23	27/33	50/10	34/26
5/55	60/0	30/30	39/21	37/23	30/30	50/10	35/25
10/50	50/10	30/30	45/15	37/23	31/29	50/10	40/20
10/50	57/3	30/30	55/5	37/23	34/26	50/10	45/15
15/45	50/10	30/30	60/0	37/23	35/25	50/10	57/3
15/45	57/3	31/29	29/31	37/23	40/20	55/5	10/50
20/40	45/15	31/29	33/27	37/23	50/10	55/5	15/45
20/40	55/5	31/29	35/25	37/23	57/3	55/5	23/37
20/40	60/0	31/29	39/21	39/21	23/37	55/5	25/35
21/39	39/21	31/29	45/15	39/21	25/35	55/5	28/32
21/39	45/15	31/29	55/5	39/21	28/32	55/5	29/31
21/39	55/5	31/29	60/0	39/21	29/31	55/5	32/28
21/39	60/0	32/28	30/30	39/21	32/28	55/5	33/27
23/37	37/23	32/28	31/29	39/21	33/27	55/5	37/23
23/37	40/20	32/28	34/26	39/21	37/23	55/5	39/21
23/37	50/10	32/28	37/23	39/21	40/20	55/5	50/10
23/37	57/3	32/28	40/20	39/21	50/10	55/5	57/3
25/35	37/23	32/28	50/10	39/21	57/3	57/3	3/57
25/35	40/20	32/28	57/3	40/20	20/40	57/3	5/55
25/35	50/10	33/27	28/32	40/20	21/39	57/3	20/40
25/35	57/3	33/27	29/31	40/20	26/34	57/3	21/39
26/34	35/25	33/27	32/28	40/20	27/33	57/3	26/34
26/34	39/21	33/27	34/26	40/20	30/30	57/3	27/33
26/34	45/15	33/27	37/23	40/20	31/29	57/3	30/30
26/34	55/5	33/27	40/20	40/20	34/26	57/3	31/29
26/34	60/0	33/27	50/10	40/20	35/25	57/3	34/26
27/33	33/27	33/27	57/3	40/20	45/15	57/3	35/25
27/33	35/25	34/26	26/34	40/20	55/5	57/3	40/20
27/33	39/21	34/26	27/33	40/20	60/0	57/3	45/15
27/33	45/15	34/26	30/30	45/15	15/45	57/3	60/0
27/33	55/5	34/26	31/29	45/15	23/37	60/0	0/60
27/33	60/0	34/26	35/25	45/15	25/35	60/0	10/50
28/32	32/28	34/26	39/21	45/15	28/32	60/0	15/45
28/32	34/26	34/26	45/15	45/15	29/31	60/0	23/37
28/32	37/23	34/26	55/5	45/15	32/28	60/0	25/35
28/32	40/20	34/26	60/0	45/15	33/27	60/0	28/32
28/32	50/10	35/25	25/35	45/15	37/23	60/0	29/31
28/32	57/3	35/25	28/32	45/15	39/21	60/0	32/28
29/31	32/28	35/25	29/31	45/15	55/5	60/0	33/27
29/31	34/26	35/25	32/28	45/15	60/0	60/0	37/23
29/31	37/23	35/25	33/27	50/10	20/40	60/0	39/21
29/31	40/20	35/25	39/21	50/10	21/39	60/0	50/10
29/31	50/10	35/25	45/15	50/10	26/34	60/0	55/5
29/31	57/3	35/25	55/5	50/10	27/33		
30/30	31/29	35/25	60/0	50/10	30/30		

Table 6.2: Mini-DGs used in the experiment (dictator's/recipients' payoffs).

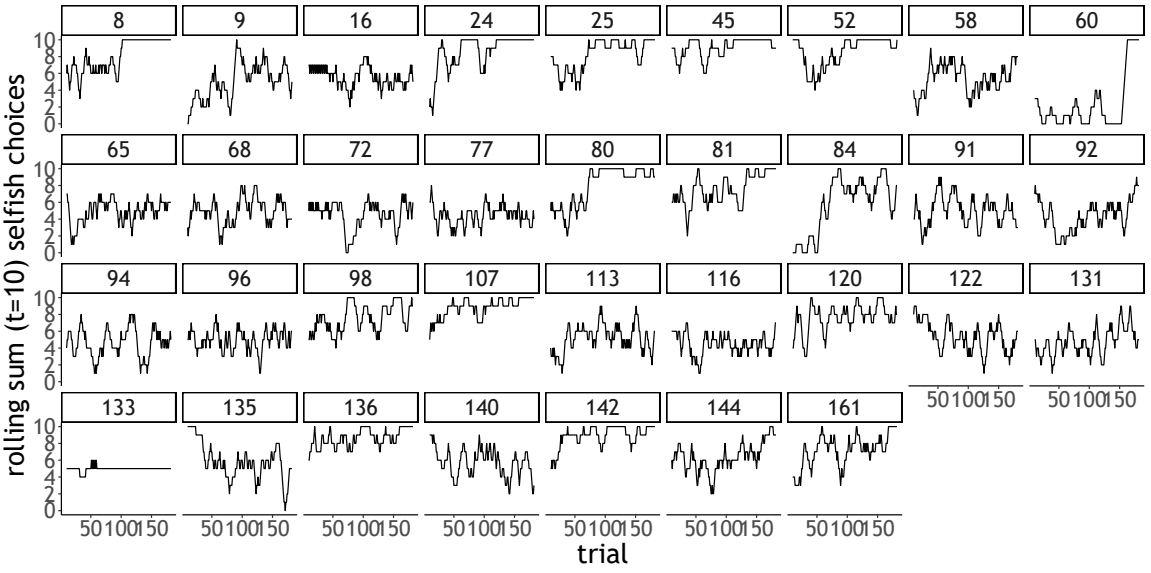

Figure 6.7: To test whether participants' did not pay attention or reduced their attention to allocations during the task, we searched for four possible patterns of behaviour: choosing always the allocation on the left of the screen, choosing always the allocation on the right of the screen, choosing always selfishly, choosing at random. We computed the sum of selfish (or left) choices using a sliding window of ten trials (i.e. for trials from 1 to 10, from 2 to 11, etc.) for each participant. The graph represents all participants that display a consistent change towards selfish behaviour (e.g. participant 8) or random responses (e.g. participant 133) alongside a selection of participants displaying a consistent pattern of behaviour. No anomalous patterns are observed for left/right choices. Visual inspection suggests that less than ten participants out of 166 present a behaviour that can be associated to a lack or reduction of attention during the task.

Allocation		A			B		
A	B	mean	SE	median	mean	SE	median
0/60	60/0	-0.570	0.060	-1.000	-0.622	0.057	-1.000
15/45	55/5	-0.213	0.044	-0.333	-0.546	0.050	-1.000
21/39	57/3	0.092	0.042	0.333	-0.610	0.051	-1.000
25/35	55/5	0.321	0.035	0.333	-0.554	0.049	-1.000
26/34	35/25	0.325	0.037	0.333	0.482	0.026	0.333
27/33	33/27	0.498	0.036	0.333	0.606	0.027	0.333
28/32	37/23	0.486	0.037	0.333	0.337	0.032	0.333
30/30	34/26	0.847	0.028	1.000	0.502	0.028	0.333
30/30	39/21	0.867	0.025	1.000	0.257	0.036	0.333
30/30	60/0	0.859	0.027	1.000	-0.719	0.050	-1.000
30/30	45/15	0.880	0.024	1.000	-0.096	0.039	-0.333
31/29	37/23	0.703	0.028	1.000	0.341	0.033	0.333
32/28	35/25	0.606	0.030	0.333	0.438	0.031	0.333
32/28	57/3	0.614	0.032	0.333	-0.566	0.055	-1.000
33/27	34/26	0.550	0.027	0.333	0.482	0.030	0.333
33/27	45/15	0.586	0.031	0.333	-0.092	0.039	-0.333
55/5	34/26	-0.526	0.051	-1.000	0.526	0.030	0.333
60/0	45/15	-0.683	0.053	-1.000	-0.068	0.039	-0.333

Table 6.3: Items used in the norm elicitation task and the rating summary.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
a_{adv}	–1.18	–1.04	–0.97	–0.69	–0.70	1.09
b_{adv}	–1.29	1.44	1.78	1.52	1.99	2.40
c_{adv}	0.00	0.98	1.49	8.53	2.14	538.38
a_{dis}	–1.14	–1.01	–0.95	–0.58	–0.34	1.08
b_{dis}	–0.86	1.11	1.65	1.44	1.94	2.24
c_{dis}	0.00	1.35	2.29	44.16	6.83	3637.47

Table 6.4: Summary of the individual parameters for the advantageous and disadvantageous norm functions.

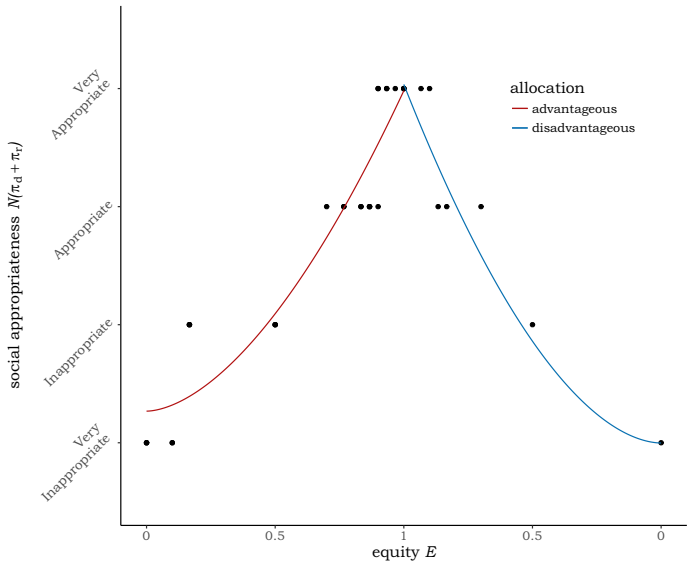

Figure 6.8: An example of a power function fitting (participant 17).

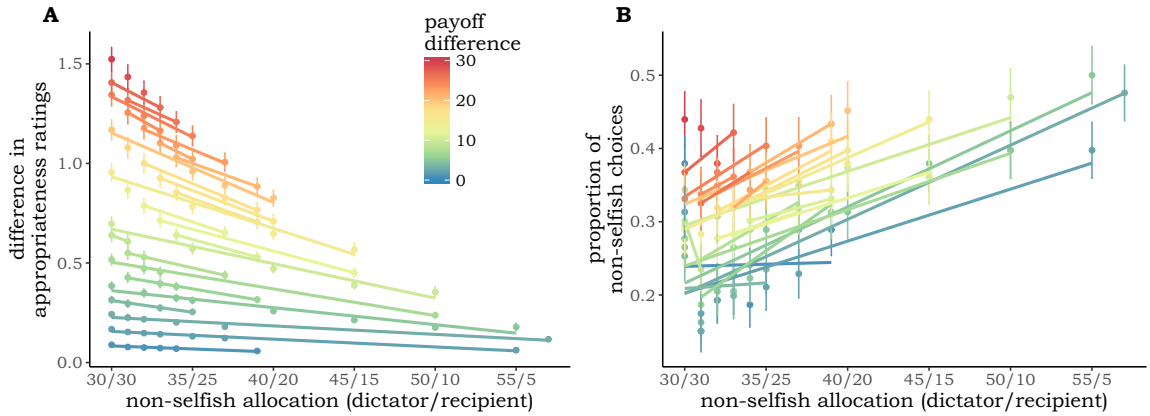

Figure 6.9: Difference in appropriateness rating and proportion of non-selfish choices for mini-DGs with fixed payoff differences (similar to Figure 3.3).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\pi_d^{nonself}$	0.026***	0.011***	0.010***	0.009**	0.016***	0.000	0.000
intercept	–0.862***	–0.316***	–0.246***	–0.284**	–0.541***	0.000	–0.125***
N observations	12,376	12,376	12,376	12,376	12,376	12,376	14,144

Table 6.5: OLS regression of the residuals from different models. (1) standard norm-dependent utility; (2) degree 3 polynomial utility; (3) concave norm utility $c = 0.1$; (4) concave norm utility $c = 0.5$; (5) concave norm utility $c = 0.8$; (6) standard norm-dependent utility with relative cost; (7) standard norm-dependent utility, disadvantageous allocations; ***,** stand for $p < 0.001$ and $p < 0.01$.

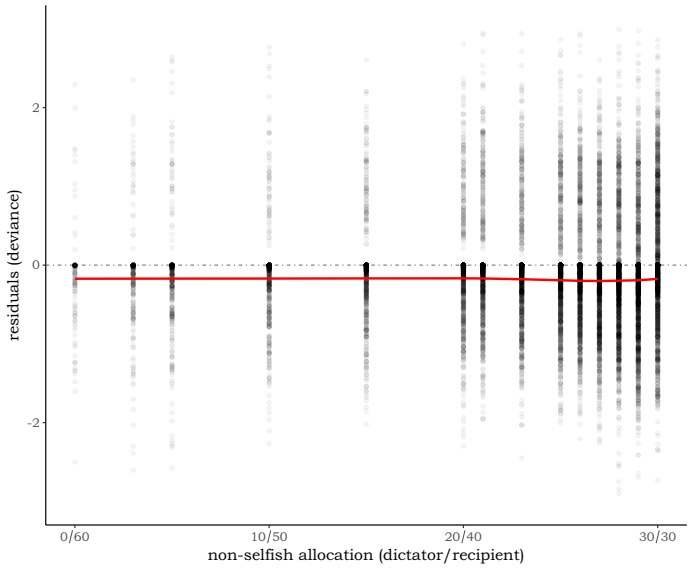

Figure 6.10: Locally weighted scatter-plot smoothing of deviance residuals for disadvantageous mini-DGs.

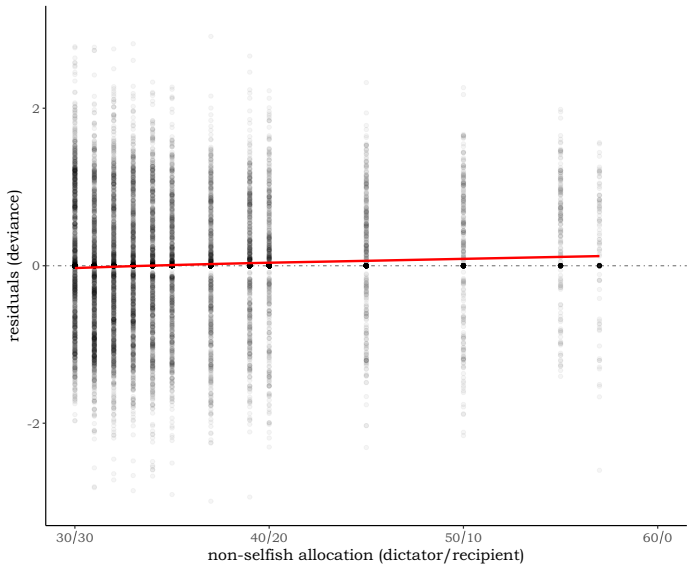

Figure 6.11: Locally weighted scatter-plot smoothing of deviance residuals across all individual regressions in polynomial regression.

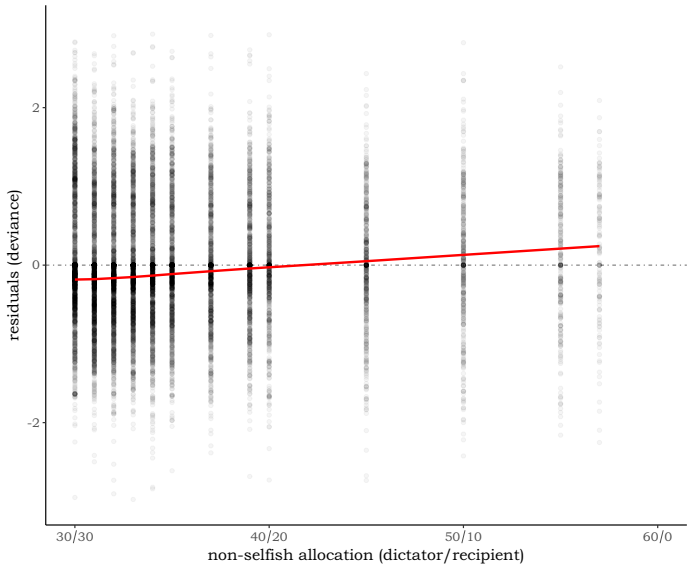

Figure 6.12: Locally weighted scatter-plot smoothing of deviance residuals across all individual regressions in concave norm regression.

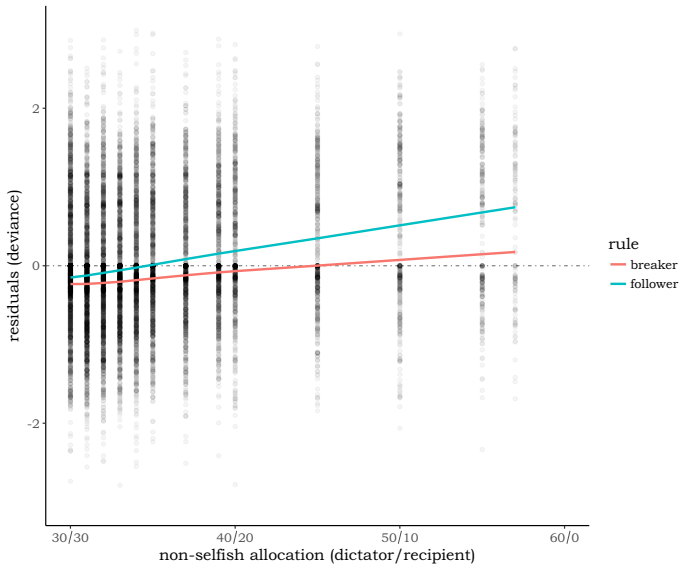

Figure 6.13: Locally weighted scatter-plot smoothing of deviance residuals across all individual regressions: regressions for rule-followers and rule-breakers.

6.4 Chapter 4: Supplementary Materials and Methods

6.4.1 Original Preregistration

We report here the main hypotheses of the original pre-registered protocol and the related figures and tables in the main text. In addition, we list any of the manuscript's departures from the protocol.

- Hypotheses:
  1. H1a: Participants will recall being more altruistic than they actually were. → Figure 4.5A
  2. H1b: Altruistic decisions will be remembered with higher accuracy than selfish or spiteful decisions → Table 4.2
  3. H1c: Participants with higher moral standards will display stronger memory biases than participants with lower moral standards → Could not be tested due to strong homogeneity in fairness self-report ratings
  4. H1d: Memory performance is worse in fairness violators (defined as participants showing a discrepancy between moral standards and actual choices) than in fairness upholders → Figure 4.4
  5. H2a: Contextual information that are relevant for choosing will be remembered with higher accuracy when the output of the decision is altruistic rather than selfish or spiteful. → Figure 4.4
  6. H2b: Contextual information that are relevant for choosing will be remembered with lower accuracy in case of violation of the individual fairness threshold → not reported, but results do not differ from hypothesis H2a
  7. H2c: Contextual information that are not relevant for choosing will be remembered with higher accuracy when the output of the decision is altruistic rather than selfish or spiteful. → Figure 4.4
  8. H2d: Contextual information that are not relevant for choosing will be remembered with lower accuracy in case of violation of the individual fairness threshold → not reported, but results do not differ from hypothesis H2c
- Indices:
  1. We changed the measure of objective fairness violation as it was based on computational models (see Analyses below). To define objective fairness violations we counted instead the number of participants' non-equitable choices.
  2. Memory accuracy: false positives and false negatives are presented in Supplementary Figure 6.14.
- Analyses:

1. General Linear Mixed Effects Models (GLMM): We did not include response time as a predictor variable as response times and choices are co-occurring events in the experiment. We have instead included confidence and choice frequency as covariates.
2. Computational Modelling: We fit the same cognitive models of Chapter 2 to participants' choices. However, we do not report the results in the thesis because estimates of participants' social attitude were extreme (very high α) and homogeneous across the sample due to participants' extremely prosocial behaviour in the task. This being said, we find that fairness upholders recall having had a more prosocial attitude in the task than their actual attitude as estimated from their choices, in accordance to results in Figure 4.5. Results are available upon request.
3. Confidence comparison between fairness violators and fairness upholders has been moved to Supplementary Material 6.4.8.

6.4.2 Allocation Selection

Allocations are sorted from an allocation space with points for the participant π_y on the x axis and points for the other π_o on the y axis. Allocations were sorted from circumferences of radius $r \in \{20, 40, 60, 80\}$ and centre $(x = r, y = r/2)$. Different circumferences allowed to sort several allocations without overlapping numbers, which could have acted as a confound for the memory task. Allocations were sorted according to the following requirements (20):

$$(r-1)^2 \leq (\pi_y - r)^2 + (\pi_o - r/2)^2 \leq (r+1)^2, \pi_y \geq r, \pi_o > 0, \tag{20}$$

Allocations in the fairness self-report are generated adopting the same procedure around a circumference of radius $r = 1$, with the sole difference that values are expressed in pounds and not in points.

Resource allocation trials were created by pairing allocations on the same circumference, thus creating four different subsets of trials. Allocations pairs in the different subsets had similar point trade-offs, so that choices represented clear fairness dilemmas, while changing only the magnitude of points allocated. See the Supplementary Spreadsheet at osf.io/g5u73/ for a list of the allocation pairs used during the task.

6.4.3 Fairness self-report allocations

6.4.4 Subjective Fairness

In each choice of the resource-allocation game we define which of the two alternatives is the most fair option according to the participant's subjective moral standard. To this end, we need to compute a distance between the two alternatives and the participant's answer in the fairness self report. Since all allocations are sorted around circumferences with centre $(x = r, y = r/2)$, we assume that it is possible

Allocation							
self	other	self	other	self	other	self	other
2.00	0.50	1.90	0.90	1.60	1.30	1.20	1.50
2.00	0.55	1.90	0.95	1.55	1.35	1.15	1.50
2.00	0.60	1.85	1.00	1.50	1.35	1.10	1.50
2.00	0.65	1.85	1.05	1.45	1.40	1.05	1.50
2.00	0.70	1.80	1.10	1.40	1.40	1.00	1.50
1.95	0.75	1.75	1.15	1.35	1.45		
1.95	0.80	1.70	1.20	1.30	1.45		
1.95	0.85	1.65	1.25	1.25	1.45		

Table 6.6: Allocations from the self-report measure, ordered from more selfish to more prosocial (increasing gains for the other), from top to bottom, left to right.

to identify each allocation simply by its respective position on its circumference, which we express as an angle α (21):

$$\alpha = \text{atan}\frac{\pi_o - r/2}{\pi_y - r}, \tag{21}$$

Where π_y are the points for the participant and π_o are the points for the other. An allocation with $\alpha = 0^\circ$ gives more points to the participant relative to all other allocations on the same circle. As α increases ($\alpha > 0^\circ$), more points are allocated to the other whereas less points are given to the participant (pro-social allocations); as α decreases ($\alpha < 0^\circ$), less points are allocated to both players (antisocial allocations). We measure the angle α_{fairness} of the allocation chosen in the self-report, and the angles α_1 and α_2 of the two allocations presented in each trial of the resource-allocation game; we then compute the distance of the two allocations from α_{fairness}. To measure subjective violations of fairness, we consider choices in the resource-allocation game where the participant prefers the allocation more distant from the fairness reference point(22):

$$\max\left(|\alpha_1 - \alpha_{\text{fairness}}|, |\alpha_2 - \alpha_{\text{fairness}}|\right), \tag{22}$$

For instance, let us assume that a participant's moral standard in the fairness self-report is £1.40 for self and £1.40 for other, which has an angle $\alpha_{\text{fairness}} \approx 66^\circ$. In one trial of the resource-allocation game, the participant observes the following two allocations which are drawn from the circle of $r = 20$: 20 points for self and 31 for the other (allocation 1, with angle $\alpha_1 = 90^\circ$), 39 points for self and 2 for the other (allocation 2, with angle $\alpha_2 = -23^\circ$). Since α_{fairness} is closer to α_1 than to α_2, then choosing the second allocation over the first one would count as a fairness violation.

6.4.5 Symbol Selection

Before each choice in the resource-allocation game, participants observed a symbol sorted from a predefined set. Symbols were 90° clockwise-rotated ideograms

of the Nuosu language, and differed from trial to trial. Given that in the Irrelevant Context participants had to recall whether or not they observed a symbol in the resource-allocation game, symbols were pre-selected based on a separate pilot experiment where participants rated what symbols were similar and what were not. Similarity ratings served to select symbols that could not easily be mistaken for each other. The final set was composed of 70 ideograms (Supplementary Spreadsheet at https://osf.io/g5u73/), 50 of which were paired randomly for each participant to allocations in the resource-allocation game, whereas the remaining 20 were used as decoy in the following cued recall memory task for the Irrelevant Context condition.

6.4.6 Relevant Context Allocations

In the Relevant Context condition, participants were presented with pairs of allocations present in the resource-allocation game and some entirely new allocation pairs, and had to recall whether they saw the presented pair or not. The new allocations were sampled from the same circumferences used for the resource-allocation game. To decrease recall difficulty, however, no new allocation presented the same points for self π_y or for the other π_o of an old allocation. In addition, whereas allocations in the game were paired with allocations on the same circumference, new pairs matched allocations from different circumferences. The list of new allocation pairs for the Relevant Context condition is shown in a Supplementary Spreadsheet at osf.io/g5u73/.

6.4.7 ROC curves for old/new recognition tasks

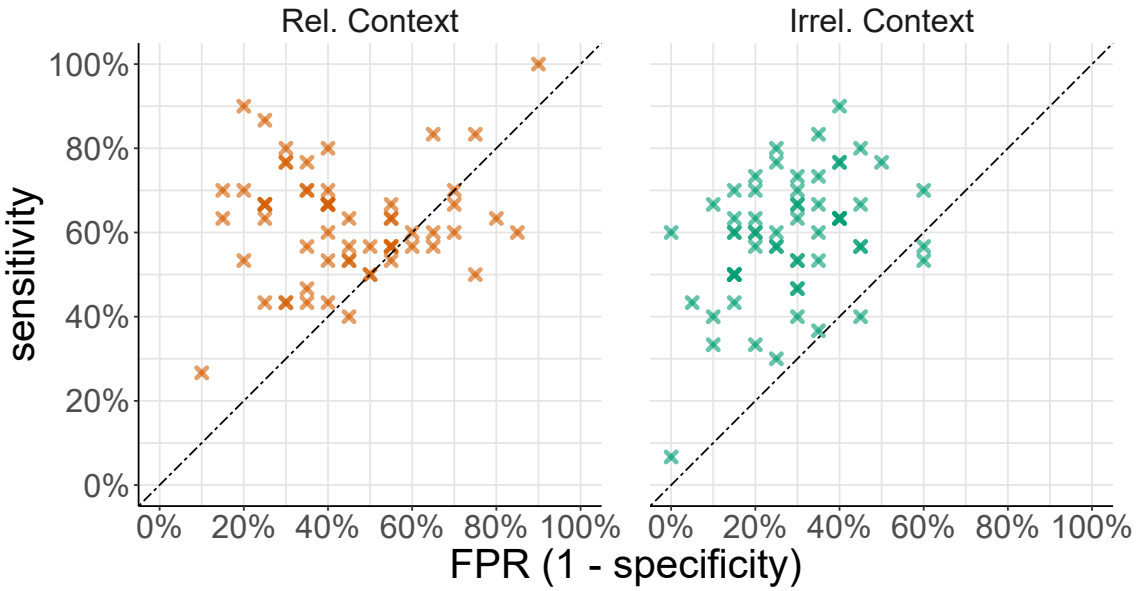

Figure 6.14: Receiver Operating Characteristic (ROC) curves at a sample level for the Relevant Context and Irrelevant Context conditions, each cross represents a participant. on the x axis is the False Positive Rate (FPR, proportion of new stimuli reported as old), on the y axis is sensitivity (recognising correctly an old stimulus as old). A participant with perfect memory performance would be marked on the top left corner, whereas any deviation represents a reduction in performance (towards bottom: decreased sensitivity; towards right: increase of false positives). The diagonal dashed line represents chance performance when accounting for responses in both 'new' and 'old' trials. Keep note however that results and analyses reported in the main text refer only to 'old' trials, and therefore on differences along the y axis; our research question in fact specifically asks whether participants have a biased recall of the already observed (old) information.

6.4.8 between-subject analyses of confidence

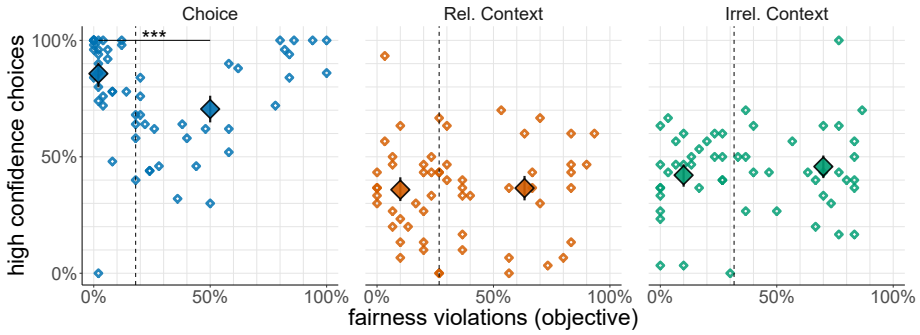

Figure 6.15: Percentage of high confidence answers by percentage of fairness violations: participants in each condition are divided based on the same median split (dashed line) as in Figure 4.4. The similarity with the test using memory accuracy as the dependent variable can be explained by the strong correlation between confidence and accuracy. Error bars indicate *t*-adjusted, 95% Gaussian confidence intervals. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

References

Agerström, J., Carlsson, R., Nicklasson, L., and Guntell, L. (2016). Using descriptive social norms to increase charitable giving: The power of local norms. *Journal of Economic Psychology*, 52:147–153.

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

Anderson, A. A., Yeo, S. K., Brossard, D., Scheufele, D. A., and Xenos, M. A. (2016). Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research*, 30(1):156–168.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of memory and language*, 49(4):415–445.

Anderson, M. C. and Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends in Cognitive Sciences*, 18(6):279–292.

Anderson, M. C. and Levy, B. J. (2009). Suppressing unwanted memories. *Current Directions in Psychological Science*, 18(4):189–194.

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

Andreozzi, L., Ploner, M., Soraperra, I., et al. (2013). Justice among strangers. on altruism, inequality aversion and fairness. Technical report, Cognitive and Experimental Economics Laboratory, Department of Economics ....

Apps, M. A. J. and Ramnani, N. (2017). Contributions of the medial prefrontal cortex to social influence in economic decision-making. *Cerebral Cortex*, 27(9):4635–4648.

Arendt, H. (2006). *Eichmann in Jerusalem: A Report on the Banality of Evil*. Penguin.

Attanasi, G. and Boun My, K. (2016). Jeu du dictateur et jeu de la confiance: préférences distributives vs préférences dépendantes des croyances. *L'Actualité économique*, 92(1-2):249–287.

Attanasi, G., Hopfensitz, A., Lorini, E., and Moisan, F. (2014). The effects of social ties on coordination: conceptual foundations for an empirical analysis. *Phenomenology and the cognitive sciences*, 13(1):47–73.

Attanasi, G., Hopfensitz, A., Lorini, E., and Moisan, F. (2016). Social connectedness improves co-ordination on individually costly, efficient outcomes. *European Economic Review*, 90:86–106.

Baader, M. and Vostroknutov, A. (2017). Interaction of reasoning ability and distributional preferences in a social dilemma. *Journal of Economic Behavior & Organization*, 142:79–91.

Bandura, A. (1986). Social foundations of thought and action. *Englewood Cliffs, NJ*, 1986.

Bardsley, N. (2007). Dictator game giving: altruism or artefact? *Experimental Economics*, 11(2):122–133.

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

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Battigalli, P. and Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144(1):1–35.

Battigalli, P. and Dufwenberg, M. (2019). Psychological game theory. Technical report, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

Battigalli, P., Dufwenberg, M., and Smith, A. (2019). Frustration, aggression, and anger in leader-follower games. *Games and Economic Behavior*, 117:15–39.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Benoit, R. G. and Anderson, M. C. (2012). Opposing mechanisms support the voluntary forgetting of unwanted memories. *Neuron*, 76(2):450–460.

Bernheim, B. D. (1994). A theory of conformity. *Journal of political Economy*, 102(5):841–877.

Bhati, A. and McDonnell, D. (2019). Success in an online giving day: The role of social media in fundraising. *Nonprofit and Voluntary Sector Quarterly*, page 0899764019868849.

Bhatia, S. (2018). Semantic processes in preferential decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Bhatia, S. and Loomes, G. (2017). Noisy preferences in risky choice: A cautionary note. *Psychological review*, 124(5):678.

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

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Blanco, M., Engelmann, D., Koch, A. K., and Normann, H.-T. (2014). Preferences and beliefs in a sequential social dilemma: a within-subjects analysis. *Games and Economic Behavior*, 87:122–135.

Bogaert, S., Boone, C., and Declerck, C. (2008). Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology*, 47(3):453–480.

Bolton, G. E. (1991). A comparative model of bargaining: Theory and evidence. *The American Economic Review*, pages 1096–1136.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.

Bolton, G. E. and Zwick, R. (1995). Anonymity versus punishment in ultimatum bargaining. *Games and Economic behavior*, 10(1):95–121.

Brañas-Garza, P. (2007). Promoting helping behavior with framing in dictator games. *Journal of Economic Psychology*, 28(4):477–486.

Brosig, J., Riechmann, T., and Weimann, J. (2007). Selfish in the end?: An investigation of consistency and stability of individual behavior.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Calluso, C., Tosoni, A., Fortunato, G., and Committeri, G. (2017). Can you change my preferences? effect of social influence on intertemporal choice behavior. *Behavioural Brain Research*, 330:78–84.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

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

Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B., and Tyran, J.-R. (2013). Give and take in dictator games. *Economics Letters*, 118(2):280–283.

Carlson, R. W., Marechal, M., Oud, B., Fehr, E., and Crockett, M. (2018). Motivated misremembering: Selfish decisions are more generous in hindsight.

Carlsson, F., Johansson-Stenman, O., and Nam, P. K. (2014). Social preferences are stable over long periods of time. *Journal of Public Economics*, 117:104–114.

Cason, T. N. and Mui, V.-L. (1998). Social influence in the sequential dictator game. *Journal of Mathematical Psychology*, 42(2-3):248–265.

Chang, D., Chen, R., and Krupka, E. (2018). Rhetoric matters: A social norms explanation for the anomaly of framing. *Games and Economic Behavior*.

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

Charpentier, C. J. and O'Doherty, J. P. (2018). The application of computational models to social neuroscience: promises and pitfalls. *Social neuroscience*, 13(6):637–647.

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

Chlaß, N. and Moffatt, P. G. (2012). Giving in dictator games: Experimenter demand effect or preference over the rules of the game? Technical report, Jena Economic Research Papers.

Chung, D., Christopoulos, G. I., King-Casas, B., Ball, S. B., and Chiu, P. H. (2015). Social signals of safety and risk confer utility and have asymmetric effects on observers' choices. *Nature Neuroscience*, 18(6):912–916.

Cialdini, R. B. and Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55(1):591–621.

Cooper, L. A. (1975). Mental rotation of random two-dimensional shapes. *Cognitive psychology*, 7(1):20–43.

Cox, J. C., List, J. A., Price, M., Sadiraj, V., and Samek, A. (2016). Moral costs and rational choice: Theory and experimental evidence.

d'Adda, G., Drouvelis, M., and Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, 62:1–7.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

De Oliveira, A. C., Eckel, C., and Croson, R. T. (2012). The stability of social preferences in a low-income neighborhood. *Southern Economic Journal*, 79(1):15–45.

De Quidt, J., Haushofer, J., and Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302.

Depue, B. E., Curran, T., and Banich, M. T. (2007). Prefrontal regions orchestrate suppression of emotional memories via a two-phase process. *science*, 317(5835):215–219.

Detert, J. R., Treviño, L. K., and Sweitzer, V. L. (2008). Moral disengagement in ethical decision making: a study of antecedents and outcomes. *Journal of Applied Psychology*, 93(2):374.

Devaine, M. and Daunizeau, J. (2017). Learning about and from others' prudence, impatience or laziness: The computational bases of attitude alignment. *PLOS Computational Biology*, 13(3):e1005422.

Duchowski, A. T. (2007). *Eye Tracking Methodology: Theory and Practice*. Springer.

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

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution. 94(4):857–869.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191.

Fehr, E. and Krajbich, I. (2014). Social preferences and the brain. In *Neuroeconomics*, pages 193–218. Elsevier.

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

Festinger, L. (1962). *A theory of cognitive dissonance*, volume 2. Stanford university press.

Fetchenhauer, D. and Huang, X. (2004). Justice sensitivity and distributive decisions in experimental games. *Personality and Individual Differences*, 36(5):1015–1029.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178.

Fleming, P. and Zizzo, D. J. (2014). A simple stress test of experimenter demand effects. *Theory and Decision*, 78(2):219–231.

Fleming, P. and Zizzo, D. J. (2015). A simple stress test of experimenter demand effects. *Theory and Decision*, 78(2):219–231.

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

Franzen, A. and Pointner, S. (2012a). Anonymity in the dictator game revisited. *Journal of Economic Behavior & Organization*, 81(1):74–81.

Franzen, A. and Pointner, S. (2012b). The external validity of giving in the dictator game. *Experimental Economics*, 16(2):155–169.

Garvert, M. M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T. E. J., and Dolan, R. J. (2015). Learning-induced plasticity in medial prefrontal cortex predicts preference malleability. *Neuron*, 85(2):418–428.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and economic Behavior*, 1(1):60–79.

González-Vallejo, C. (2002). Making trade-offs: A probabilistic and context-sensitive model of choice behavior. *Psychological Review*, 109(1):137.

Griesinger, D. W. and Livingston Jr, J. W. (1973). Toward a model of interpersonal motivation in experimental games. *Behavioral science*, 18(3):173–188.

Gronau, Q. F., Ly, A., and Wagenmakers, E.-J. (2017). Informed Bayesian t-tests. *arXiv preprint arXiv:1704.02479*.

Guala, F. and Mittone, L. (2010). Paradigmatic experiments: The dictator game. *The Journal of Socio-Economics*, 39(5):578–584.

Gul, F. and Pesendorfer, W. (2006). Random expected utility. *Econometrica*, 74(1):121–146.

Güroğlu, B., Will, G.-J., and Crone, E. A. (2014). Neural correlates of advantageous and disadvantageous inequity in sharing decisions. *PLoS One*, 9(9).

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, 3(4):367–388.

Haley, K. J. and Fessler, D. M. T. (2005). Nobody's watching? *Evolution and Human Behavior*, 26(3):245–256.

Hanslmayr, S. and Staudigl, T. (2014). How brain oscillations form memories—a processing based perspective on oscillatory subsequent memory effects. *Neuroimage*, 85:648–655.

Harbaugh, W. T., Mayr, U., and Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, 316(5831):1622–1625.

Hare, T. A., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P., and Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience*, 30(2):583–590.

Haruno, M. and Frith, C. D. (2010). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nature neuroscience*, 13(2):160–161.

He, L., Golman, R., and Bhatia, S. (2019). Variable time preference. *Cognitive psychology*, 111:53–79.

Henninger, F., Shevchenko, Y., Mertens, U., Kieslich, P. J., and Hilbig, B. E. (2019). Lab. js: A free, open, online study builder.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., and Tracer, D. (2005). "economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(06).

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., and Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature neuroscience*, 20(8):1142.

Hoffman, E., McCabe, K., and Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *The American economic review*, 86(3):653–660.

Huber, J., Payne, J. W., and Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of consumer research*, 9(1):90–98.

Hunt, L. T. and Hayden, B. Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. *Nature Reviews Neuroscience*, 18(3):172.

Hutcherson, C. A., Bushong, B., and Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, 87(2):451–462.

Hwang, H., Kim, Y., and Huh, C. U. (2014). Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media*, 58(4):621–633.

Ivaturi, K. and Chua, C. (2019). Framing norms in online communities. *Information & Management*, 56(1):15–27.

Izuma, K., Akula, S., Murayama, K., Wu, D.-A., Iacoboni, M., and Adolphs, R. (2015). A causal role for posterior medial frontal cortex in choice-induced preference change. *Journal of Neuroscience*, 35(8):3598–3606.

Jimenez-Buedo, M. and Guala, F. (2016). Artificiality, reactivity, and demand effects in experimental economics. *Philosophy of the Social Sciences*, 46(1):3–23.

Juechems, K. and Summerfield, C. (2019). Where does value come from? *Trends in cognitive sciences*.

Kahneman, D., Knetsch, J. L., and Thaler, R. (1986a). Fairness as a constraint on profit seeking: Entitlements in the market. *The American economic review*, pages 728–741.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1986b). Fairness and the assumptions of economics. *The Journal of Business*, 59(S4):S285.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1986c). Fairness and the assumptions of economics. *Journal of business*, pages S285–S300.

Kensinger, E. A. and Ford, J. H. (2020). Retrieval of emotional events from memory. *Annual review of psychology*, 71:251–272.

Kenward, B., Karlsson, M., and Persson, J. (2010). Over-imitation is better explained by norm learning than by distorted causal learning. *Proceedings of the Royal Society B: Biological Sciences*, 278(1709):1239–1246.

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

Kimbrough, E. and Vostroknutov, A. (2018a). A normative model of social behavior. mimeo, Chapman University and University of Trento.

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

Kimbrough, E. and Vostroknutov, A. (2019). Injunctive norms and moral heuristics. mimeo, Chapman University and Maastricht University.

Kimbrough, E. O., Miller, J. B., and Vostroknutov, A. (2018). Norms, frames, and prosocial behavior in games. mimeo, Chapman University, Bocconi University, University of Trento.

Kimbrough, E. O. and Vostroknutov, A. (2015). The social and ecological determinants of common pool resource sustainability. *Journal of Environmental Economics and Management*, 72:38–53.

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

Kitayama, S., Chua, H. F., Tompson, S., and Han, S. (2013). Neural mechanisms of dissonance: An fMRI investigation of choice justification. *NeuroImage*, 69:206–212.

Koller, M. and Stahel, W. A. (2017). Nonsingular subsampling for regression s estimators with categorical predictors. *Computational Statistics*, 32(2):631–646.

Koppen, M. (2001). Characterization theorems in random utility theory. *Smelser, NJ; Baltes, PB (ed.), International Encyclopedia of the Social & Behavioral Sciences, Section Mathematics and Computer Sciences*, pages 1646–1651.

Kouchaki, M. and Gino, F. (2016). Memories of unethical actions become obfuscated over time. *Proceedings of the National Academy of Sciences*, 113(22):6166–6171.

Kriegeskorte, N. and Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–412.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

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

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

Kuss, K., Falk, A., Trautner, P., Elger, C. E., Weber, B., and Fliessbach, K. (2013). A reward prediction error for charitable donations reveals outcome orientation of donators. *Social cognitive and affective neuroscience*, 8(2):216–223.

Kuss, K., Falk, A., Trautner, P., Montag, C., Weber, B., and Fliessbach, K. (2015). Neuronal correlates of social decision making are influenced by social value orientation—an fmri study. *Frontiers in behavioral neuroscience*, 9:40.

Lacetera, N., Macis, M., and Mele, A. (2016). Viral altruism? charitable giving and social contagion in online networks. *Sociological Science*, 3.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of economic dynamics*, 1(3):593–622.

Liberman, V., Samuels, S. M., and Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisonerâ€™s dilemma game moves. *Personality and social psychology bulletin*, 30(9):1175–1185.

Liebrand, W. B. G. (1984). The effect of social motives, communication and group size on behaviour in an N-person multi-stage mixed-motive game. *European journal of social psychology*, 14(3):239–264.

Lightner, A. D., Barclay, P., and Hagen, E. H. (2017). Radical framing effects in the ultimatum game: the impact of explicit culturally transmitted frames on economic decision-making. *Royal Society Open Science*, 4(12):170543.

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

Loewenstein, G. F., Thompson, L., and Bazerman, M. H. (1989). Social utility and decision making in interpersonal contexts. *Journal of Personality and Social psychology*, 57(3):426.

Loftus, E. F., Loftus, G. R., and Messo, J. (1987). Some facts about "weapon focus". *Law and human behavior*, 11(1):55–62.

Loomes, G., Moffatt, P. G., and Sugden, R. (2002). A microeconometric test of alternative stochastic theories of risky choice. *Journal of risk and Uncertainty*, 24(2):103–130.

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

Lotz, S., Baumert, A., Schlösser, T., Gresser, F., and Fetchenhauer, D. (2011). Individual differences in third-party interventions: How justice sensitivity shapes altruistic punishment. *Negotiation and Conflict Management Research*, 4(4):297–313.

Lotz, S., Schlösser, T., Cain, D. M., and Fetchenhauer, D. (2013). The (in) stability of social preferences: Using justice sensitivity to predict when altruism collapses. *Journal of Economic Behavior & Organization*, 93:141–148.

Lu, J. and Saito, K. (2018). Random intertemporal choice. *Journal of Economic Theory*, 177:780–815.

MacFarquhar, L. (2016). *Strangers drowning: Impossible idealism, drastic choices, and the urge to help*. Penguin.

Malmasi, S. and Zampieri, M. (2017). Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.

Mather, M. (2006). Why memories may become more positive as people age. *Memory and emotion: Interdisciplinary perspectives*, pages 135–158.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research*, 45(6):633–644.

McClintock, C. G. and Allison, S. T. (1989). Social value orientation and helping behavior 1. *Journal of Applied Social Psychology*, 19(4):353–362.

McFadden, D. L. (1976). Quantal choice analaysis: A survey. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 363–390. NBER.

Meder, D., Kolling, N., Verhagen, L., Wittmann, M. K., Scholl, J., Madsen, K. H., Hulme, O. J., Behrens, T. E. J., and Rushworth, M. F. S. (2017). Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. *Nature communications*, 8(1):1942.

Messick, D. M. and McClintock, C. G. (1968). Motivational bases of choice in experimental games. *Journal of experimental social psychology*, 4(1):1–25.

Mischel, W., Ebbesen, E. B., and Zeiss, A. M. (1976). Determinants of selective memory about the self. *Journal of consulting and clinical Psychology*, 44(1):92.

Mohan, S., Guha, A., Harris, M., Popowich, F., Schuster, A., and Priebe, C. (2017). The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence*, pages 51–56. Springer.

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., and Grafman, J. (2006). Human fronto–mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences*, 103(42):15623–15628.

Morgenstern, O. and Von Neumann, J. (1953). *Theory of games and economic behavior*. Princeton university press.

Mormann, M. M., Malmaud, J., Huth, A., Koch, C., and Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, 5(6):437–449.

Moutoussis, M., Dolan, R. J., and Dayan, P. (2016). How people use social information to find out what to want in the paradigmatic case of inter-temporal preferences. *PLOS Computational Biology*, 12(7):e1004965.

Mullen, E. and Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual review of psychology*, 67:363–385.

Murphy, R. O. and Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, 18(1):13–41.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. (2011). Measuring social value orientation. *SSRN Electronic Journal*.

Nax, H. H., Murphy, R. O., and Ackermann, K. A. (2015). Interactive preferences. *Economics Letters*, 135:133–136.

Nook, E. C., Ong, D. C., Morelli, S. A., Mitchell, J. P., and Zaki, J. (2016). Prosocial conformity. *Personality and Social Psychology Bulletin*, 42(8):1045–1062.

Oechssler, J. (2010). Searching beyond the lamppost: Let's focus on economically relevant questions. *Journal of Economic Behavior & Organization*, 73(1):65–67.

Orne, M. T. (2009). Demand characteristics and the concept of quasi-controls. *Artifacts in behavioral research: Robert Rosenthal and Ralph L. Rosnow's classic books*, 110:110–137.

Paller, K. A. and Wagner, A. D. (2002). Observing the transformation of experience into memory. *Trends in cognitive sciences*, 6(2):93–102.

Park, S. A., Goïame, S., O'Connor, D. A., and Dreher, J.-C. (2017). Integration of individual and social information for decision-making in groups of different sizes. *PLoS biology*, 15(6):e2001958.

Patil, I. (2018). *ggstatsplot: 'ggplot2' Based Plots with Statistical Details*.

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., and Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, 1(11):803–809.

Plummer, M. (2003). JAGS: a program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124. Vienna, Austria.

Plummer, M. (2019). *rjags: Bayesian graphical models using MCMC*.

Polania, R., Woodford, M., and Ruff, C. C. (2019). Efficient coding of subjective value. *Nature neuroscience*, 22(1):134.

Pryor, C., Perfors, A., and Howe, P. D. L. (2019). Even arbitrary norms influence moral decision-making. *Nature human behaviour*, 3(1):57.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Rangel, A. and Clithero, J. A. (2012). Value normalization in decision making: theory and evidence. *Current opinion in neurobiology*, 22(6):970–981.

Regenwetter, M., Cavagnaro, D. R., Popova, A., Guo, Y., Zwilling, C., Lim, S. H., and Stevens, J. R. (2018). Heterogeneity and parsimony in intertemporal choice. *Decision*, 5(2):63.

Regenwetter, M., Dana, J., and Davis-Stober, C. P. (2010). Testing transitivity of preferences on two-alternative forced choice data. *Frontiers in Psychology.*, 1.

Regenwetter, M., Dana, J., and Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, 118(1):42–56.

Regenwetter, M. and Robinson, M. M. (2017). The construct–behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review*, 124(5):533.

Richter, M. K. (2008). *Revealed Preference Theory*, pages 5569–5574. Palgrave Macmillan UK, London.

Rizio, A. A. and Dennis, N. A. (2013). The neural correlates of cognitive control: successful remembering and intentional forgetting. *Journal of Cognitive Neuroscience*, 25(2):297–312.

Rosenthal, R. and Rosnow, R. L. (1977). *People studying people: Artifacts and ethics in behavioral research*.

Rossell, S. L. and Nobre, A. C. (2004). Semantic priming of different affective categories. *Emotion*, 4(4):354.

Rotello, C. M. and Macmillan, N. A. (2007). Response bias in recognition memory. *Psychology of Learning and Motivation*, 48:61–94.

Ruff, C. C. and Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, 15(8):549–562.

Saucet, C. and Villeval, M. C. (2019). Motivated memory in dictator games. *Games and Economic Behavior*.

Sawyer, J. (1966). The altruism scale: A measure of co-operative, individualistic, and competitive interpersonal orientation. *American Journal of Sociology*, 71(4):407–416.

Sedikides, C. and Green, J. D. (2000). On the self-protective nature of inconsistency-negativity management: Using the person memory paradigm to examine self-referent memory. *Journal of personality and social psychology*, 79(6):906.

Sedikides, C. and Green, J. D. (2009). Memory as a Self-Protective Mechanism. *Social and Personality Psychology Compass*, 3(6):1055–1068.

Sedikides, C., Green, J. D., and Pinter, B. (2004). Self-protective memory. In Beike, D. R., Lampinen, J. M., and Behrend, D. A., editors, *The Self and Memory*, pages 161–179. Psychology Press, Philadelphia, PA.

Shamay-Tsoory, S. G., Saporta, N., Marton-Alper, I. Z., and Gvirts, H. Z. (2019). Herding brains: A core neural mechanism for social alignment. *Trends in cognitive sciences*.

Shiffrin, R., Lee, M., Kim, W., and Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cognitive Science: A Multidisciplinary Journal*, 32(8):1248–1284.

Shu, L. L. and Gino, F. (2012). Sweeping dishonesty under the rug: How unethical actions lead to forgetting of moral rules. *Journal of Personality and Social Psychology*, 102(6):1164–1177.

Shu, L. L., Gino, F., and Bazerman, M. H. (2011). Dishonest Deed, Clear Conscience: When Cheating Leads to Moral Disengagement and Motivated Forgetting. *Personality and Social Psychology Bulletin*, 37(3):330–349.

Smeets, P., Bauer, R., and Gneezy, U. (2015). Giving behavior of millionaires. *Proceedings of the National Academy of Sciences*, 112(34):10641–10644.

Smith, V. L. (2010). Theory and experiment: What are the questions? *Journal of Economic Behavior & Organization*, 73(1):3–15.

Sonnemans, J., Van Dijk, F., and Van Winden, F. (2006). On the dynamics of social ties structures in groups. *Journal of economic psychology*, 27(2):187–204.

Stewart, N., Reimers, S., and Harris, A. J. L. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*, 61(3):687–705.

Su, Y.-S. and Yajima, M. (2015). *R2jags: Using R to run 'JAGS'*.

Sugden, R. (1995). A theory of focal points. *The Economic Journal*, pages 533–550.

Sugden, R. (2000). Team preferences. *Economics & Philosophy*, 16(2):175–204.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Suzuki, S., Jensen, E. L. S., Bossaerts, P., and O'Doherty, J. P. (2016). Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. *Proceedings of the National Academy of Sciences*, 113(14):3755–3760.

Swann Jr, W. B., Rentfrow, P. J., and Guinn, J. S. (2003). Self-verification: The search for coherence. *Handbook of self and identity*, pages 367–383.

Tadelis, S. (2011). The power of shame and the rationality of trust. *Haas School of Business working paper*, 3(2).

Tajfel, H., Turner, J. C., Austin, W. G., and Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56:65.

Templeton, E. M., Stanton, M. V., and Zaki, J. (2016). Social norms shift preferences for healthy and unhealthy foods. *PLOS ONE*, 11(11):e0166286.

Thomsson, K. and Vostroknutov, A. (2017). Small-world conservatives and rigid liberals: Attitudes towards sharing in self-proclaimed left and right. 135:181–192.

Toelch, U. and Dolan, R. J. (2015). Informational and normative influences in conformity from a neurocomputational perspective. *Trends in Cognitive Sciences*, 19(10):579–589.

Tricomi, E., Rangel, A., Camerer, C. F., and O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature*, 463(7284):1089–1091.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

Ugazio, G., Grueschow, M., Polania, R., Lamm, C., Tobler, P. N., and Ruff, C. C. (2019). Neuro-computational foundations of moral preferences. *bioRxiv*, page 801936.

Urminsky, O. and Zauberman, G. (2015). The psychology of intertemporal preferences. *The Wiley Blackwell handbook of judgment and decision making*, 2:141–181.

van Baar, J. M., Chang, L. J., and Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature communications*, 10(1):1–14.

van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, 77(2):337–349.

Van Winden, F., Stallen, M., and Ridderinkhof, K. R. (2008). On the nature, modeling, and neural bases of social ties. *Neuroeconomics*, 20:125–159.

Voigt, K., Murawski, C., Speer, S., and Bode, S. (2019). Hard decisions shape the neural coding of preferences. *Journal of Neuroscience*, 39(4):718–726.

Volk, S., Thöni, C., and Ruigrok, W. (2012). Temporal stability and psychological foundations of cooperation preferences. *Journal of Economic Behavior & Organization*, 81(2):664–676.

Weiner, B. (1968). Motivated forgetting and the study of repression. *Journal of Personality*.

Wood, W. (2000). Attitude change: Persuasion and social influence. *Annual review of psychology*, 51(1):539–570.

Yoo, S. B. M. and Hayden, B. Y. (2018). Economic choice as an untangling of options into actions. *Neuron*, 99(3):434–447.

Zaki, J. and Mitchell, J. P. (2011). Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences*, 108(49):19761–19766.

Zaki, J., Schirmer, J., and Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psychological Science*, 22(7):894–900.

Zhou, W. and Hey, J. (2018). Context matters. *Experimental economics*, 21(4):723–756.

Zizzo, D. J. (2009). Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1):75–98.

Zizzo, D. J. (2013). Claims and confounds in economic experiments. *Journal of Economic Behavior & Organization*, 93:186–195.

Zizzo, D. J. and Fleming, P. (2011). Can experimental measures of sensitivity to social pressure predict public good contribution? *Economics Letters*, 111(3):239–242.