Moral Rules and Social Preferences in Cooperation problems

BY ERNESTO M. GAVASSA-PÉREZ*

7TH NOVEMBER, 2022

CLICK HERE FOR THE LATEST VERSION

Abstract. Models of prosocial preferences allow a trade-off between one's own social orientation and one's selfishness. This paper presents a new model where trade-offs are explicitly excluded by moral constraints to the choice space of individuals. I present two moral constraints formulated from an impartial spectator standpoint: blame avoidance and praise seeking. I test at the individual level, and against canonical models of social preferences, the extent to which moral rules can predict attitudes to cooperation in public goods games. I find that moral rules successfully predict cooperation, and that are complementary to social preferences in our understanding of social behaviour. JEL codes. C91, D01, D91. H41

Keywords. Behavioural Economics; Constraints; Experiments; Morality; Public goods; Social Preferences.

---

* Gavassa-Pérez: University of Nottingham, School of Economics, Sir Clive Granger Building, University Park, Nottingham, NG7 2RD, United Kingdom (e-mail: ernesto.gavassaperez1@nottingham.ac.uk). Support from the School of Economics in the form of a PhD Scholarship, and of CeDEx for funding the experiment, is gratefully acknowledged. The author also thanks Robin Cubitt, Astri Drange Hole, Simon Gächter, Konstantinos Ioannidis, Alex Possajennikov, Rui Silva, Silvia Sonderegger, Chris Starmer, and Bertil Tungodden for in-depth comments of earlier drafts, and Robert Sugden, Martin Sefton, and Vernon Smith, for constructive comments on the experimental design and methodology that we hereby present. Additionally, the author wants to thank the audience in the ESA Europe (Bologna) Conference, the ESA Job Market Candidates seminar series attendees, the UAM's seminar series attendees, the IRNEP workshop attendees, the CeDEx colleagues attending various seminars on this project, and all the faculty members at FAIR (NHH, Norway) that gave very constructive comments on the work during my stay; especially, but not only, to Alexander Cappelen, Erik Sørensen, and Pablo Soto-Mota.

Classical economic theory assumes that people maximise their own utility, and that the content of that utility is more selfish than, on retrospect and given experimental data, it is. Since the 1980's, behavioural economics provided ample evidence showing that people rejected unfair offers in ultimatum games, donated money in dictator games, and did not

undersupply public goods[1]. To rationalise this behaviour, in the last forty years economists have retained the assumption that people maximise their own utility and, rather, challenged the content of the utility function that is supposed to govern subjects' choices[2]. Yet, moral philosophy has, for long, suggested that moral choices are driven at least partially by strictly disinterested motives (that is, detached from one's own pleasures). Indeed, the two greatest and opposing contemporary moral philosophies – utilitarianism and deontology – coincide on this point: utilitarianism is based on strict impartiality wen making assessments over choices that influence the society, and Kant posits that only behaviour coming from duties, and not from one's own inclinations, has moral worth[3].

Economists are not alien to this literature, and some have tried to introduce them in the economics discourse through concepts like '*Ethical Preferences*', strictly independent from traditional preferences (see Harsanyi, 1955); '*Commitment*', which is supposed to be crucially counter-preferential (see Sen, 1977); or '*Sympathy*' (see Smith and Wilson, 2017, and 2019), drawing on Adam Smith (1982) and the literature within experimental economics presenting the concept of impartial spectators (see Konow, 2009 and 2012)[4]. In this paper I build on this literature and present a new model of decision-making where the concept of morality is introduced via constraints to the choice set of an individual, who will maximise over the set of actions that survive an agents' moral constraints. In doing so, I not only provide a new representation of altruistic choice but present a substantially different model that departs from models of social preferences in two key aspects.

First, whereas social preferences are based on trade-offs between different courses of action, based on different social characteristics of the actions (the inequality they generate, how reciprocal my action is, and so on), my new model captures morality as a rigid concept that strictly prohibits any trade-off between an agent's moral values and their monetary

[1] For initial evidence of behaviour in ultimatum games, see Güth et al. (1982), for initial evidence of behaviour in dictator games see Forsythe et al. (1994), and for an early survey on behaviour in public goods games see Ledyard (1995). More generally, for a survey of altruistic behaviour and its link to prosocial preferences see Camerer (2003, ch.2).

2 For a survey of theoretical models of prosocial preferences, see Sobel (2005) and Cooper and Kagel (2017). For theoretical models built to accommodate this evidence, see Fehr and Schmidt (1999) and Bolton and Ockelfels (2000) for inequality aversion motives, Sugden (1984), Rabin (1993), Dufwenberg and Kirchsteiger (2004), and Cox et al (2007) for reciprocity motives, Falk and Fischbacher (2006) for a mixture of inequality aversion and reciprocity motives, Charness and Rabin (2002) for a mixture of social efficiency and maximin motives, Batigalli and Dufwenberg (2007) for guilt aversion motives, McKelvey and Palfrey (1995) for confusion motives, Cappelen et al. (2007) for egalitarian, libertarian and liberal egalitarian concerns, Andreoni (1990) for impure altruistic concerns, and Levine (1998) for spiteful concerns. For different theoretical approaches trying to incorporate morality or normative concerns within a utility framework, one can see, most notably, Brekke et al. (2003); Bénabou and Tirole (2006, and 2011); Roemer (2010); Alger and Weibull (2013); and, more recently, Masclet and Dickinson (2019).

3 Protoutilitarianism was deeply belligerent with the idea of equating morality to one's self interest, howsoever broadly construed. See, for instance, Hutcheson's (2002 and 2004) distinction of Subordinate and Ultimate Desires, and David Hume (1739, and especially 1983's Appendix II). Also, see Kant's (2012) 4:398-399 passage on one's own Inclinations and Duties, and how morality is solely based on the latter.

4 For other discussions on the topic, one can refer to Prelec (1991); Tungodden (2004); or Vanberg (2008).

payoff. Evidence in moral psychology has accumulated suggesting that this way of thinking indeed represents the way some subjects think, as they do not want to even consider doing actions that they believe to be very wrong[5].

Second, whereas social preferences can be seen as a slow, reflective path to cooperation, assuming subjects spend time deliberating and calculating benefits and costs to their own pleasure of different courses of action, moral constraints are supposed to kick in quickly, quasi-subconsciously, restricting the set of actions an individual can take. So to speak, social preferences are de facto increasing the cognitive efforts subjects must face to solve a problem, whilst at the same time in economics we have been presenting evidence documenting that subjects normally do abide by rules of thumb, or heuristics and biases, when making decisions (see, for instance, Herbert Simon, 1995; Tversky and Kahneman, 1974; Gigerenzer and Selten, 2002; and Gilovich et al., 2002, to name a few). As of lately, moral psychology and neuroscience have proposed dual path models to moral thinking, and argue that i) quick responses to moral dilemmas tend to be more deontological than slow responses[6]; ii) own-conscious emotions and disinterested, other-conscious emotions lie on different places in the brain[7]; and iii) different brain regions are associated with either calculating value or evolutionary-driven behaviour[8]. Evolutionary psychology has also

5 For work on protected values, see Baron and Spranca (1997), and Baron (2017). For work on taboo trade-offs, see Tetlock (2003), Schoemaker and Tetlock (2012), and Tetlock et al. (2017). For work on moral conviction, and how individuals universalize their moral values, see Skitka et al. (2005), and Skitka (2010).

6 See Zajonc and Markus (1982), and Railton (2014) for the differential influences of emotions and cognition in behaviour; see Evans (2008) for a discussion on dual-processes in social cognition; see Harenski et al. (2010) for how moral deliberation and intuition are different neural constructs; see Rand et al. (2012), Declerck et al. (2013), and Rand et al. (2014), for dual-processing in co-operation; see Calvillo and Burgeno (2015), and Feng et al. (2015) for dual-processing in the ultimatum game; and see FeldmanHall et al. (2013) for the differential functions of the temporoparietal junction and the ventromedial prefrontal cortex in decision-making. See also Duc et al. (2013) for how sacred values and non-sacred values correlate differently with intuition and deliberation; see Kluver et al. (2014) for a discussion of dual processing related to economics and ethics. Finally, see Hallsson et al. (2018) for a review on the evidence of fairness and dual-processing mechanisms. Of especial relevance to morality is Antonio Damasio's (1995) Somatic Marker hypothesis, and how quick subconscious information that manifests itself to the sensitive world can constrain what we believe possible. See Bechara and Damasio (2005) for a coverage of their theory in an economic journal.

7 For the rising influence of emotions in decision-making, see Dukes et al (2021). For a generic coverage of the roles of emotions in moral judgments, see, as well, Pizarro (2000), Nichols (2002), Shweder et al (2008), and Zahn et al. (2013). For research on neuroscience and self-conscious and other conscious emotions, see Beer et al. (2003), Finger et al. (2006), Takahasi et al. (2004), Algoe and Haidt (2009), Green et al. (2010), Ebstein et al. (2011), Morey et al. (2012), Yu et al. (2013), Jankowski and Takahashi (2014), FeldmanHall and Mobbs (2015), Fox et al. (2015), Bastin et al. (2016), Gilead et al. (2016), Moll et al. (2016), Bas-hoogendam et al. (2017), and Saarimäki et al. (2018). For how self- versus other agency plays a differential role in the link between morality and co-operation, see Tomasello and Vaish (2013). For how emotions are different from deliberative thinking, and neuroscientific evidence of this differential path in moral judgments, see Wiech et al. (2013). For a specific discussion of empathy, which is the emotion closest to Adam Smith's sympathy – the construct related to the impartial spectator position my theory is based on – see Jackson et al. (2005), Decety and Jackson (2006), Decety and Lamm (2006), de Waal (2009), Decety and Ickes (2009), Singer and Lamm (2009), Zahn et al. (2009), Gleidhgerrcht and Young (2013), Gonzalez-Liencres et al. (2013), Klimecki and Singer (2013), and Stevens and Katherine (2021). For the relation between moral emotions and behaviour, see Keltner and Haidt (2001) and Tangney et al. (2007). For how moral and transgressions and transgressions of social norms differ, see Song et al. (1987).

8 For the relation of the amygdala to the social life, see Angrilli et al. (1996), Adolphs et al. (1998), Adolphs et al. (2001), Anderson and Phelps (2001), Adolphs et al. (2002), Sander et al. (2003), Adolphs and Spezio (2006), Mendez (209), Adolphs (2010) and Bickart et al. (2014). For literature related to the differential role of the amygdala and the ventromedial prefrontal cortex in social and moral cognition, see Bechara et al. (1999), Bechara et al. (2000), Berthoz et al. (2002), Bechara et al. (2003), Blair et al. (2006), Blair (2007), Haruno et al. (2014), and Shenhav and Greene (2014). For a broader literature on

provided evidence showing that humans are innately tuned for social doing without heavy reflection[9], and animal and children's prosocial behaviour has accumulated, suggesting that if human behaviour has its basis in ontogeny and phylogeny, then it is better understood as being driven by a quick, subconscious process available, as well, to children and animals[10]. The model I present herein fills a gap in the prosocial literature within economics as it provides a model for the '*quick, intuitive*' system making prosocial choices, thereby synthesizing the literature in other disciplines in a model of decision-making that can make precise predictions in economically relevant decision situations, and that can be seen as complementary to all the models of social preferences already available in the literature.

In this paper I additionally bring the new theory of moral rules to the test, not only against the void but against five canonical models of social preferences, plus the classical benchmark of a selfish Homo Economicus. To do so, I use experimental methods to investigate the predictive power of moral rules and social preferences in predicting behaviour in two public goods games. The two public goods games are isomorphic in all respects but in the marginal per capita return to contribution (i.e., m henceforth. How much a subject gets from the public good). On one game, which I coin the *social dilemma game*, I set $m = 0.6 < 1$, meaning a subject is better off by free riding. In contrast, on another game, which I coin the *common interest game*, I set $m = 1.2 > 1$, meaning a subject is better off by fully contributing to the public good[11]. By focusing on best response functions

morality and neuroscience, see Moll et al. (2002), Greene (2003), Moll et al. (2003), Moll et al. (2005), Killen and Smetana (2008), Lindquist et al. (2012), Young and Dungan (2012), Eres et al. (2018) and Eslinger et al. (2021). For a broader literature of brain regions and social behaviour, see Anderson et al. (1999), Sanfey et al. (2003), Moll et al. (2006), Sanfey (2007), Harensky et al. (2010), Trevor and Sanfey (2010), Gospic et al. (2011), Rilling and Sanfey (2011), Gospic et al. (2013), Hinterbuchinger et al. (2018), and Massen et al. (2019).

9 See Cosmides and Tooby (1994) for an article in an economic journal defending the evolutionary explanation of social behaviour; see Barkow et al. (1992) for a book-length approach to evolutionary psychology and the adaptation of humans for prosocial behaviour; and see Tooby and Cosmides (1990), Cosmides and Tooby (1992), Öhman et al. (2001), Cosmides et al. (2005), Delton et al. (2012), and Cosmides and Tooby (2013) for work on evolutionary psychology related to the evolutionary basis of human prosocial behaviour. See also Haidt (2001) for a model proposing intuition as the main basis of moral judgments; and see Moll et al. (2002), Chang et al. (2015), Basile et al. (2020) and Zahn et al. (2020) for work on how morality is based on brain structures favoured by evolution.

10 See Dunbar (1991) for the social meaning of grooming in primates; see de Waal (1997), and de Waal and Brosnan (2006) for evidence of reciprocity in primates; see Brosnan and de Waal (2003), Wynne (2004), Brosnan (2006), van Wonkenten et al (2007), Brosnan et al. (2010) for evidence of inequality aversion in monkeys; see Kimberly and de Waal (2000), Mendres and de Waal (2000), Kappeler and Schaik (2006), and Brosnan (2011) for evidence of cooperation among primates; see Warneken et al (2007), Brosnan (2013), and hopper et al. (2013) for instances of altruistic and fair behaviour among chimpancees; and see Milinski (2013), Proctor et al. (2013a), and Proctor et al. (2013b) for primates fairness in the ultimatum game. See Allison et al. (2000), Barger et al. (2014) and Wittman et al. (2018) for the neural correlates of social behaviour favouring an evolutionary explanation of such behaviour. See Warneken and Tomasello (2007 and 2009), LoBue et al. (2011), House et al. (2012), Wittig et al. (2013), and McAuliffe et al. (2015) for evidence of prosocial behaviour in children. For phylogenetic adaptations in the human brain, see Allman et al. (2002). For how the human brain is specialized to recognized social stimuli, and how moral cognition is organized in the brain, see Allison et al. (2000), and Moll et al. (2008). See also Delton et al. (2011) for how reciprocity under uncertainty can be an evolutionarily stable rationale for one-shot behaviour.

11 One can refer to Bohm (1972); Dawes et al. (1977); Marwell and Ames (1979) and Isaac et al. (1984) for early evidence on contributions to social dilemmas; and see also Saijo and Nakamura (1995), Palfrey and Prisbey (1997); Brunton et al. (2001); Brandts et al. (2004); and Reuben and Riedl (2009) for evidence on common interest games. For seminal

of the simultaneous versions of both games (i.e., what is called *strategy method* within the experimental economics literature, following Fischbacher et al., 2001), rather than on unconditional contributions, I manage to provide a clean theoretical ground for the test, as each of the five social preferences I consider make five different predictions about the joint best response functions in both games[12].

A key feature of the empirical test is that it is within-subjects, meaning that the test is done at the individual level. To achieve this, I extend a revealed preference approach, based on the methodology presented in Blanco et al. (2011), and present several games to each subject that allow me to restrict the range of feasible parameter values for each social preference if subject's choices are well-behaved (i.e., transitive). These *parameter-elicitation games* are designed so that the range of values elicited at the individual level allows us to make an individual, unambiguous prediction for the best response functions in both the social dilemma and the common interest game. Additionally, I ask subjects to state their moral judgments of each strategy combination of both games from an impartial spectator viewpoint (i.e., they make the judgments of other people), and I use those moral judgments to make individual-level predictions about the best response functions of both public goods based on two moral rules: *blame avoidance*, or a principle that states that subjects ought to avoid doing what they perceive as blameworthy actions from an impartial perspective, and *praise seeking*, or a principle that states that subjects ought to seek doing what they perceive as the most praiseworthy actions from an impartial perspective. By eliciting parameters, moral judgments, and play in the strategy method of the social dilemma and common interest game, I can compare, for each subject taking part in the experiment, the prediction of each of the theories under test with their actual play in the games.

The results in the experiment show that both moral rules and social preferences contribute to our understanding of attitudes to cooperation in cooperation problems. Maximin, inequality aversion, and blame avoidance achieve the greatest percentages of successful predictions and unique successful predictions (i.e., % of total data that is only predicted correctly by one theory) for joint play. When looking at each game individually,

---

experimental papers on public goods, one can see, as well, Andreoni (1988, 1990 and 1995); Croson (1996); Ferraro and Vossler (2010); Palfrey and Prisbey (1996 and 1997); and Anderson et al. (1998).

12 See Chaudhuri, 2011 for a review on conditionally cooperative behaviour, and also Fischbacher and Gächter (2010), and Thöni and Volk (2018). For the development of, and further literature on, contribution attitudes to public goods, one can refer to one can additionally refer to Weimann (1994); Bardsley (2000); Keser and Van Winden (2000); Frey and Meier (2004); Croson et al. (2005); Herrmann and Thöni (2009); Neugebauer et al. (2009); Smith (2011); Cartwright and Lovett (2014); Hartig et al. (2015); Gächter et al. (2017); Andreozzi et al. (2020); and Eichenseer and Moser (2000) among others.

it is noteworthy to mention that blame avoidance achieves a substantial percentage of unique successful prediction, amounting to around one fifth of the total subjects for the social dilemma, showing the contribution to our knowledge to attitudes to cooperation that moral rules make.

My paper makes a substantial contribution to the literature of prosociality in economics by presenting a substantially new model of moral decision-making, bringing it to the test against several canonical models within the discipline and demonstrating that we need non-trade-offs models if we want to improve our understanding of the individual heterogeneity that belies subjects' choices in cooperation problems. In this sense, the paper follows the ambition of Abeler et al. (2019), as it brings to the test several models to understand better which precise motivations are indeed causal drivers of behavioural choices in economic environments. It also builds on Miettinen et al. (2020) in performing a quantitative test of canonical models of social preferences, but unlike them the test is at the individual level. Whereas Miettinen et al. (2020) observed data and calculated the maximum predicted power of each theory, by calculating the parameter value that would maximise a theory's predictive power, my paper empirically measures parameters and confronts theories with the actual revealed parameter for each subject, thereby providing an accurate representation of a theory's predictive power at the individual level rather than a lowest upper boundary on such predictive power.

Furthermore, my paper presents a juxtaposition of related, but fundamentally different, cooperation problems and brings them to the test. Not only their difference is interesting because they are two different types of public goods, as discussed in Olson (1965, pp. 49-50) and Reuben and Reidl (2009), but their difference is interesting by what it reveals about current theories of prosocial preferences. Given the parameters I outline, neither inequality aversion nor reciprocity can predict conditional cooperation in the social dilemma and unconditional cooperation in the common interest game, and yet this is the second most popular joint behaviour in both games (25% of subjects in the experiment). This reveals a limitation, or paradox, of some of the key models of social preferences as currently formulated: their tendency to predict antisocial behaviour (i.e., not contributing in the common interest game and thereby destroying social welfare) as driven by a motivation that was supposed to be prosocial in nature (i.e., avoiding inequality, being reciprocal, and so on). Although this is by no means new, as people are known to burn money in

experiments and reject unequal offers in ultimatum games[13], it perhaps stresses that the underlying theories that were built to explain altruism imply that benevolent and malevolent behaviour are supposed to stem from the same underlying motivation, something at odds with the folk intuition of the concept of morality.

The remainder of the paper is structured as follows. The next section presents the experimental design. The second section presents the new moral theories under a generic framework, focusing the theoretical discussion on the two experimental games that are the core of the paper. The third section presents the social preferences I consider for the test, and presents their theoretical predictions, relegating all the proofs to an online appendix. Section four presents the results and section five concludes.

I. Experimental design

Each subject completed eight experimental tasks. Three of them – an *ultimatum game* (henceforth, UG), and a set of *modified dictator games* (henceforth, MDG) and *reciprocity games* (henceforth, RG) – were designed to elicit the parameters of a set of social preferences. Two experimental tasks involved two different versions of a two-person, one-shot, simultaneous move public goods game. I refer to these versions as a *social dilemma game* (henceforth, SDG) and a *common interest game* (henceforth, CIG), and to the tasks related to these versions as *P-experiments*. They elicited each subject's *cooperation attitudes (*as defined above – a subject's desired schedule of contributions for each contribution of the other group member*)*. Additionally, subjects had to complete what I refer to as two *M-experiments*, one related to the SDG and another related to the CIG. The M-experiments elicited each subject's moral judgments of each strategy combination of the SD and the CIG. Finally, subjects also completed a sociodemographic questionnaire.

For the remainder of the paper, I refer to all tasks related to the SDG (the relevant P- and M-experiments) as the *social dilemma tasks* and to all tasks related to the CIG (the relevant P- and M-experiments) as the *common interest game tasks*. I also refer to tasks involving UG, MDG and RG as *parameter-elicitation tasks*.

The order in which subjects performed the experimental tasks was as follows. Everyone answered the sociodemographic questionnaire at the end and the parameter-elicitation tasks after all the social dilemma and common interest game tasks had been completed. The

13 See, for instance, Zizzo and Oswald (2001); Zizzo (2003); Abbink and Sadrieh (2009); and Abbink and Herrmann (2011).

sequence in which all subjects answered the parameter elicitation tasks was kept the same for all: they completed the UG first, followed by the RG and, finally, the MDG. In contrast, I manipulated two aspects of the order of tasks: (i) whether the social dilemma tasks preceded or followed the common interest game tasks; and (ii) whether the M-experiments preceded or followed the P-experiments. This led to four different sequences in which tasks could be presented.

This manipulation led to a mixed design, where each subject had to complete all the tasks (*within-subjects* component) and subjects were randomly assigned to a treatment arm with a particular sequence (*between-subjects* component). The rationale for this design choice is threefold. First, moral suasion in public goods has been documented previously (see Dal Bó and Dal Bó, 2014). I wanted to control for any spillover effects between the M-experiments and the P-experiments to clearly identify any relation between moral judgments and cooperation attitudes beyond that captured by order effects in the presentation of the tasks. Second, I wanted to control for spillover effects between social dilemma tasks and common interest game tasks. Since they are very similar games, I want to be sure I can control for any anchoring effect that may arise by having been exposed to a similar game before when analysing cooperation attitudes. Third, by eliciting the P-experiments, M-experiments, and the parameters for each subject I was able to get each subject's observed cooperation attitudes of the SDG and the CIG and the predictions that each of the considered models make for those cooperation attitudes. The within-subjects element of the design allowed us, thus, to have all the necessary information to test the theories at the individual level.

To ensure that subjects understood the incentives of the SDG and the CIG, they had to answer some control questions after reading the instructions but before completing the M- and P-experiments. Only after they answered all control questions correctly they could proceed to complete those tasks. Subjects were allowed to participate in the experiment once only, and they received no feedback on their earnings and co-player's decisions until all tasks had been completed. This procedure is similar to that of Blanco et al (2011) and minimizes the chance of learning about the co-player's choices between tasks.

Only the two P-experiments and the parameter-elicitation games were incentivized. The incentivization scheme was as follows. Subjects played different games, each game had different roles and two games (RG and MDG) had different versions with different payoff allocations. I first gathered all the data, and, at the end of the experiment, I randomly assigned subjects to games, and all subjects assigned to a given game were randomly

matched into pairs. Once subjects were matched into pairs, I randomly assigned each pair member to one of the two possible roles for the game they had been allocated to. Lastly, for games with several versions (RG and MDG) one of the versions was randomly chosen to be relevant for each pair. Only the relevant actions arising from the randomization procedure implemented determined our subjects' final payoffs. Subjects were briefed about the procedure and knew how payoff were calculated. They also knew that all games, roles, and versions had the same probability of being chosen.

In the next subsections I provide a description of all tasks subjects had to complete. Given that one of the aims of the paper is to study the motivations behind cooperation attitudes in social dilemmas and common interest games, I start by giving a detailed account of the public goods game I used in the experiment prior to briefly presenting each experimental task.

A. The public goods game

The two cooperation problems I study – SDG and CIG – are based on the same decision situation: a linear, one-shot, simultaneous move, two-person public goods game. In the public goods game versions I implemented, each of the group members is endowed with 30 tokens and must decide how many to contribute to a group project (the public good). The material payoff function of a generic subject i is:

$$30 - c_i + m * \left(c_i + c_j\right) \tag{1}$$

Where c_i (c_j) refers to the token contributions of i (i's co-player) to the public good. A subject's feasible contribution levels are constrained to 0, 10, 20 or 30 tokens. For each token a subject does not contribute to the public good, that subject gets 1 token, and all the other group members get nothing. For each token a subject contributes to the public good, every member gets $m \in \{\underline{m}, \overline{m}\}$ tokens – that is, the benefits of the public good are non-excludable.

For the social dilemma I set $\underline{m}$ to 0.6, and for the common interest game I set $\overline{m}$ to 1.2[14]. Although the functional form of the payoff function is the same for both games, the qualitative incentive structure of the games is different because of the difference in the

14 More generally, for a SDG, then $\frac{1}{n} < \underline{m} < 1$ and for a CIG, then $\overline{m} > 1$

value of m. In the SDG, a subject gets more by not contributing a token to the public good (as $1 > 0.6$) whereas the total social payoff is maximized by contributing that token (as $1.2 > 1$). In contrast, in the CIG both the individual and total social payoff are maximized by contributing the token to the public good ($1.2 > 1$, and $2.4 > 1$ respectively).

B. Experimental tasks

The M-experiments

I use the survey method introduced by Cubitt et al (2011), and used in Gavassa-Pérez (2022), and adapt it to systematically elicit people's personal normative views of each strategy combination of the SD and the CIG.

Each M-experiment starts by presenting a given game to our subjects as an interaction between Person A and Person B. Then, I present each subject with several scenarios. Each scenario presents the contributions made by Person A and Person B to the public good and asks subjects to rate the morality of Person A on a scale ranging from -50 (extremely bad) to +50 (extremely good). A moral judgment of 0 is labelled as neutral. I run two M-experiments, one regarding the SDG and another one regarding the CIG. Each M-experiment consists of 16 scenarios, as I present to subjects one scenario for each strategy combination of Person A and Person B and the M-experiments are based on the SDG and CIG described earlier, where two players interact, each having only 4 feasible contribution levels (0, 10, 20, and 30). Figure 4.1.a provides a screenshot of how a set of scenarios of the SDG were presented to subjects, with Person B's contribution held constant but Person A's contribution varied across the scenarios in a given set. Recall that it is always Person A who is being judged.

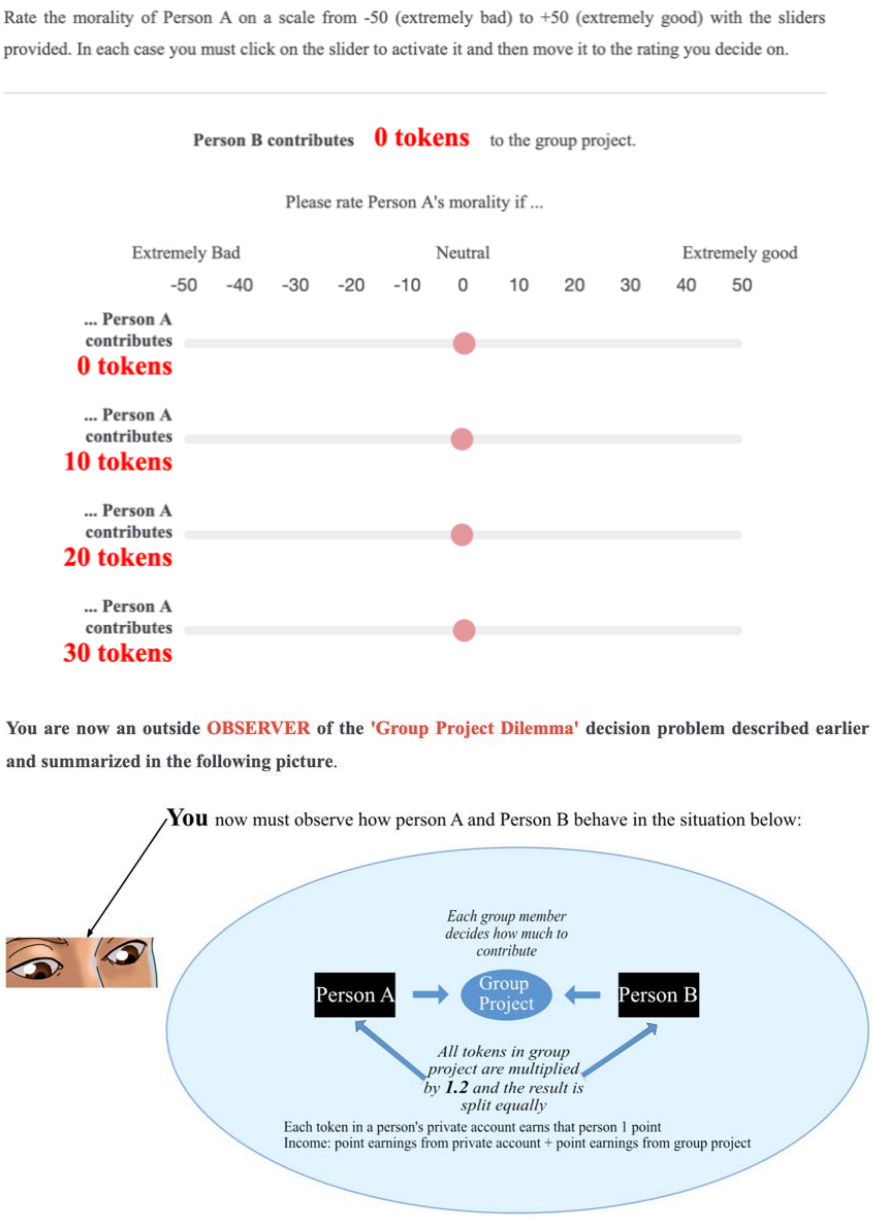


FIGURE 1. (A) TOP PANEL: SCREENSHOTS OF SCENARIOS; (B) BOTTOM PANEL: IMPLEMENTATION OF THE IMPARTIAL SPECTATOR FEATURE

Three characteristics of the M-experiment are worthy of discussion. First, I told subjects that they are neither Person A nor Person B and, rather, they are giving their moral views as an outside observer (an *impartial spectator*). This design choice aims to capture impartiality in moral judgments typical of the moral theories, among others, of Adam Smith

(see Konow, 2009, 2012 for discussion of the topic). A third party or a spectator has been used in the economics literature previously (see, for instance, Fehr and Fischbacher, 2004, for the use of third parties and, more recently, Konow, 2009, Smith and Wilson, 2014, Cappelen et al, 2019 and Almas et al, 2020). It is because the theories I develop are based on the moral judgments that one forms as an impartial spectator guiding one's own behaviour that I implemented this design choice. Figure 1.b summarizes how I introduced this feature to subjects in the M-experiment for the SDG.

Second, subjects were explicitly told to give their own moral views rather than society's normative opinions about the scenarios. I use this approach as the theories I present in this paper are based on an individual's moral code rather than the social moral conventions. This follows the tradition of an important part of moral philosophy (see Russell, 2009, ch.42, p.334-344 for a discussion).

Third, the M-experiments are not incentivized. I made this decision so that I did not confound subjects' true moral views with some hypothetical moral views that, if reported, would have maximized their payoff in the M-experiment given the incentive structure I would have chosen for it (see Cubitt at al, 2011 for discussion of this topic)[15]. This departs from what is currently done in the literature of social norms, where incentivized coordination games are used to elicit subjects' beliefs about the norms in their group (see Krupka and Weber, 2013 for one such approach). As good as this procedure sounds in the right context, it would not be appropriate for my design as I focus on subject's individual views rather than on their perceptions of the average social or moral conventions.

The P-experiments

I implement two tasks for both the SDG and the CIG: an *unconditional contribution* and a *contribution table task*. In the unconditional contribution task, a subject has to choose their contribution level without knowing what the other group member will choose. In the contribution table task, each subject must state their desired contribution per each feasible contribution of the other player. As each subject has four potential contribution levels (0, 10, 20, or 30), the contribution table task elicits four contributions per subject, one for each contribution level of the other player. It is this schedule of contributions from the contribution table task that I refer to as the subject's cooperation attitudes, and which

15 Additionally, there exists preliminary evidence suggesting that self-reported data contains important information aligning with subjects' attitudes in prosocial environments (see, for instance, Cappelen et al, 2011).

constitutes the dependent variable in our statistical analyses. Implementing the contribution table task in the SDG and CIG allows me to elicit such attitudes for both cooperation problems. The joint incentive-compatible elicitation of both tasks per each game constitutes the core methodology developed in Fischbacher, Gächter and Fehr (2001)[16], to which I refer to as the P-experiment.

To fix some notation, I define a free rider as a subject whose contributions are of the type $c_i^* = 0 \forall c_j$; a perfect conditional cooperator as a subject whose contributions are of the type $c_i^* = c_j \forall c_j$; and an unconditional cooperator as a subject whose contributions are of the type $c_i^* = 30 \forall c_j$.

Parameter-elicitation games

Subjects played three different games to elicit the parameters of a set of social preference theories. One such game was the two-person, *ultimatum game*. In the generic ultimatum game (Güth et al, 1982), two players – a proposer and a responder – interact. In the first stage, the proposer's decision is the number of monetary units out of a total pie P to offer to the responder. In the second stage, the responder's decision is whether to accept the offer. Letting o denote the offer, the respondent's acceptance of the offer implies the proposer gets $P - o$ and the responder gets o units as payoff. If, however, the responder rejects the offer, both players get nothing. In essence, the respondent gets to decide between two allocations – $(P - o, o)$ and $(0,0)$ – where the first (last) entry in each of the allocations defines the proposer's (respondent's) material payoff. I impose the following restrictions to the parameters of the game: (i) $o \in \mathbb{N}^*$; (ii) $o \in \left[0, \frac{P}{2}\right]$, and iii) $P = 14$. Each subject had to make their decision as a proposer and decide whether to accept the offer for each potential o that the proposer can send.

I also presented to subjects a set of *modified dictator games* based on the ones described in Blanco et al (2011). In these games, the dictator must choose between keeping the full pie (denoted P, as before) for himself or split another pie $(2x)$ into two equal shares. In essence, it is a decision between two allocations – $(P, 0)$ and (x, x) – where the first (last) entry in each of the allocations defines the dictator's (recipient's) payoff. Implementing

16 To make both tasks incentive compatible, Fischbacher, Gächter and Fehr (2001) impose, to each group member, a probability p for the unconditional contribution task to be payoff relevant and a probability $1 - p$ for the contribution table task to be payoff relevant. The probability p is known ex ante, but the realization of who will have the unconditional contribution and who will have the contribution table task as relevant is only realized after each subject has played both games.

several versions of this game in which I keep P fixed and vary x allows me to elicit each subject's willingness to pay to implement an equal split of income. I impose the following restrictions when setting all the implementations of the game: i) $x \in \mathbb{N}^*$; ii) x is an even number; iii) $P = 20$; and iv) $x \in [0,32]$. Restriction iv) is a significant one as it allows subjects to reveal negative willingness to pay for implementing an equal split of the total pie for any $x > P$[17].

The *reciprocity games* I implemented followed the ones presented in Bruhin et al (2019). Each reciprocity game is a two-stage, sequential game. In the first stage, the first mover decides whether to implement the allocation – $(5,95)$ – or pass on that allocation. In the second stage, the second mover only gets to choose if the first mover passes from implementing $(5,95)$, in which case he can select one of two alternative allocations – (x_4, x_2) and $(0,0)$, where I only vary the alternative allocation (x_4, x_2) between versions of the reciprocity game. Across all reciprocity games, I impose $x_2 < 95$ so that the first mover's decision to pass on implementing the allocation $(5,95)$ is unambiguously unkind for the second mover (as either of the alternative distributions gives him/her a lower payoff). Each subject had to state, per each version, whether to pass on $(5,95)$ when playing the role of the first mover and which of the alternative allocations to select as the second mover.

I follow Blanco et al (2011) in using a revealed-preference approach based on the games just described to calibrate the parameters of all the social preference models I consider. Using this approach for all the choices made, the revealed-preference approach reveals a range of values for the relevant parameter – provided that the subject's responses are compatible with any (i.e., if choices do not violate any axiom underlying preference relations). In online Appendix A. I present propositions showing the inequalities, for all the parameters of the social preference theories I consider, that are revealed given subjects' behaviour in the parameter elicitation games.

Sociodemographic questionnaire

Once subjects had finished all the previous tasks, I presented them several questions about their background characteristics. More specifically, I asked them about their gender, age, political identification (ranging from very left to very right), religiosity (ranging from

17 The direct implication is that, unlike Blanco et al (2011), I am explicitly able to detect subjects with spiteful preferences (i.e., subjects that derive pleasure for being ahead of others, and would need to be paid extra to accept an equal split of resources).

not religious at all to very religious), the community size (in number of inhabitants) where they lived most of their life, their field of study and presented them with the big five personality traits questionnaire.

C. Participants and procedures

Due to Covid restrictions, I ran the experiment online during May 2021 using Qualtrics. I recruited 318 students from the University of Nottingham using the ORSEE platform (Greiner, 2015). The number of participants was determined by a power calculation aiming to achieve 80% power given available estimates from previous data of Gavassa-Pérez (2022. See the pre-registration document for more details[18]). The average earning per subject being £7.88.

The average age of subjects was 21.4 years, 56.7% of subjects were female, another 51.9% identified as left and a further 42.5% self-reported as being religious.

II. The MRC framework: from Morality to Rules to Choices

A. Motivation

The MRC framework models individuals as having impartial moral judgments (i.e., personal normative evaluations) of all strategy combinations of the decision situation of interest. It assumes that subjects have a moral rule that receives those moral judgments as inputs and outputs a set of normative prescriptions for desired play at the relevant decision situation. In the case there is more than one suggested way to proceed, material selfishness acts as a tiebreaker to decide which, among all the morally suggested actions, to choose. My methodological framework owes intellectually to the contribution of Smith and Wilson (2019), which transformed Adam Smith's moral theory into an economically tractable framework, and to Francis Hutcheson's (2004) and David Hume's (1960 and 1983) works. The framework I present is novel as it mixes some concepts of the latter philosophers to the general theory of Smith and Wilson (2019) to be able, for the first time, to use a theory

18 The pre-registration document is available at: https://osf.io/z2yrh/

of personal moral judgments to make precise, testable predictions of behaviour at the individual level.

The MRC framework departs from the classical way to model social preferences, which revolve around self-centered individuals pursuing the maximization of their own broadened utility, normally containing their material payoff along with a specific social goal (e.g., inequality aversion, reciprocity, social efficiency, maximin, spite, and so on). My framework, instead, is based on subjects whose impartial judgments influence the way they ought to act. There are three main points of departure with the classical way in which social preferences are modelled, which I proceed to discuss below.

Self-centeredness has been proven an undesirable feature of some of those models (i.e., models of direct reciprocity), as evidenced, for instance, by people's tendency to punish as third parties (see, most notably, Fehr and Fischbacher, 2004): it is because subjects cannot consider a harmful action geared towards another person as unkind why reciprocity cannot predict to engage in costly punishment as a third party. By modelling the way in which morality drives behaviour as impartial, I allow people to base their behaviour on how a situation is perceived regardless of whether it involves them.

Additionally, my framework assumes that it is not the properties of the social interaction that directly feed one's choice deliberation. Rather, it is subjects' implicit judgments about those properties that are relevant for their decisions: I assume that it is not because some outcomes are unequal why subjects avoid inequality; but, rather, that only if those unequal outcomes are morally blameworthy subjects will avoid them. Modelling morality in this way I allow subjects to act differently in payoff-equivalent situations to the extent that those situations are evaluated differently from a moral perspective, thereby allowing framing effects even when beliefs are held constant.

Lastly, as far as the suggestion from the moral rule is a unique choice, my framework assumes that it is only a subject's morality that drives their behaviour, rather than being a mixture of a social goal and material selfishness. This feature of morality as the only input to the decision-making process is a unique feature of the MRC framework and can capture deontological attitudes that have been widely documented in the moral psychology literature in the form of taboo trade-offs (for work on protected values, see Baron and Spranca, 1997 and Baron, 2017. For work on taboo trade-offs, see Tetlock, 2003; Schoemaker and Tetlock, 2012; and Tetlock et al, 2017. For work on moral conviction, see Skitka et al, 2005, and Skitka, 2010. For work on morality as constraining the possible actions to be taken, see, more recently, Cushman, 2015; and Phillips and Cushman, 2017).

B. An illustrative example: the social dilemma game

To explain the intuition of my new framework, my starting point is the social dilemma game I presented in the previous section. Game theory typically assumes that a game is defined by the players, the set of strategies of each player and the utility functions of each player, that map each strategy combination into a given utility. Table 1.A below presents the normal form matrix of the social dilemma game under the assumption that both players' utility depends exclusively on the material payoffs of the game. The row player is person i and the column player is i's opponent, which I name '$-i$'. Both players have free riding as a strictly dominating strategy, so the benchmark of material selfishness predicts free riding regardless of the contribution of the other player.

TABLE 1. NORMAL FORM MATRIX OF THE SDG UNDER MATERIAL SELFISHNESS (A) AND INEQUALITY AVERSION (B)

Normal form matrix of the Social Dilemma Game …

a. … assuming material self interest

$i \setminus j$	$c_{-i} = 0$	$c_{-i} = 10$	$c_{-i} = 20$	$c_{-i} = 30$
$c_i = 0$	30,30	36,26	42,22	48,18
$c_i = 10$	26,36	32,32	38,28	44,24
$c_i = 20$	22,42	28,38	34,34	40,30
$c_i = 30$	18,48	24,44	30,40	36,36

b. … assuming Fehr-Schmidt preferences

$i \setminus j$	$c_{-i} = 0$	$c_{-i} = 10$	$c_{-i} = 20$	$c_{-i} = 30$
$c_i = 0$	30,30	$36 - \beta_i 10, 26 - \alpha_j 10$	$42 - \beta_i 20, 22 - \alpha_j 20$	$48 - \beta_i 30, 18 - \alpha_j 30$
$c_i = 10$	$26 - \alpha_i, 10, 36 - \beta_j 10$	32,32	$38 - \beta_i 10, 28 - \alpha_j 10$	$44 - \beta_i 20, 24 - \alpha_j 20$
$c_i = 20$	$22 - \alpha_i 20, 42 - \beta_j 20$	$28 - \alpha_i 10, 38 - \beta_j 10$	34,34	$40 - \beta_i 10, 30 - \alpha_j 10$
$c_i = 30$	$18 - \alpha_i 30, 48 - \beta_j 30$	$24 - \alpha_i 20, 44 - \beta_j 20$	$30 - \alpha_i 10, 40 - \beta_j 10$	36,36

Table 1.B transforms the material payoffs to account for inequality aversion as modelled by Fehr and Schmidt (1999). And, more generally, any social preference model changes this game theoretical benchmark by modifying the utility function of the players, thereby transforming the normal form matrix of material selfishness into a normal form matrix representing subjects' final utilities of every strategy combination of the game. In the case of inequality aversion, note that neither player will contribute more than the other player, as doing so decreases one's own material payoff and can only increase one's disadvantageous inequality, as $\alpha_i \geq 0$. However, inequality aversion deviates from the classical material selfishness assumption in the SDG whenever $\beta_i > 0.4$, as, in that case,

each player's best response is to contribute the same as the other player ($c_i^* = c_j \forall c_j \in C$). Hence, inequality aversion can predict free riding or perfect conditional cooperation in the social dilemma game; and, crucially, the prediction will depend on the strength of a subject's aversion towards advantageous inequality.

In contrast, the MRC framework elicits the moral judgments of every strategy combination in the social dilemma game, from an impartial perspective. Recall that moral judgments are on a scale from -50 (extremely bad) to +50 (extremely good). I represent such moral judgments in Table 2, setting the moral judgments to be the average moral judgments of the SDG in my experiments, rounded to the nearest integer, so that they are representative for the example.

TABLE 2. *i*'S MORAL JUDGMENTS OF PERSON A IN THE SDG

***i*'s Moral judgments of a Person A in the Social Dilemma Game …**				
$a \setminus b$	$c_b = 0$	$c_b = 10$	$c_b = 20$	$c_b = 30$
$c_a = 0$	-3	-15	-25	-34
$c_a = 10$	$+12$	$+7$	-8	-17
$c_a = 20$	$+24$	$+20$	$+12$	-2
$c_a = 30$	$+37$	$+32$	$+29$	$+20$

The first evident difference with classical models of social preferences is that the matrix in Table 2 does not regard subject i, which is the focus of our attention. Social preferences are self-centered as they assume that i's worry about inequality is born out of how inequality influences him\her. Rather, the MRC framework contemplates morality as arising from a disinterested stance. To do this, I assume subject i rates the morality of a generic player, Person A, when playing against another generic player, Person B, in the same decision situation that person i will play. That is, the moral judgments of Person A are done in an environment where the set of strategies of Person A and Person B, and the payoff consequences of all strategy combinations, are the same as in the game that i plays against j. The crucial assumption is that moral judgments are impartial. Thus, I assume that Person i will judge him/herself in the same way as he/she judges Person A. So, I can derive Table 3 from Table 2, where the moral judgments are kept the same, but now the players are i and j.

TABLE 3. I'S MORAL JUDGMENTS OF HIM/HERSELF IN THE SDG

***i*'s Moral judgments of *i* in the Social Dilemma Game …**				
$i \setminus j$	$c_{-i} = 0$	$c_{-i} = 10$	$c_{-i} = 20$	$c_{-i} = 30$
$c_i = 0$	-3	-15	-25	-34
$c_i = 10$	$+12$	$+7$	-8	-17
$c_i = 20$	$+24$	$+20$	$+12$	-2
$c_i = 30$	$+37$	$+32$	$+29$	$+20$

The MRC assumes that the way subjects come to act is by following a moral rule. Following Smith and Wilson (2019), I propose two such rules within the MRC framework: blame avoidance and praise seeking. Both moral rules use the relevant moral judgments as inputs to produce a given choice, or set of choices, that are morally suggested.

Blame avoidance states that a person ought to avoid doing blameworthy actions (i.e., actions with negative moral judgments). In this example, then, blame avoidance suggests that a subject ought to avoid doing $c_i = 0$ against $c_j = 0$, $c_i = 0$ against $c_j = 10$, $c_i \in \{0,10\}$ against $c_j = 20$ and $c_i \in \{0,10,20\}$ against $c_j = 30$, as all are strategy combinations for which, by impartiality, I assume i will judge him/her as being blameworthy (i.e., with negative moral judgments).

Praise seeking states that a person ought to choose the most praiseworthy actions (i.e., actions with the highest moral judgment). Hence, this rule suggests that a person ought to choose $c_i = 30$ against $c_j \in \{0,10,20,30\}$, as $c_i = 30$ has the highest rating attached to it for every value of c_j.

In practice, these rules constrain the set of possible strategies to choose against each strategy combination, and I can represent their output with a modified Table 1.A matrix in Tables 4.A and 4.B.

Table 4.A represents the normal form matrix of the SDG with all the cells representing strategy combinations not suggested by blame avoidance shaded in black. Similarly, Table 4.B represents the normal form matrix of the SDG with all the cells representing strategy combinations not suggested by praise seeking shaded in grey. Cells shaded in grey are cells that cannot be chosen by an individual if he/she decides to follow the relevant moral rule (blame avoidance for table 4.A; praise seeking for table 4.B).

TABLE 4. NORMAL FORM MATRIX OF THE SDG UNDER BLAME AVOIDANCE (A) AND PRAISE SEEKING (B)

Modified normal form matrix of the Social Dilemma Game …

a. … assuming blame avoidance

$i \setminus j$	$c_{-i} = 0$	$c_{-i} = 10$	$c_{-i} = 20$	$c_{-i} = 30$
$c_i = 0$				
$c_i = 10$	26,36	32,32		
$c_i = 20$	22,42	28,38	34,34	
$c_i = 30$	18,48	24,44	30,40	36,36

b. … assuming praise seeking

$i \setminus j$	$c_{-i} = 0$	$c_{-i} = 10$	$c_{-i} = 20$	$c_{-i} = 30$
$c_i = 0$				
$c_i = 10$				
$c_i = 20$				
$c_i = 30$	18,48	24,44	30,40	36,36

Whenever a moral rule suggests a single strategy to be taken, as is the case with praise seeking in Table 4.B, then no further work is needed, and the relevant moral rule would predict those strategies to be chosen. In the case of praise seeking, it would imply that person i ought to be an unconditional co-operator (i.e., $c_i = 30 \forall c_j$). If, however, more than one strategy is plausible given the output of a moral rule, as is the case with blame avoidance, then I use material selfishness as a tiebreaker to make a point prediction about i's play in the game. In the case of Table 4.A, person i ought to choose $c_i = 10$ against $c_j \in \{0,10\}$; choose $c_i = 20$ against $c_j = 20$; and choose $c_i = 30$ against $c_j = 30$.

C. A formal presentation of the MRC framework: praise seeking and blame avoidance

Preliminaries

Let $I \coloneqq \{i, j\}$ be the set of players and $G \coloneqq \{SDG, CIG\}$, with g as its typical element, be the set of games; where SDG is the social dilemma and CIG is the common interest game. Let $M \coloneqq \{-50, \ldots, 0, \ldots, +50\}$ be the judgment space. Let $C \coloneqq \{0,10,20,30\}$ be the individual contributions space in the public goods games presented earlier. It is the set of strategies (feasible contributions) for each hypothetical agent (Person A and Person B), for person i and for '$-i$'. Let the Cartesian product $C \times C$, with typical ordered pair $\langle c_a, c_b \rangle$, be the set of all strategy combinations in the public goods games I study; where

c_a and c_b denote, respectively, the contributions of Person A (the judged person) and Person B (the non-judged person) to the public good. As $C \times C$ is also the set of strategy combinations of i and j, I shall also use, without any loss of generality, the notation $\langle c_i, c_j \rangle$ to refer to a typical ordered pair of $C \times C$. Let $m: C \times C \times G \times I \rightarrow M$ be the moral judgments of an impartial spectator of the set of the strategy combinations of the relevant games. Let, m depend on the strategy combination, the game being played and the identity of the person standing on the role of an impartial spectator: $m(\langle c_a, c_b \rangle, g, i)$. The variable i captures a subject i's biases that he/she cannot get rid of when entering the impartial spectator stance. Also, let $m_i: C \times C \times G \rightarrow M$ denote a function from the set of strategy combinations of relevant games to the judgment space. m_i is the function of the moral judgments that subject i holds about him/herself in game g for a strategy combination $\langle c_i, c_j \rangle$. It follows that $m(\langle c_a, c_b \rangle, g, i) \in M$ represents the moral judgment that subject i has, as an impartial spectator, of Person A given the strategy combination $\langle c_a, c_b \rangle$ in game g. Similarly, $m_i\big(\langle c_i, c_j \rangle, g\big) \in M$ represents the moral judgment that subject i has of him/herself given the strategy combination $\langle c_i, c_j \rangle$ in game g. Lastly, denote $R: G \times C \rightarrow C$ as a function whose domain is all the combinations of strategies of a given player and relevant games and whose range is the set of strategies, common to all relevant games. Then, a function R can be understood as the rule that selects a given strategy against each strategy of the other player in each game. The functions of the type R, thus, represent the predicted schedules of contributions against each potential contribution of the other player in each game.

Assumptions of the MRC framework and predictions

The MRC framework is based on five main assumptions: (1) impartiality in judgments; (2) subjectivity in judgments; (3) moral rules as constraints in choices; (4) material selfishness as a tiebreaker; and (5) rule-following. Below I present the five assumptions together with the predictions that blame avoidance and praise seeking make about cooperation attitudes in the SDG and CIG. I discuss how each assumption is applied to both praise seeking and blame avoidance when the assumption is specific to each theory.

Assumption 1. Impartiality in judgments.

Assumption 1 says that subjects form moral judgments from the stance of an impartial spectator. Put differently, subjects evaluate the moral judgment of a given scenario imagining how they would judge such scenario if they would not take part in it. Then, they ascribe to themselves the same moral rating as they ascribed to the relevant player from the impartial spectator stance. This assumption is most prominent in Adam Smith's Theory of Moral Sentiments, but it also appears in other theories of moral philosophy, such as Hume's *judicious spectator* in the Treatise of Human Nature (1739, Book III, Part I, Sect. II., pp. 472), or Rawls' *veil of ignorance* within the original position proposed in A Theory of Justice (1999, pp.118-123). Given my notation, this assumption can be written as:

(2) $$If\ \langle c_a, c_b\rangle = \langle c_i, c_j\rangle, then\ m(\langle c_a, c_b\rangle, g, i) \equiv m_i\big(\langle c_i, c_j\rangle, g\big)$$

I use this assumption in the experiments to infer each subject's moral judgments of him/herself in all strategy combinations of the SDG and CIG from the moral judgments that they ascribed to Person A in the M-experiments (see discussion in subsection 2.2, where I go from Table 4.2 to Table 4.3). It is this assumption that makes the MRC framework to depart from the self-centeredness of classical models of social preferences, as I move the focus from analysing a social situation with respect to oneself (as social preferences do) to analysing the moral aspect of a scenario without subjects making any reference to themselves.

Assumption 2. Subjectivity in judgments.

Assumption 2 says that, although subjects put themselves in an impartial position when making judgments, nothing ensures that they can abstract from all their own characteristics when making judgments. Given my notation, I can capture Assumption 2 as:

(3) $$\frac{\partial m(\langle c_a, c_b\rangle, g, i)}{\partial i} \gtreqless 0$$

As far as the bias that two subjects bring to the impartial spectator stance is different, then their moral judgment of the same scenario will be different. In my notation,

(4) $$If\ m(\langle c_a, c_b\rangle, g, i) \neq m(\langle c_a, c_b\rangle, g, j), then\ m_i\big(\langle c_i, c_j\rangle, g\big) \neq m_{-i}\big(\langle c_i, c_j\rangle, g\big)$$

Thus, Assumption 2's contribution to the MRC framework is to state that $m(\langle c_a, c_b\rangle, g, i) = m(\langle c_a, c_b\rangle, g, \mathrm{j})$ is not necessarily true. This feature of moral judgments is especially present in the works of Francis Hutcheson (2002) and David Hume (1739), who held a view that paralleled aesthetics with ethics. They conceived that people may have different perceptions of *good* and *wrong*, just as they had different perceptions of *beauty* and *deformity*[19]. It is this assumption that makes the MRC framework different from Smith and Wilson (2019)'s Humanomics framework, as I consider subject's moral judgments – and, hence, potentially their predicted choices – to differ.

Assumption 3. Moral Rules as constraints to choices.

This assumption says that moral rules constrain the set of strategies to a subset of strategies that a subject can make in a game. I initially include two moral rules within the MRC framework: praise seeking and blame avoidance.

The rule of praise seeking states that subjects ought to seek choosing strategy combinations that they perceive as most praiseworthy as impartial spectators. Given my previous notation, I can define the subset of strategies suggested by the rule of praise seeking for individual i against strategy c_{-i} in game g as:

$$(5) \quad B_{i,c_{\mathrm{j}},g} \coloneqq \left\{c_i \in C \,|\, (\forall c_i' \in C)\left(m_i\left(\langle c_i, c_{\mathrm{j}}\rangle, g\right) \geq m_i\left(\langle c_i', c_{\mathrm{j}}\rangle, g\right)\right)\right\}$$

Where B stands for 'best' and $B_{i,c_j,g} \subseteq C$ is the subset of strategies that praise seeking suggests an agent i to take against c_j in game g. They are those strategies with the highest moral judgment for the relevant c_j and g.

The *rule of blame avoidance* states that subjects ought to avoid choosing strategy combinations that they perceive as blameworthy as impartial spectators. Given my previous notation, I can define the subset of strategies suggested by the rule of blame avoidance for individual i against strategy c_j in game g as:

19 Read, for instance, Hume's (1998, pp.134) sentence: "There are certain terms in language which import blame, and others praise; and all men who use the same tongue must agree in their application of them. … But when critics come to particulars, this seeming unanimity vanishes; and it is found, that they had affixed a very different meaning to their expressions. … Those who found morality on sentiment, more than on reason, are inclined to comprehend ethics under the former observation, and to maintain, that, in all questions which regard to conduct and manners, the difference among men is really greater than at first sight it appears"

(6) $$U_{i,c_j,g} \coloneqq \{c_i \in C | m_i(\langle c_i, c_j \rangle, g) \geq 0\},$$

where U stands for 'un-condemned' and $U_{i,c_j,g} \subseteq C$ is the subset of strategies that blame avoidance suggests an agent i to take against c_j in game g. These are those strategies that have a non-negative moral judgment for the relevant c_j and g.

Assumption 4. Material selfishness as a tiebreaker.

This assumption says that with respect to their material payoffs subjects are strictly monotonous, locally insatiable individuals. Hence, in the absence of moral considerations they prefer to choose strategies that yield them a higher material payoff. In other words:

(7) $$c_i \succ c_i' \; iff \; \pi_i(\langle c_i, c_j \rangle, g) > \pi_i(\langle c_i', c_j \rangle, g)$$

Where $\pi_i(\langle c_i, c_j \rangle, g)$ refers to the material payoff that subject i gets given the strategy combination $\langle c_i, c_j \rangle$ in game g.

Whenever the sets $B_{i,c_j,g}$ or $U_{i,c_j,g}$ contain a single element, that is, $\left|B_{i,c_j,g}\right| = 1$ or $\left|U_{i,c_j,g}\right| = 1$ respectively, then subject i's choices against c_j in game g will be uniquely determined by praise seeking or blame avoidance, respectively. However, whenever more than one strategy lies within $B_{i,c_j,g}$ or $U_{i,c_j,g}$, then I apply material selfishness as a tiebreaker to decide the predicted strategy for subject i against c_j in game g. More formally,

(8) $$B'_{i,c_j,g} \coloneqq \left\{ c_i \in B_{i,c_j,g} | \left(\forall c_i' \in B_{i,c_j,g}\right), c_i \succ c_i' \right\}$$

(9) $$U'_{i,c_j,g} \coloneqq \left\{ c_i \in U_{i,c_j,g} | \left(\forall c_i' \in U_{i,c_j,g}\right), c_i \succ c_i' \right\}$$

Where the set $B'_{i,c_j,g} \subseteq B_{i,c_j,g}$ (resp. $U'_{i,c_j,g} \subseteq U_{i,c_j,g}$) represents a set with a single element, the element being the strategy that yields the highest payoff within all the strategies allowed by praise seeking (resp. blame avoidance) against c_j in game g.

Assumption 5. *Rule-following*.

This assumption says that subjects make their choices according to their moral rules and, when the tiebreaker is needed, refined by material self-interest. The rules for praise seeking and blame avoidance for subject i when playing against c_{-i} in game g can be defined as:

$$PS_{i,c_j,g} \coloneqq \begin{cases} B_{i,c_j,g} \; if \; \left|B_{i,c_j,g}\right| = 1 \\ B'_{i,c_j,g} \; if \; \left|B_{i,c_j,g}\right| > 1 \end{cases} \tag{10}$$

$$BA_{i,c_j,g} \coloneqq \begin{cases} B'_{i,c_j,g} \; if \; U_{i,c_j,g} = \emptyset \\ U_{i,c_j,g} \; if \; \left|U_{i,c_j,g}\right| = 1 \\ U'_{i,c_j,g} \; if \; \left|U_{i,c_j,g}\right| > 1 \end{cases} \tag{11}$$

Where $PS_{i,c_j,g}$ (resp. $BA_{i,c_j,g}$) is a set with a single element, that element representing subject i's predicted strategy against c_j in game g if i follows the rule of praise seeking (resp. blame avoidance). Whenever $B_{i,c_j,g}$ and $U_{i,c_j,g}$ contain a single element, then the values of the functions $PS_{i,c_j,g}$ and $BA_{i,c_j,g}$ are uniquely based on the moral constraints imposed on choice by blame avoidance and praise seeking. Whenever $B_{i,c_j,g}$ and $U_{i,c_j,g}$ contain more than one element, then the values of the functions $PS_{i,c_j,g}$ and $BA_{i,c_j,g}$ are based on the most selfish actions out of the ones allowed by praise seeking and blame avoidance. Whenever all moral judgments are negative, then $U_{i,c_j,g}$ will be empty, and hence a subject's suggestion will be to do that action which minimizes blameworthiness when performed. In the case where all feasible strategies are blameworthy, that suggestion will be the same as the one of praise seeking, as the strategy with the highest moral judgment will be the least negative one.

I can, then, use sets of the type $PS_{i,c_j,g}$ and $BA_{i,c_j,g}$ to define praise seeking and blame avoidance's predicted vector of contributions for subject i in game g as:

$$\overrightarrow{\boldsymbol{PS}}_{i,g} = \left(PS_{i,0,g}, PS_{i,10,g}, PS_{i,20,g}, PS_{i,30,g}\right) \tag{12}$$

$$\overrightarrow{\boldsymbol{BA}}_{i,g} = \left(BA_{i,0,g}, BA_{i,10,g}, BA_{i,20,g}, BA_{i,30,g}\right) \tag{13}$$

It is these two vectors per each subject i and per each game g that form the predictions of praise seeking and blame avoidance regarding cooperation attitudes in the SDG and CIG.

III. Social preferences and cooperation attitudes

In this section I present all the utility functions of the social preference models I consider as candidates to predict cooperation attitudes in both the social dilemma and the common interest game, alongside their theoretical predictions, relegating the proofs to the online appendix A. The utility functions we focus on are:

$$U_i^{FS}(\pi_i, \pi_j) \coloneqq \pi_i - \alpha_i * Max\{\pi_j - \pi_i, 0\} - \beta_i * Max\{\pi_i - \pi_j, 0\} \quad (14)$$

$$U_i^{DK}(\pi_i, \pi_j) \coloneqq \pi_i\left(a_i(h), b_{i,j}(h)\right) + Y_{i,j} * \kappa_{i,j}\left(a_i(h), b_{i,j}(h)\right) * \lambda_{i,j,i}\left(b_{i,j}(h), c_{i,j,i}(h)\right) \quad (15)$$

$$U_i^{S}(\pi_i, \pi_j) \coloneqq \pi_i - \beta_i \times Max\{\pi_i - \pi_j, 0\} \quad (16)$$

$$U_i^{SE}(\pi_i, \pi_j) \coloneqq (1 - p_i)\pi_i + p_i \times (\pi_i + \pi_j) \quad (17)$$

$$U_i^{MM}(\pi_i, \pi_j) \coloneqq (1 - q_i)\pi_i + q_i \times Min\{\pi_i, \pi_j\} \quad (18)$$

Where α_i and β_i in eq. (14) capture the weight a subject puts to his own experienced pain of disadvantageous and advantageous inequality, respectively; $Y_{i,j}$ in eq. (15) captures the weight a subject puts to his own experienced pain of reciprocal concerns of matching kindness and perceived kindness; β_i in eq. (16) captures the weight a subject puts to his own experienced pleasure of having more payoff than others; and p_i and q_i are the weight a subject puts to the experienced pleasure a subject derives from the total payoff in a society and the payoff of the person worse-off in an interaction, respectively.

Equation (14) captures inequality aversion motives as presented in Fehr and Schmidt (1999), where $\alpha_i \geq \beta_i \geq 0$ and $\beta_i < 1$. Equation (15) captures reciprocal concerns as modelled by Dufwenberg and Kirchsteiger (2004), where $Y_{i,j} \geq 0$. Equation (16) captures spiteful motivations, where eq. (16) is constructed by changing the range of parameter values of the Fehr and Schmidt (1999) model, such that $\alpha_i = 0$ and $\beta_i \leq 0$.

TABLE 5. THEORETICAL PREDICTIONS OF SOCIAL PREFERENCES

Motivation	Predictions in …	
	… Social Dilemma Game ($m = 0.6$)	… Common Interest Game ($m = 1.2$)
Homo Economicus	1. $c_i = 0 \; \forall \; c_j \in C$	1. $c_i = 30 \; \forall \; c_j \in C$
Inequality Aversion *Equation (14)*	1. $If \; \beta_i < 0.4,$ $then \; c_i = 0 \; \forall \; c_j \in C$ 2. $If \; \beta_i > 0.4,$ $then \; c_i = c_j \; \forall \; c_j \in C$	1. $If \; \alpha_i < 0.2,$ $then \; c_i = 30 \; \forall \; c_j \in C$ 2. $If \; \alpha_i > 0.2,$ $then \; c_i = c_j \; \forall \; c_j \in C$
Maximin *Equation (18)*	1. $If \; q_i < 0.4,$ $then \; c_i = 0 \; \forall \; c_j \in C$ 2. $If \; q_i > 0.4,$ $then \; c_i = c_j \; \forall \; c_j \in C$	1. $c_i = 30 \; \forall \; c_j \in C$
Social Efficiency *Equation (17)*	1. $If \; p_i < 2/3,$ $then \; c_i = 0 \; \forall \; c_j \in C$ 2. $If \; \beta_i > 2/3,$ $then \; c_i = 30 \; \forall \; c_j \in C$	1. $c_i = 30 \; \forall \; c_j \in C$
Reciprocity *Equation (15)*	1. $If \; Y_{i,j} < 0.4/5.4,$ $then \; c_i = 0 \; \forall \; c_j \in C$ 2. $If \;\; 0.4/5.4 < Y_{i,j} < 0.4/1.8,$ $then \; c_i = \begin{cases} 0 \; \forall \; c_j \in \{0,10,20\} \\ 30 \;\; iff \; c_j \in \{30\} \end{cases}$ 3. $If \; Y_{i,j} > 0.4/1.8,$ $then \; c_i = \begin{cases} 0 \; \forall \; c_j \in \{0,10\} \\ 30 \;\; \forall \; c_j \in \{20,30\} \end{cases}$	1. $If \; Y_{i,j} < 0.2/43.4,$ $then \; c_i = 30 \; \forall \; c_j \in C$ 2. $If \;\; 0.2/43.4 < Y_{i,j} < 0.2/28.8,$ $then \; c_i = \begin{cases} 0 \; iff \; c_j \in \{0\} \\ 30 \;\; \forall \; c_j \in \{10,20,30\} \end{cases}$ 3. $If \;\; 0.2/28.8 < Y_{i,j} < 0.2/14.4,$ $then \; c_i = \begin{cases} 0 \; \forall \; c_j \in \{0,10\} \\ 30 \;\; \forall \; c_j \in \{20,30\} \end{cases}$ 4. $If \; Y_{i,j} > 0.2/14.4,$ $then \; c_i = \begin{cases} 0 \; \forall \; c_j \in \{0,10,20\} \\ 30 \;\; iff \; c_j \in \{30\} \end{cases}$
Spitefulness *Equation (16)*	1. $c_i = 0 \; \forall \; c_j \in C$	1. $If \; \beta_i > -0.2,$ $then \; c_i = 30 \; \forall \; c_j \in C$ 2. $If -0.3 < \beta_i < -0.2,$ $then \; c_i = \begin{cases} 0 \; iff \; c_j \in \{30\} \\ 30 \;\; \forall \; c_j \in \{0,10,20\} \end{cases}$ 3. $If -0.6 < \beta_i < -0.3,$ $then \; c_i = \begin{cases} 0 \; \forall \; c_j \in \{20,30\} \\ 30 \;\; \forall \; c_j \in \{0,10\} \end{cases}$ 4. $If \; \beta_{i,j} < -0.6,$ $then \; c_i = \begin{cases} 0 \; \forall \; c_j \in \{10,20,30\} \\ 30 \;\; iff \; c_j \in \{0\} \end{cases}$

Notes: c_i refers to the predicted contribution of subject i, c_j refers to the contribution of i's co-player, and $C \coloneqq \{0,10,20,30\}$ refers to the strategy space of both i and j.

Equations (17) and (18) disaggregate the two social motivations of the model presented in Charness and Rabin (2002) into two utility functions, where $p_i, q_i \in [0,1]$.

Table 5 below presents the theoretical predictions of the five social preference models we focus on this paper, together with the prediction of material selfishness (labelled as *Homo Economicus*). Crucially, the table manifests that most models can capture heterogeneity in cooperation attitudes in both games through different parameter values for different subjects.

As we measure all the parameters of equations (14) to (18) at the individual level, we get a clear-cut prediction for each of the theories per each subject, and then compare those individual-level predictions with individual-level behaviour at the P-experiments of the social dilemma and the common interest game.

IV. Results

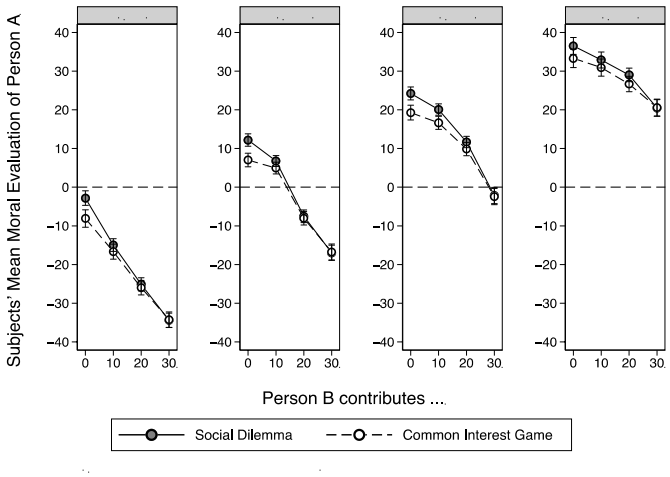
A. Descriptive statistics

Before the main statistical analysis, we report the main descriptive statistics of the key experimental variables. Figure 2 plots information about moral judgments and contribution attitudes in social dilemma and common interest games in three panels.

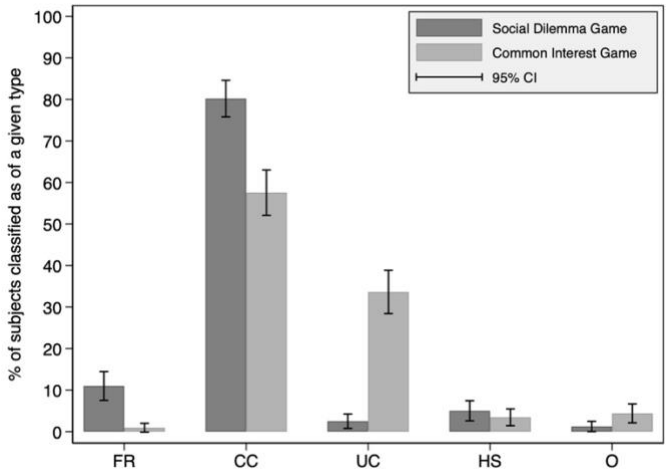
Panel A plots the average moral judgments (with 95% confidence intervals) of all scenarios of M-experiments. I display them in 4 sub-panels, each containing all average moral judgments corresponding to scenarios based on the same contribution level of Person A (the judged Person. For short, c_a). I then arrange (within each subpanel) the average moral judgments as an increasing function of the contribution of the non-judged player (Person B) and connect the average moral judgments with a line for each game, hence providing two lines per subpanel. As a benchmark, I plot – in each panel – a black, dotted horizontal line at a moral rating of 0.

Four features of Panel A are especially striking. First, average moral judgments different from 0 imply that subjects perceive the SDG and the CIG as situations of moral significance. Second, MEF's are increasing in c_a (the contribution of the judged person), suggesting an *increasing approbation of Person A* the more he/she contributes to the public good.

A.



B.



C.

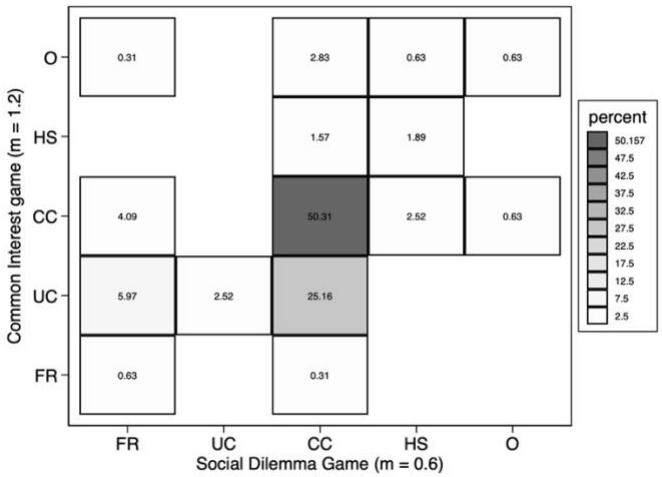


FIG. 2. A. AVERAGE MORAL JUDGMENTS OF ALL CONTRIBUTIONS OF PERSON A. **B.** % OF SUBJECTS WITH A GIVEN CONTRIBUTION TYPE FOR EACH GAME. **C.** HEATMAP OF THE JOINT DISTRIBUTION OF CONTRIBUTION TYPES IN SOCIAL DILEMMA AND COMMON INTEREST GAMES.

Third, MEF's are decreasing in c_b (the contribution of the non-judged person), suggesting an *increasing condemnation of Person A* the higher the contribution of relevant others to the public good. And fourth – and perhaps more strikingly –, MEF's of social dilemmas and common interest games are remarkably similar.

Applying the framework presented in Section 2 to the data displayed in Panel A, we find that, for our sample, the average predictions of praise seeking are unconditional cooperation in both games, whilst the average predictions of blame avoidance are a form of conditional cooperation in the social dilemma and unconditional cooperation in the common interest game. This serves to highlight a key point of interest of the theories, especially of blame avoidance: even when moral judgments are the same for the two games, the theories have the potential to make different predictions for both games.

Panel B reports the distribution of subjects' cooperation attitudes in the SDG and CIG, as measured by Thöni and Volk's (2018) classification of free riders (FR), conditional co-operators (CC), unconditional co-operators (UC), hump-shaded (HS) and others (O), along with 95% confidence intervals for the proportion of each type. As Panel B clearly shows, I find a significantly lower number of free riders and conditional co-operators and a significantly higher number of unconditional co-operators in the common interest game relative to the social dilemma.

Panel C reports the joint distribution of contribution types with a heat map, where a darker colour represents a higher proportion of subjects with a given joint distribution of types. The heatmap reveals that the two main joint distributions present in our data are those of conditional cooperation in both games (around 50% of subjects) and conditional cooperation in the social dilemma and unconditional cooperation in the common interest game (around 25% of subjects). It is also important to note that selfishness cannot account for the high prevalence of unconditional cooperation in the common interest game (more than 30% of subjects), as only 5% of the total data represents subjects who are free riders in the social dilemma and unconditional co-operators in the common interest games.

Additionally, we report in Table 6 the descriptive statistics of the social preference parameters, elicited at the individual level with the parameter-elicitation games. On average, the parameters of inequality aversion, social efficiency, and maximin are bigger than those of reciprocity and/or spite. In terms of behaviour, the average parameter values of inequality aversion and reciprocity imply free riding in the social dilemma and

a form of conditional cooperation in the common interest game. The average spite parameter is very close to 0 (-0.02), which implies the same predictions as material selfishness: free riding in social dilemmas and unconditional cooperation in common interest games. The average values of the social efficiency ($p_i = 0.47$) and the maximin ($q_i = 0.38$) parameters imply free riding in the social dilemma and unconditional cooperation in the common interest game[20].

TABLE 6. DESCRIPTIVE STATISTICS OF ELICITED PARAMETERS OF OTHER-REGARDING PREFERENCES

		Theoretical Range	Empirical Range	25th Percentile	Mean	75th Percentile	St. Dev.
Inequality Aversion							
	α_i	$[0,\infty)$	[0,3]	0.52	1.21	2.13	0.95
	β_i	[0,1)	[0,1]	0.05	0.38	0.55	0.35
Spite							
	β_i	$(-\infty,0]$	$[-0.61,0]$	0.00	-0.02	0.00	0.09
Reciprocity							
	$Y_{i,j}$	$[0,\infty)$	[0,3.9]	0.00	0.16	0.02	0.75
Social Efficiency							
	p_i	[0,1]	[0,1]	0.06	0.47	1.00	0.43
Maximin							
	q_i	[0,1]	[0,1]	0.05	0.38	0.55	0.35

Notes: The values of this table are computed without using the data of subjects with multiple switches in either of the three games. I maintain all remaining subjects regardless of whether they violate a condition of the theory (e.g., $\beta_i > \alpha_i$). For people with no switches, I impute values at the extreme of the theoretical range. For inequality aversion (resp. spite), I impute $\beta_i = 0$ whenever I observe $\beta_i < 0$ (resp. $\beta_i > 0$).

B. Do social preferences and moral rules influence cooperation attitudes?

For the analysis of the experimental data, Table 7 below presents the percentage of successful predictions of the different models under test i) for both games separately; and ii) for joint behaviour in both games. Given the nature of the strategy method data of the P-experiments, that gives me four behavioural data points per each game (i.e., a chosen contribution for each of the possible contributions of the other co-player), I consider a theory to be successful in predicting a subject's choices when such a theory's predictions exactly coincide with the four choices of the subject. I also report a random benchmark that

20 A mean outside the interquartile range in the spite and reciprocity parameters deserves some discussion. In the case of spite, most subjects in the modified dictator games elicited a positive β_i. I imputed a value of 0 for the spite parameter to all subjects who revealed a positive β_i, hence the skewed distribution. The distribution of the reciprocity parameter was also skewed as subjects showed extreme reciprocal attitudes in the reciprocity games. The high number of subjects with low revealed reciprocity dragged the mean downwards.

uses a uniform distribution to predict each of the choices per individual[21]. Additionally, I compare the percentage of unique successful predictions of the different theories. That is, the percentage of individual's choices that can only be explained by one of the theories at the individual level when all the other theories make wrong predictions. When calculating these two ratios, I impose that a violation of an assumption of a given theory for an individual renders null any predictive power that the theory has.

TABLE 7. PERCENTAGE OF TOTAL AND UNIQUE SUCCESSFUL PREDICTIONS OF THE THEORIES UNDER TEST

	Social Dilemma		Common Interest		Joint analysis	
	% of … successful predictions		% of … successful predictions		% of … successful predictions	
	Total…	Unique…	Total …	Unique …	Total …	Unique …
Randomness						
Uniform distrib.	0.39%		0.39%		0.002%	
Moral Rules						
Blame Avoidance	**26.42%**	**17.30%**	11.32%	3.46%	5.97%	4.09%
Praise Seeking	2.52%	0.94%	26.10%	0.00%	0.94%	0.31%
Homo Econ.						
Selfishness	11.01%	0.00%	**33.02%**	0.00%	5.97%	0.31%
Social Prefs.						
Inequality Aversion	17.92%	0.00%	18.87%	0.00%	7.23%	5.97%
Reciprocity	10.06%	0.00%	24.84%	0.00%	4.40%	0.00%
Social Efficiency	10.69%	0.63%	31.45%	0.00%	6.29%	0.94%
Maximin	**26.42%**	4.09%	31.45%	0.00%	**12.89%**	**6.92%**
Spite	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Notes: I print in bold the highest percentage in each column.

Looking at the data for the social dilemma shows that blame avoidance and maximin have the highest % of successful predictions (26.42% of all subjects). However, blame avoidance has a substantially higher % of unique successful predictions than maximin (17.30% vs 4.09%). This can be because maximin and inequality aversion's predictions for the social dilemma are perfect conditional cooperation if concerns for the social motive are high enough, and there is a high correlation between the parameters of advantageous inequality and maximin, hence generating a high correlation in prediction between both theories. One should, then, interpret the high value of the percentage of unique predictions of blame avoidance not so much in comparison with other theories but on its own: blame avoidance makes distinct predictions than the ones of social preference models and systematically gets that distinct prediction right 17.30% of the times. Taken at face value,

21 Given that each person can choose any of four strategies per choice, the probability of the random benchmark getting the four choices exactly right per each game is $0.25^4 = 0.0039$, and the probability of getting the eight choices of both games exactly right is $0.25^8 = 0.000015$.

it indicates that to understand cooperation attitudes in social dilemmas when social preference models fail at the individual level we need to make recourse of the novel models of moral rules presented herein, highlighting their contribution to the explanation of attitudes to cooperation in social dilemmas. Other social preferences, like reciprocity and social efficiency, have a lower predictive power, and almost negligible % of unique successful predictions, implying they are secondary motivations to explain attitudes to cooperation in social dilemmas. Finally, spite and praise seeking have a very small predictive power, them being the worse theories in terms of explaining attitudes to cooperation in social dilemmas.

When looking at the data of common interest games, the story changes. Blame avoidance has a lower predictive power than all models of social preferences and selfishness, highlighting the importance of the latter in explaining behaviour in common interest games. However, blame avoidance, again, ranks first in the percentage of unique successful predictions, pointing out that we need to make recourse to blame avoidance if we want to understand 3.46% of the attitudes to cooperation in common interest games.

Finally, we look at the joint play in both games. Here, we can see three theories with high % of successful and unique successful predictions: maximin, inequality aversion, and blame avoidance. Other models of social preferences have high levels of successful predictions (e.g., reciprocity, social efficiency), but with negligible levels of unique successful predictions, telling us that most of the data they predict could also be accounted for some of the other models in the test.

Overall, the results suggest that both moral rules and social preferences are good predictors of attitudes to cooperation in social dilemmas and common interest games, and that their predictive success is, in most cases, substantially higher than what would be expected from randomly generated data. Moral rules, esp. blame avoidance, performs very well in social dilemmas, even when brought to the test not only against the void but against several canonical models of social preferences in the literature. This highlights the fact that we should consider people's normative judgments are a driver of choice. Social preferences, esp. maximin and inequality aversion, are also good explanations of attitudes to cooperation, and other behavioural motivations (e.g., reciprocity, social efficiency, praise seeking) are more game-dependent, as they only have some substantial level of predictive power in common interest games.

V. Concluding remarks and implications of the results

In this paper I have analysed the likelihood of a set of social preference and moral rule theories in explaining cooperation attitudes of two cooperation problems: social dilemmas and common interest games. To achieve this, I have measured (i) cooperation attitudes with P-experiments; (ii) the parameters of several social preference models with parameter-elicitation games; and (iii) the moral judgments of each strategy combination of social dilemmas and common interest games with M-experiments. The latter two measurements have been used to generate predictions of five social preference models (inequality aversion, reciprocity, social efficiency, maximin, and spite) and two novel moral rules (blame avoidance and praise seeking). Using these theoretical predictions, I have tested the seven theories against each other, and against the benchmark of material selfishness, to determine the likelihood of each of them as explanations of cooperation attitudes in cooperation problems.

The results can be best summarized as follows. I can group the theories into three groups according to their consistency with the data. The first group, formed by maximin, inequality aversion, and blame avoidance, receives high levels of successful and unique successful prediction of joint play. The second group, including social efficiency, reciprocity, praise seeking, and material selfishness, get high levels of substantial prediction in common interest games. The third group, formed by spite, contains the theories that get little successful predictions in either game. In conclusion, cooperation attitudes of cooperation problems are likely to be driven by several heterogeneous motivations.

These results have two major implications that I proceed to discuss in detail now. One implication is that no unique motivation – at least from the ones considered in this experiment – can explain people's cooperation attitudes. A second implication, more important in my view, is that the data does not support a single modelling strategy for representing subjects' social behaviour. The main modelling strategy in the social preferences literature relies on self-centered agents that derive pleasure from both material selfishness and a social goal. In contrast, the two moral rules within the model presented herein – praise seeking and blame avoidance – are models that represent an individual's motivation for the social as coming from a disinterested, impartial perspective. It is the individual who constrains his choice space so as to tone it down to what, from an impartial perspective, he considers admissible given an underlying moral rule that forms his own

motivational basis for action. This study demonstrates that both the classical, self-centered models and my new, impartial, moral judgment-based models are compatible with observed behaviour when the other models aren't, revealing two different paths to shaping cooperation attitudes in social dilemmas and common interest games.

Appendix A: Regression analysis

I support my analysis by presenting random effects estimates of the data from the SDG and CIG separately. The equation I estimate uses the observed cooperation attitudes as the dependent variable and the predicted cooperation attitudes of most of the theories presented in the two previous sections as dependent variables. Additionally, in the estimated equation I also use the contribution of the other co-player (c_j) to control for the potential effect of other relevant social preference theories in cooperation attitudes, and two dummies to control for the order effects of moral judgments (whether moral judgments preceded or followed the P-experiment) and games (whether the SDG tasks preceded or followed the CIG tasks)[22]. Columns '*Estimates*' in Table A1 report the regression estimates.

Four patterns can be inferred from the data. First, only inequality aversion and blame avoidance are statistically significant in both games, which I take as a signal of them being more universal motives of cooperation attitudes. Second, spite and social efficiency were statistically significant in the only regression in which they were included (CIG and SDG respectively). I take this as initial evidence of their role in explaining cooperation attitudes. Third, reciprocity is statistically significant only in common interest games, suggesting that it is a specific motivation of cooperation attitudes in the CIG. Four, only blame avoidance has a similar

[22] My rationale is as follows. First, note that guilt aversion's prediction, in social dilemmas, of cooperation attitudes for subjects with a high concern for avoiding guilt is contributing according to their second-order belief (see Dufwenberg et al, 2011). Assuming a high probability of playing against a conditional co-operator, it is reasonable to believe that the other co-player's contribution is increasing in that co-player's expectation about their contribution. Second, a central concept in social norms is empirical expectations (see Bicchieri, 2005 and 2017), which have been shown to be important drivers of behaviour even when they conflict with normative expectations (see Bicchieri and Xiao, 2009). As the contribution of others (c_j) represents a subject's empirical expectations of his/her co-player behaviour I see a reasonable conjecture the statement that social norms' predictions will vary in proportion to c_j.

coefficient in both regressions, suggesting its effect is more stable than that of the other social preferences. More specifically, inequality aversion and reciprocity have a significantly greater coefficient in CIG, suggesting they play a greater role in explaining cooperation attitudes in CIG.

Table A1. *Regression estimates and decomposition of explained variance*

Dependent variable: Cooperation attitudes (elicited in the contribution table task of the P-experiments)				
	Social dilemma game		*Common interest game*	
Independent variables	Estimates	Decomposition of R^2	Estimates	Decomposition of R^2
Constant	1.591		8.676***	
	(1.34)		(1.769)	
c_{-i}	0.585***	**52.58%**	0.213***	**20.77%**
	(0.031)		(0.05)	
Predictions				
Moral Rules				
Blame Avoidance	0.094***	**24.26%**	0.094***	**19.61%**
	(0.033)		(0.036)	
Praise Seeking	-0.011	0.40%	-0.021	0.33%
	(0.042)		(0.045)	
Social Preferences				
Inequality Aversion	0.11***	**17.46%**	0.225***	**34.34%**
	(0.034)		(0.044)	
Reciprocity	-0.006	0.71%	0.105***	**15.32%**
	(0.051)		(0.026)	
Social Efficiency	0.075**	4.07%		
	(0.03)			
Spite			0.06**	9.02%
			(0.026)	
Controls				
Social Dilemmas first	-0.596	0.24%	0.039	0.02%
	(0.743)		(0.931)	
Moral Judgments first	-0.873	0.28%	0.863	0.58%
	(0.741)		(0.929)	

*Notes: * p<0.1 ** p<0.05 *** p<0.01.* Percentages higher than 10% are printed in bold.

Additionally, I report the estimates of the decomposition of explained variance in columns '*Decomposition of* R^2' of Table A1. I decompose the explained overall variance in shares by applying the hierarchical partitioning method proposed in Chevan and Sutherland (1991) to the data. The share of all the independent variables adds up to one, each share representing the relative importance of each of the independent variables in explaining cooperation attitudes.

It is remarkable to see that more than 50% of the explained variation in cooperation attitudes of the SDG is captured by the c_j control variable. As explained above, I used it as a proxy for the effect that other theories not included in the test had in cooperation attitudes. More specifically, I conjectured guilt aversion and social norms to be the two main theories that could be represented within the control. The high relative importance in both games, together with statistical significance in both games, suggests that these alternative theories play an important role in cooperation attitudes.

Going back to the theories I do test, blame avoidance appears as the clear winner in the SDG: its relative importance is higher than the aggregate relative importance of all the remaining theories (24.26% vs. 22.64%). Only inequality aversion gets close, capturing 17.46% of the explained variation of cooperation attitudes in social dilemmas. Out of the remaining variables, only social efficiency has a non-negligible relative importance, although its role in explaining cooperation attitudes is substantially lower than inequality aversion and blame avoidance.

Data from the CIG reveal a different picture, revealing inequality aversion as of greater relative importance than blame avoidance (34.34% vs 19.61%). Again, both theories share the first and second place of relative importance in the CIG. Reciprocity (15.32%) and spite (9.02%), this time, have a substantial degree of relative importance, strengthening my previous claim suggesting their game-specific role in explaining cooperation attitudes of cooperation problems.

Overall, most of the qualitative findings in the results section are in line with what is reported in the regression analysis. First, out of the theories tested blame avoidance and inequality aversion are explanations of cooperation attitudes in both cooperation problems. Second, reciprocity, social efficiency, and spite are game-specific explanations of cooperation attitudes and play a minor role relative to that of blame avoidance and inequality aversion. Third, moral rules play a greater role than social preferences in explaining cooperation attitudes of social dilemmas, and social preferences play a greater role than moral rules in explaining cooperation attitudes of common interest games.

Online Appendix A: Proofs

https://egavassaperez.org/files/ernesto-mgp/files/gavassaperezernesto_jmp_proofs.pdf

Online Appendix B: Instructions

https://egavassaperez.org/files/ernesto-mgp/files/gavassaperezernesto_jmp_instructions.pdf

REFERENCES

Abbink, Klaus and Benedikt Herrmann. 2011. "The Moral Costs of Nastiness." *Economic Inquiry*, 49(2), 631-33.

Abbink, Klaus and Abdolkarim Sadrieh. 2009. "The Pleasure of Being Nasty." *Economics Letters*, 105(3), 306-08.

Abeler, Johannes; Daniele Nosenzo and Collin Raymond. 2019. "Preferences for Truth-Telling." *Econometrica*, 87(4), 1115-53.

Adolphs, R.; L. Sears and J. Piven. 2001. "Abnormal Processing of Social Information from Faces in Autism." *Journal of Cognitive Neuroscience*, 13(2), 232-40.

Adolphs, Ralph. 2010. "What Does the Amygdala Contribute to Social Cognition?" *Annals of the New York Academy of Sciences*, 1191(1), 42-61.

Adolphs, Ralph; Simon Baron-Cohen and Daniel Tranel. 2002. "Impaired Recognition of Social Emotions Following Amygdala Damage." *Journal of Cognitive Neuroscience*, 14(8), 1264-74.

Adolphs, Ralph and Michael Spezio. 2006. "Role of the Amygdala in Processing Visual Social Stimuli," S. Anders, G. Ende, M. Junghofer, J. Kissler and D. Wildgruber, *Progress in Brain Research.* Elsevier, 363-78.

Adolphs, Ralph; Daniel Tranel and Antonio R. Damasio. 1998. "The Human Amygdala in Social Judgment." *Nature*, 393(6684), 470-74.

Alger, Ingela and Jörgen W. Weibull. 2013. "Homo Moralis—Preference Evolution under Incomplete Information and Assortative Matching." *Econometrica*, 81(6), 2269-302.

Algoe, Sara B. and Jonathan Haidt. 2009. "Witnessing Excellence in Action: The 'Other-Praising' Emotions of Elevation, Gratitude, and Admiration." *The Journal of Positive Psychology*, 4(2), 105-27.

Allison, Truett; Aina Puce and Gregory McCarthy. 2000. "Social Perception from Visual Cues: Role of the Sts Region." *Trends in Cognitive Sciences*, 4(7), 267-78.

Allman, John; Atiya Hakeem and Karli Watson. 2002. "Book Review: Two Phylogenetic Specializations in the Human Brain." *The Neuroscientist*, 8(4), 335-46.

Almås, Ingvild; Alexander W. Cappelen and Bertil Tungodden. 2020. "Cutthroat Capitalism Versus Cuddly Socialism: Are Americans More Meritocratic and Efficiency-Seeking Than Scandinavians?" *Journal of Political Economy*, 128(5), 1753-88.

Anderson, Adam K. and Elizabeth A. Phelps. 2001. "Lesions of the Human Amygdala Impair Enhanced Perception of Emotionally Salient Events." *Nature*, 411(6835), 305-09.

Anderson, Simon P.; Jacob K. Goeree and Charles A. Holt. 1998. "A Theoretical Analysis of Altruism and Decision Error in Public Goods Games." *Journal of Public Economics*, 70(2), 297-323.

Anderson, Steven W.; Antoine Bechara; Hanna Damasio; Daniel Tranel and Antonio R. Damasio. 1999. "Impairment of Social and Moral Behavior Related to Early Damage in Human Prefrontal Cortex." *Nature Neuroscience*, 2(11), 1032-37.

Andreoni, James. 1995. "Cooperation in Public-Goods Experiments: Kindness or Confusion?" *The American Economic Review*, 85(4), 891-904.

____. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *The Economic Journal*, 100(401), 464-77.

____. 1988. "Why Free Ride?: Strategies and Learning in Public Goods Experiments." *Journal of Public Economics*, 37(3), 291-304.

Andreozzi, Luciano; Matteo Ploner and Ali Seyhun Saral. 2020. "The Stability of Conditional Cooperation: Beliefs Alone Cannot Explain the Decline of Cooperation in Social Dilemmas." *Scientific Reports*, 10(1), 13610.

Angrilli, Alessandro; Alessandra Mauri; Daniela Palomba; Herta Flor; Niels Birbaumer; Giuseppe Sartori and Francesco di Paola. 1996. "Startle Reflex and Emotion Modulation Impairment after a Right Amygdala Lesion." *Brain*, 119(6), 1991-2004.

Bardsley, Nicholas. 2000. "Control without Deception: Individual Behaviour in Free-Riding Experiments Revisited." *Experimental Economics*, 3(3), 215-40.

Barger, Nicole; Kari L. Hanson; Kate Teffer; Natalie M. Schenker-Ahmed and Katerina Semendeferi. 2014. "Evidence for Evolutionary Specialization in Human Limbic Structures." *Frontiers in Human Neuroscience*, 8.

Baron, Jonathan. 2017. "Protected Values and Other Types of Values." *Analyse & Kritik*, 39(1), 85-100.

Baron, Jonathan and Mark Spranca. 1997. "Protected Values." *Organizational Behavior and Human Decision Processes*, 70(1), 1-16.

Bas-Hoogendam, Janna Marie; Henk van Steenbergen; Tanja Kreuk; Nic J. A. van der Wee and P. Michiel Westenberg. 2017. "How Embarrassing! The Behavioral and Neural Correlates of Processing Social Norm Violations." *PLOS ONE*, 12(4), e0176326.

Basile, Benjamin M.; Jamie L. Schafroth; Chloe L. Karaskiewicz; Steve W. C. Chang and Elisabeth A. Murray. 2020. "The Anterior Cingulate Cortex Is Necessary for Forming Prosocial Preferences from Vicarious Reinforcement in Monkeys." *PLOS Biology*, 18(6), e3000677.

Bastin, Coralie; Ben J. Harrison; Christopher G. Davey; Jorge Moll and Sarah Whittle. 2016. "Feelings of Shame, Embarrassment and Guilt and Their Neural Correlates: A Systematic Review." *Neuroscience & Biobehavioral Reviews*, 71, 455-71.

Battigalli, Pierpaolo and Martin Dufwenberg. 2007. "Guilt in Games." *American Economic Review*, 97(2), 170-76.

Becchara, Antoine; Hanna Damasio and Antonio R. Damasio. 2003. "Role of the Amygdala in Decision-Making." *Annals of the New York Academy of Sciences*, 985(1), 356-69.

Bechara, Antoine and Antonio R. Damasio. 2005. "The Somatic Marker Hypothesis: A Neural Theory of Economic Decision." *Games and Economic Behavior*, 52(2), 336-72.

Bechara, Antoine; Hanna Damasio and Antonio R. Damasio. 2000. "Emotion, Decision Making and the Orbitofrontal Cortex." *Cerebral Cortex*, 10(3), 295-307.

Bechara , Antoine; Hanna Damasio ; Antonio R. Damasio and Gregory P. Lee 1999. "Different Contributions of the Human Amygdala and Ventromedial Prefrontal Cortex to Decision-Making." *The Journal of Neuroscience*, 19(13), 5473-81.

Beer, Jennifer S.; Erin A. Heerey; Dacher Keltner; Donatella Scabini and Robert T. Knight. 2003. "The Regulatory Function of Self-Conscious Emotion: Insights from Patients with Orbitofrontal Damage." *Journal of Personality and Social Psychology*, 85(4), 594-604.

Bénabou, Roland and Jean Tirole. 2011. "Identity, Morals, and Taboos: Beliefs as Assets *." *The Quarterly Journal of Economics*, 126(2), 805-55.

____. 2006. "Incentives and Prosocial Behavior." *American Economic Review*, 96(5), 1652-78.

Berthoz, S.; J. L. Armony; R. J. R. Blair and R. J. Dolan. 2002. "An Fmri Study of Intentional and Unintentional (Embarrassing) Violations of Social Norms." *Brain*, 125(8), 1696-708.

Bicchieri, Cristina. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

____. 2017. *Norms in the Wild. How to Diagnose, Measure, and Change Social Norms*. Corby: Oxford University Press.

Bicchieri, Cristina and Erte Xiao. 2009. "Do the Right Thing: But Only If Others Do So." *Journal of Behavioral Decision Making*, 22(2), 191-208.

Bickart, Kevin C.; Bradford C. Dickerson and Lisa Feldman Barrett. 2014. "The Amygdala as a Hub in Brain Networks That Support Social Life." *Neuropsychologia*, 63, 235-48.

Blair, James; A. A. Marsh; E. Finger; K. S. Blair and J. Luo. 2006. "Neuro-Cognitive Systems Involved in Morality." *Philosophical Explorations*, 9(1), 13-27.

Blair, R. J. R. 2007. "The Amygdala and Ventromedial Prefrontal Cortex in Morality and Psychopathy." *Trends in Cognitive Sciences*, 11(9), 387-92.

Blanco, Mariana; Dirk Engelmann and Hans Theo Normann. 2011. "A within-Subject Analysis of Other-Regarding Preferences." *Games and Economic Behavior*, 72(2), 321-38.

Bohm, Peter. 1972. "Estimating Demand for Public Goods: An Experiment." *European Economic Review*, 3(2), 111-30.

Bolton, Gary E. and Axel Ockenfels. 2000. "Erc: A Theory of Equity, Reciprocity, and Competition." *American Economic Review*, 90(1), 166-93.

Brandts, Jordi; Tatsuyoshi Saijo and Arthur Schram. 2004. "How Universal Is Behavior? A Four Country Comparison of Spite and Cooperation in Voluntary Contribution Mechanisms." *Public Choice*, 119(3), 381-424.

Brekke, Kjell Arne; Snorre Kverndokk and Karine Nyborg. 2003. "An Economic Model of Moral Motivation." *Journal of Public Economics*, 87(9), 1967-83.

Brosnan, Sarah. 2011. "A Hypothesis of the Co-Evolution of Cooperation and Responses to Inequity." *Frontiers in Neuroscience*, 5.

Brosnan, Sarah F. 2013. "Justice- and Fairness-Related Behaviors in Nonhuman Primates." *Proceedings of the National Academy of Sciences*, 110(Supplement 2), 10416-23.

____. 2006. "Nonhuman Species' Reactions to Inequity and Their Implications for Fairness." *Social Justice Research*, 19(2), 153-85.

Brosnan, Sarah F. and Frans B. M. de Waal. 2003. "Monkeys Reject Unequal Pay." *Nature*, 425(6955), 297-99.

Brosnan, Sarah F.; Catherine Talbot; Megan Ahlgren; Susan P. Lambeth and Steven J. Schapiro. 2010. "Mechanisms Underlying Responses to Inequitable Outcomes in Chimpanzees, Pan Troglodytes." *Animal Behaviour*, 79(6), 1229-37.

Bruhin, Adrian; Ernst Fehr and Daniel Schunk. 2018. "The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences." *Journal of the European Economic Association*, 17(4), 1025-69.

Brunton, Douglas; Rabia Hasan and Stuart Mestelman. 2001. "The 'Spite' Dilemma: Spite or No Spite, Is There a Dilemma?" *Economics Letters*, 71(3), 405-12.

Calvillo, Dustin P. and Jessica N. Burgeno. 2015. "Cognitive Reflection Predicts the Acceptance of Unfair Ultimatum Game Offers." *Judgment and Decision Making*, 10(4), 332-41.

Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Cappelen, Alexander W.; Gauri Gauri and Bertil Tungodden. 2019. "Cooperation Creates Special Moral Obligations." *CESifo Working Paper*, No. 7052.

Cappelen, Alexander W.; Astri Drange Hole; Erik Ø Sørensen and Bertil Tungodden. 2011. "The Importance of Moral Reflection and Self-Reported Data in a Dictator Game with Production." *Social Choice and Welfare*, 36(1), 105-20.

____. 2007. "The Pluralism of Fairness Ideals: An Experimental Approach." *American Economic Review*, 97(3), 818-27.

Cartwright, Edward J. and Denise Lovett. 2014. "Conditional Cooperation and the Marginal Per Capita Return in Public Good Games." *Games*, 5(4), 234-56.

Chang, Steve W. C.; Nicholas A. Fagan; Koji Toda; Amanda V. Utevsky; John M. Pearson and Michael L. Platt. 2015. "Neural Mechanisms of Social Decision-

Making in the Primate Amygdala." *Proceedings of the National Academy of Sciences*, 112(52), 16012-17.
Charness, Gary and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests*." *The Quarterly Journal of Economics*, 117(3), 817-69.
Chaudhuri, Ananish. 2011. "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature." *Experimental Economics*, 14(1), 47-83.
Chevan, Albert and Michael Sutherland. 1991. "Hierarchical Partitioning." *The American Statistician*, 45(2), 90-96.
Cooper, Davi J. and John H. Kagel. 2017. "Other-Regarding Preferences: A Selective Survey of Experimental Results," J. H. Kagel and A. E. Roth, *The Handbook of Experimental Economics, Volume 2.* Princeton, New Jersey: Princeton University Press, 217-89.
Cosmides, Leda and John Tooby. 1994. "Better Than Rational: Evolutionary Psychology and the Invisible Hand." *The American Economic Review*, 84(2), 327-32.
____. 1992. "Cognitive Adaptations for Social Exchange," J. H. Barkow, L. Cosmides and J. Tooby, *The Adapte Mind: Evolutionary Psychology and the Generation of Culture.* New York, United States of America: Oxford University Press, 163-228.
____. 2013. "Evolutionary Psychology: New Perspectives on Cognition and Motivation." *Annual Review of Psychology*, 64(1), 201-29.
Cosmides, Leda; John Tooby; Laurence Fiddick and Gregory A. Bryant. 2005. "Detecting Cheaters." *Trends in Cognitive Sciences*, 9(11), 505-06.
Cox, James C.; Daniel Friedman and Steven Gjerstad. 2007. "A Tractable Model of Reciprocity and Fairness." *Games and Economic Behavior*, 59(1), 17-45.
Croson, Rachel. 2007. "Theories of Commitment, Altruism and Reciprocity: Evidence from Linear Public Goods Games." *Economic Inquiry*, 45(2), 199-216.
Croson, Rachel; Enrique Fatas and Tibor Neugebauer. 2005. "Reciprocity, Matching and Conditional Cooperation in Two Public Goods Games." *Economics Letters*, 87(1), 95-101.
Croson, Rachel T. A. 1996. "Partners and Strangers Revisited." *Economics Letters*, 53(1), 25-32.
Cubitt, Robin P.; Michalis Drouvelis; Simon Gächter and Ruslan Kabalin. 2011. "Moral Judgments in Social Dilemmas: How Bad Is Free Riding?" *Journal of Public Economics*, 95(3), 253-64.
Cushman, Fiery. 2015. "From Moral Concern to Moral Constraint." *Current Opinion in Behavioral Sciences*, 3, 58-62.
Dal Bó, E. and P. Dal Bó. 2014. ""Do the Right Thing:" The Effects of Moral Suasion on Cooperation." *Journal of Public Economics*, 117, 28-38.
Damasio, Antonio R. 1995. *Descartes' Error: Emotion, Reason and the Human Brain*. New York, United States of America: Avon Books.
David Sander; Jordan Grafman and Tiziana Zalla. 2003. "The Human Amygdala: An Evolved System for Relevance Detection." *Reviews in the Neurosciences*, 14(4), 303-16.

Dawes, Robyn M; Jeanne McTavish and Harriet Shaklee. 1977. "Behavior, Communication, and Assumptions About Other People's Behavior in a Commons Dilemma Situation." *Journal of Personality and Social Psychology*, 35(1), 1.

de Waal, Frans B. M. 2009. *The Age of Empathy: Nature's Lessos for a Kinder Society*. New York, Unites States of America: Harmony Books.

____. 1997. "The Chimpanzee's Service Economy: Food for Grooming." *Evolution and Human Behavior*, 18(6), 375-86.

de Waal, Frans B. M. and Sarah F. Brosnan. 2006. "Simple and Complex Reciprocity in Primates," P. M. Kappeler and C. P. van Schaik, *Cooperation in Primates and Humans: Mechanisms and Evolution.* Berlin, Heidelberg: Springer Berlin Heidelberg, 85-105.

Decety, Jean and William John Ickes (Eds.). 2009. *The Social Neuroscience of Empathy*. Cambridge, Massachusetts: The MIT Press.

Decety, Jean and Philip L. Jackson. 2006. "A Social-Neuroscience Perspective on Empathy." *Current Directions in Psychological Science*, 15(2), 54-58.

Decety, Jean and Claus Lamm. 2006. "Human Empathy through the Lens of Social Neuroscience." *TheScientificWorldJOURNAL*, 6, 280363.

Declerck, Carolyn H.; Christophe Boone and Griet Emonds. 2013. "When Do People Cooperate? The Neuroeconomics of Prosocial Decision Making." *Brain and Cognition*, 81(1), 95-117.

Delton, Andrew W.; Leda Cosmides; Marvin Guemo; Theresa E. Robertson and John Tooby. 2012. "The Psychosemantics of Free Riding: Dissecting the Architecture of a Moral Concept." *Journal of Personality and Social Psychology*, 102(6), 1252-70.

Delton, Andrew W.; Max M. Krasnow; Leda Cosmides and John Tooby. 2011. "Evolution of Direct Reciprocity under Uncertainty Can Explain Human Generosity in One-Shot Encounters." *Proceedings of the National Academy of Sciences*, 108(32), 13335-40.

Duc, Corinne; Martin Hanselmann; Peter Boesiger and Carmen Tanner. 2013. "Sacred Values: Trade-Off Type Matters." *Journal of Neuroscience, Psychology, and Economics*, 6(4), 252-63.

Dufwenberg, Martin and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior*, 47(2), 268-98.

Dufwenberg, Martin; Simon Gächter and Heike Hennig-Schmidt. 2011. "The Framing of Games and the Psychology of Play." *Games and Economic Behavior*, 73(2), 459-78.

Dukes, Daniel; Kathryn Abrams; Ralph Adolphs; Mohammed E. Ahmed; Andrew Beatty; Kent C. Berridge; Susan Broomhall; Tobias Brosch; Joseph J. Campos; Zanna Clay, et al. 2021. "The Rise of Affectivism." *Nature Human Behaviour*, 5(7), 816-20.

Dunbar, R. I. M. 1991. "Functional Significance of Social Grooming in Primates." *Folia Primatologica*, 57(3), 121-31.

Ebstein, Richard; Simone Shamay-Tsoory and Soo Hong Chew (Eds.). 2011. *From DNA to Social Cognition*. New Jersey, United States of America: Wiley-Blackwell.

Eichenseer, Michael and Johannes Moser. 2020. "Conditional Cooperation: Type Stability across Games." *Economics Letters*, 188, 108941.
Eres, Robert; Winnifred R. Louis and Pascal Molenberghs. 2018. "Common and Distinct Neural Networks Involved in Fmri Studies Investigating Morality: An Ale Meta-Analysis." *Social Neuroscience*, 13(4), 384-98.
Eslinger, Paul J.; Silke Anders; Tommaso Ballarini; Sydney Boutros; Sören Krach; Annalina V. Mayer; Jorge Moll; Tamara L. Newton; Matthias L. Schroeter; Ricardo de Oliveira-Souza, et al. 2021. "The Neuroscience of Social Feelings: Mechanisms of Adaptive Social Functioning." *Neuroscience & Biobehavioral Reviews*, 128, 592-620.
Evans, Jonathan St. B. T. 2008. "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition." *Annual Review of Psychology*, 59(1), 255-78.
Falk, Armin and Urs Fischbacher. 2006. "A Theory of Reciprocity." *Games and Economic Behavior*, 54(2), 293-315.
Fehr, Ernst and Urs Fischbacher. 2004. "Third-Party Punishment and Social Norms." *Evolution and Human Behavior*, 25(2), 63-87.
Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation*." *The Quarterly Journal of Economics*, 114(3), 817-68.
FeldmanHall, O. and D. Mobbs. 2015. "A Neural Network for Moral Decision Making," A. W. Toga, *Brain Mapping*. Waltham: Academic Press, 205-10.
FeldmanHall, Oriel; Dean Mobbs and Tim Dalgleish. 2013. "Deconstructing the Brain's Moral Network: Dissociable Functionality between the Temporoparietal Junction and Ventro-Medial Prefrontal Cortex." *Social Cognitive and Affective Neuroscience*, 9(3), 297-306.
Feng, Chunliang; Yue-Jia Luo and Frank Krueger. 2015. "Neural Signatures of Fairness-Related Normative Decision Making in the Ultimatum Game: A Coordinate-Based Meta-Analysis." *Human Brain Mapping*, 36(2), 591-602.
Ferraro, Paul J and Christian A Vossler. 2010. "The Source and Significance of Confusion in Public Goods Experiments." *The B.E. Journal of Economic Analysis & Policy*, 10(1).
Finger, Elizabeth C.; Abigail A. Marsh; Niveen Kamel; Derek G. V. Mitchell and James R. Blair. 2006. "Caught in the Act: The Impact of Audience on the Neural Response to Morally and Socially Inappropriate Behavior." *NeuroImage*, 33(1), 414-21.
Fischbacher, Urs and Simon Gächter. 2010. "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments." *American Economic Review*, 100(1), 541-56.
Fischbacher, Urs; Simon Gächter and Ernst Fehr. 2001. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters*, 71(3), 397-404.
Forsythe, Robert; Joel L. Horowitz; N. E. Savin and Martin Sefton. 1994. "Fairness in Simple Bargaining Experiments." *Games and Economic Behavior*, 6(3), 347-69.
Fox, Glenn R.; Jonas Kaplan; Hanna Damasio and Antonio Damasio. 2015. "Neural Correlates of Gratitude." *Frontiers in Psychology*, 6.

Frey, Bruno S. and Stephan Meier. 2004. "Social Comparisons and Pro-Social Behavior: Testing "Conditional Cooperation" in a Field Experiment." *American Economic Review*, 94(5), 1717-22.
Gächter, Simon; Felix Kölle and Simone Quercia. 2017. "Reciprocity and the Tragedies of Maintaining and Providing the Commons." *Nature Human Behaviour*, 1(9), 650-56.
Gavassa-Pérez, Ernesto M. 2022. "From Morality to Rules to Choices: Introducing and Testing a New Theory on How Morals Influence Cooperation," Nottingham: University of Nottingham,
Gigerenzer, Gerd and Reinhard Selten. 2002. *Bounded Rationality: The Adaptive Toolbox*. MIT press.
Gilead, Michael; Maayan Katzir; Tal Eyal and Nira Liberman. 2016. "Neural Correlates of Processing "Self-Conscious" Vs. "Basic" Emotions." *Neuropsychologia*, 81, 207-18.
Gilovich, Thomas; Dale Griffin and Daniel Kahneman. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge university press.
Gleichgerrcht, Ezequiel and Liane Young. 2013. "Low Levels of Empathic Concern Predict Utilitarian Moral Judgment." *PLOS ONE*, 8(4), e60418.
Gonzalez-Liencres, Cristina; Simone G. Shamay-Tsoory and Martin Brüne. 2013. "Towards a Neuroscience of Empathy: Ontogeny, Phylogeny, Brain Mechanisms, Context and Psychopathology." *Neuroscience & Biobehavioral Reviews*, 37(8), 1537-48.
Gospic, Katarina; Erik Mohlin; Peter Fransson; Predrag Petrovic; Magnus Johannesson and Martin Ingvar. 2011. "Limbic Justice—Amygdala Involvement in Immediate Rejection in the Ultimatum Game." *PLOS Biology*, 9(5), e1001054.
Gospic, Katarina; Marcus Sundberg; Johanna Maeder; Peter Fransson; Predrag Petrovic; Gunnar Isacsson; Anders Karlström and Martin Ingvar. 2013. "Altruism Costs—the Cheap Signal from Amygdala." *Social Cognitive and Affective Neuroscience*, 9(9), 1325-32.
Green, Sophie; Matthew A. Lambon Ralph; Jorge Moll; Emmanuel A. Stamatakis; Jordan Grafman and Roland Zahn. 2010. "Selective Functional Integration between Anterior Temporal and Distinct Fronto-Mesolimbic Regions During Guilt and Indignation." *NeuroImage*, 52(4), 1720-26.
Greene, Joshua. 2003. "From Neural 'Is' to Moral 'Ought': What Are the Moral Implications of Neuroscientific Moral Psychology?" *Nature Reviews Neuroscience*, 4(10), 846-50.
Güth, Werner; Rolf Schmittberger and Bernd Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior & Organization*, 3(4), 367-88.
Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review*, 108(4), 814-34.
Hallsson, Bjørn G.; Hartwig R. Siebner and Oliver J. Hulme. 2018. "Fairness, Fast and Slow: A Review of Dual Process Models of Fairness." *Neuroscience & Biobehavioral Reviews*, 89, 49-60.

Harenski, Carla L.; Olga Antonenko; Matthew S. Shane and Kent A. Kiehl. 2010a. "A Functional Imaging Investigation of Moral Deliberation and Moral Intuition." *NeuroImage*, 49(3), 2707-16.

Harenski, Carla L.; Keith A. Harenski; Matthew S. Shane and Kent A. Kiehl. 2010b. "Aberrant Neural Processing of Moral Violations in Criminal Psychopaths." *Journal of Abnormal Psychology*, 119(4), 863-74.

Harsanyi, John C. 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy*, 63(4), 309-21.

Hartig, Björn; Bernd Irlenbusch and Felix Kölle. 2015. "Conditioning on What? Heterogeneous Contributions and Conditional Cooperation." *Journal of Behavioral and Experimental Economics*, 55, 48-64.

Haruno, Masahiko; Minoru Kimura and Christopher D. Frith. 2014. "Activity in the Nucleus Accumbens and Amygdala Underlies Individual Differences in Prosocial and Individualistic Economic Choices." *Journal of Cognitive Neuroscience*, 26(8), 1861-70.

Herrmann, Benedikt and Christian Thöni. 2009. "Measuring Conditional Cooperation: A Replication Study in Russia." *Experimental Economics*, 12(1), 87-92.

Hinterbuchinger, Barbara; Alexander Kaltenboeck; Josef Severin Baumgartner; Nilufar Mossaheb and Fabian Friedrich. 2018. "Do Patients with Different Psychiatric Disorders Show Altered Social Decision-Making? A Systematic Review of Ultimatum Game Experiments in Clinical Populations." *Cognitive Neuropsychiatry*, 23(3), 117-41.

Hopper, Lydia M.; Susan P. Lambeth; S.J. Schapiro; Bruce J. Bernacky and Sarah F. Brosnan. 2013. "The Ontogeny of Social Comparisons in Rhesus Macaques (Macaca Mulatta)." *J Primatol*, 2(1), 1-5.

House, Bailey R.; Joseph Henrich; Sarah F. Brosnan and Joan B. Silk. 2012. "The Ontogeny of Human Prosociality: Behavioral Experiments with Children Aged 3 to 8." *Evolution and Human Behavior*, 33(4), 291-308.

Hume, David. 1987. *An Enquiry Concerning the Principles of Morals*. Indianapolis: Hackett Pub. Co.

____. 2008. *Selected Essays*. Oxford: Oxford Univ. Press.

____. 1739. *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. Oxford, United Kingdom: Clarendon Press.

Hutcheson, Francis. 2002. *An Essay on the Nature and Conduct of the Passions and Affections, with Illustrations on the Moral Sense*. Indianapolis, United Statis of America: Liberty Fund.

____. 2004. *An Inquiry into the Original of Our Ideas of Beauty and Virtue in Two Treatises*. Indianapolis, United States of America: Liberty Fund.

Isaac, R. Mark; James M. Walker and Susan H. Thomas. 1984. "Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations." *Public Choice*, 43(2), 113-49.

Jackson, Philip L.; Andrew N. Meltzoff and Jean Decety. 2005. "How Do We Perceive the Pain of Others? A Window into the Neural Processes Involved in Empathy." *NeuroImage*, 24(3), 771-79.

Jankowski, Kathryn F. and Hidehiko Takahashi. 2014. "Cognitive Neuroscience of Social Emotions and Implications for Psychopathology: Examining Embarrassment, Guilt, Envy, and Schadenfreude." *Psychiatry and Clinical Neurosciences*, 68(5), 319-36.

Kahneman, Daniel and Amos Tversky. 1984. "Choices, Values, and Frames." *American Psychologist*, 39, 341-50.

Kant, Immanuel. 2012. *Groundwork of the Metaphysics of Morals*. Cambridge: Cambridge University Press.

Kappeler, Peter M. and Carel P. van Schaik. 2006. *Cooperation in Primates and Humans: Mechanisms and Evolution*. Berlin, Germany: Springer.

Keltner, Dacher and Jonathan Haidt. 2001. "Social Functions of Emotions," *Emotions: Currrent Issues and Future Directions.* New York, NY, US: Guilford Press, 192-213.

Keser, Claudia and Frans Van Winden. 2000. "Conditional Cooperation and Voluntary Contributions to Public Goods." *The Scandinavian Journal of Economics*, 102(1), 23-39.

Killen, Melanie and Judith Smetana. 2008. "Moral Judgment and Moral Neuroscience: Intersections, Definitions, and Issues." *Child Development Perspectives*, 2(1), 1-6.

Klimecki, Olga and Tania Singer. 2013. "Empathy from the Perspective of Social Neuroscience," J. Armony and P. Vuilleumer, *The Cambridge Handbook of Human Affective Neuroscience.* New York, United States of America: Cambridge University Press, 533-49.

Kluver, Jesse; Rebecca Frazier and Jonathan Haidt. 2014. "Behavioral Ethics for Homo Economicus, Homo Heuristicus, and Homo Duplex." *Organizational Behavior and Human Decision Processes*, 123(2), 150-58.

Konow, James. 2012. "Adam Smith and the Modern Science of Ethics." *Economics and Philosophy*, 28(3), 333-62.

____. 2009. "Is Fairness in the Eye of the Beholder? An Impartial Spectator Analysis of Justice." *Social Choice and Welfare*, 33(1), 101-27.

Krupka, Erin L. and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association*, 11(3), 495-524.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research," J. H. Kagel and A. E. Roth, *The Handbook of Experimental Economics.* Princeton, New Jersey: Princeton University Press, 111-94.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1(3), 593-622.

Lindquist, Kristen A.; Tor D. Wager; Hedy Kober; Eliza Bliss-Moreau and Lisa Feldman Barrett. 2012. "The Brain Basis of Emotion: A Meta-Analytic Review." *The Behavioral and brain sciences*, 35(3), 121-43.

LoBue, Vanessa; Tracy Nishida; Cynthia Chiong; Judy S. DeLoache and Jonathan Haidt. 2011. "When Getting Something Good Is Bad: Even Three-Year-Olds React to Inequality." *Social Development*, 20(1), 154-70.
Marwell, Gerald and Ruth E. Ames. 1979. "Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem." *American Journal of Sociology*, 84(6), 1335-60.
Masclet, David and David L. Dickinson. 2019. "Incorporating Conditional Morality into Economic Decisions." *IZA Discussion Papers*, No. 12872.
Massen, Jorg J. M.; Friederike Behrens; Jordan S. Martin; Martina Stocker and Sarah F. Brosnan. 2019. "A Comparative Approach to Affect and Cooperation." *Neuroscience & Biobehavioral Reviews*, 107, 370-87.
McAuliffe, Katherine; Jillian J. Jordan and Felix Warneken. 2015. "Costly Third-Party Punishment in Young Children." *Cognition*, 134, 1-10.
McKelvey, Richard D. and Thomas R. Palfrey. 1995. "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior*, 10(1), 6-38.
Mendez, Mario F. 2009. "The Neurobiology of Moral Behavior: Review and Neuropsychiatric Implications." *CNS Spectrums*, 14(11), 608-20.
Mendres, Kimberly A. and Frans B. M. de Waal. 2000. "Capuchins Do Cooperate: The Advantage of an Intuitive Task." *Animal Behaviour*, 60(4), 523-29.
Miettinen, Topi; Michael Kosfeld; Ernst Fehr and Jörgen Weibull. 2020. "Revealed Preferences in a Sequential Prisoners' Dilemma: A Horse-Race between Six Utility Functions." *Journal of Economic Behavior & Organization*, 173, 1-25.
Milinski, Manfred. 2013. "Chimps Play Fair in the Ultimatum Game." *Proceedings of the National Academy of Sciences*, 110(6), 1978-79.
Moll, J.; R. de Oliveira-Souza; I. E. Bramati and J. Grafman. 2002. "Functional Networks in Emotional Moral and Nonmoral Social Judgments." *NeuroImage*, 16(3 Pt 1), 696-703.
Moll, Jorge; Ricardo de Oliveira-Souza and Paul J. Eslinger. 2003. "Morals and the Human Brain: A Working Model." *NeuroReport*, 14(3), 299-305.
Moll, Jorge; Ricardo de Oliveira-Souza; Paul J. Eslinger; Ivanei E. Bramati; Janaína Mourão-Miranda; Pedro Angelo Andreiuolo and Luiz Pessoa. 2002. "The Neural Correlates of Moral Sensitivity: A Functional Magnetic Resonance Imaging Investigation of Basic and Moral Emotions." *The Journal of Neuroscience*, 22(7), 2730-36.
Moll, Jorge; Ricardo De Oliveira-Souza and Roland Zahn. 2008. "The Neural Basis of Moral Cognition." *Annals of the New York Academy of Sciences*, 1124(1), 161-80.
Moll, Jorge; Frank Krueger; Roland Zahn; Matteo Pardini; Ricardo de Oliveira-Souza and Jordan Grafman. 2006. "Human Fronto–Mesolimbic Networks Guide Decisions About Charitable Donation." *Proceedings of the National Academy of Sciences*, 103(42), 15623-28.
Moll, Jorge; Roland Zahn; Ricardo de Oliveira-Souza; Frank Krueger and Jordan Grafman. 2005. "The Neural Basis of Human Moral Cognition." *Nature Reviews Neuroscience*, 6(10), 799-809.
Moll, Jorge; Roland Zahn and Ricardo de Oliveira-Souza. 2016. "The Neural Underpinnings of Moral Values," T. Brosch and D. Sander, *Handbook of Value:*

Perspectives from Economics, Neuroscience, Philosophy, Psychology, and Sociology. New York, United States of America: Oxford University Press, 119-28.

Morey, Rajendra A.; Gregory McCarthy; Elizabeth S. Selgrade; Srishti Seth; Jessica D. Nasser and Kevin S. LaBar. 2012. "Neural Systems for Guilt from Actions Affecting Self Versus Others." *NeuroImage*, 60(1), 683-92.

Neugebauer, Tibor; Javier Perote; Ulrich Schmidt and Malte Loos. 2009. "Selfish-Biased Conditional Cooperation: On the Decline of Contributions in Repeated Public Goods Experiments." *Journal of Economic Psychology*, 30(1), 52-60.

Nichols, Shaun. 2002. "Norms with Feeling: Towards a Psychological Account of Moral Judgment." *Cognition*, 84(2), 221-36.

Öhman, Arne; Anders Flykt and Francisco Esteves. 2001. "Emotion Drives Attention: Detecting the Snake in the Grass." *Journal of Experimental Psychology: General*, 130(3), 466-78.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, Massachusetts: Harvard University Press.

Palfrey, Thomas R. and Jeffrey E. Prisbrey. 1996. "Altuism, Reputation and Noise in Linear Public Goods Experiments." *Journal of Public Economics*, 61(3), 409-27.

____. 1997. "Anomalous Behavior in Public Goods Experiments: How Much and Why?" *The American Economic Review*, 87(5), 829-46.

Phillips, Jonathan and Fiery Cushman. 2017. "Morality Constrains the Default Representation of What Is Possible." *Proceedings of the National Academy of Sciences*, 114(18), 4649-54.

Pizarro, David. 2000. "Nothing More Than Feelings? The Role of Emotions in Moral Judgment." *Journal for the Theory of Social Behaviour*, 30(4), 355-75.

Prelec, D. 1991. "Values and Principles: Some Limitations on Traditional Economic Analysis," A. Etizioni and P. Lawrence, *Socioeconomics: Towards a New Synthesis.* New York: M.E. Sharpe, 131-45.

Proctor, Darby; Sarah F. Brosnan and Frans B. M. de Waal. 2013a. "How Fairly Do Chimpanzees Play the Ultimatum Game?" *Communicative & Integrative Biology*, 6(3), e23819.

Proctor, Darby; Rebecca A. Williamson; Frans B. M. de Waal and Sarah F. Brosnan. 2013b. "Chimpanzees Play the Ultimatum Game." *Proceedings of the National Academy of Sciences*, 110(6), 2070-75.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *The American Economic Review*, 83(5), 1281-302.

Railton, Peter. 2014. "The Affective Dog and Its Rational Tale: Intuition and Attunement." *Ethics*, 124(4), 813-59.

Rand, David G.; Joshua D. Greene and Martin A. Nowak. 2012. "Spontaneous Giving and Calculated Greed." *Nature*, 489(7416), 427-30.

Rand, David G.; Alexander Peysakhovich; Gordon T. Kraft-Todd; George E. Newman; Owen Wurzbacher; Martin A. Nowak and Joshua D. Greene. 2014. "Social Heuristics Shape Intuitive Cooperation." *Nature Communications*, 5(1), 3677.

Rawls, John. 1999. *A Theory of Justice. Revised Edition*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.
Reuben, Ernesto and Arno Riedl. 2009. "Public Goods Provision and Sanctioning in Privileged Groups." *Journal of Conflict Resolution*, 53(1), 72-93.
Rilling, James K. and Alan G. Sanfey. 2011. "The Neuroscience of Social Decision-Making." *Annual Review of Psychology*, 62(1), 23-48.
Roemer, John E. 2010. "Kantian Equilibrium." *The Scandinavian Journal of Economics*, 112(1), 1-24.
Saarimäki, Heini; Lara Farzaneh Ejtehadian; Enrico Glerean; Iiro P Jääskeläinen; Patrik Vuilleumier; Mikko Sams and Lauri Nummenmaa. 2018. "Distributed Affective Space Represents Multiple Emotion Categories across the Human Brain." *Social Cognitive and Affective Neuroscience*, 13(5), 471-82.
Saijo, Tatsuyoshi and Hideki Nakamura. 1995. "The "Spite" Dilemma in Voluntary Contribution Mechanism Experiments." *Journal of Conflict Resolution*, 39(3), 535-60.
Sanfey, Alan G. 2007. "Social Decision-Making: Insights from Game Theory and Neuroscience." *Science*, 318(5850), 598-602.
Sanfey, Alan G.; James K. Rilling; Jessica A. Aronson; Leigh E. Nystrom and Jonathan D. Cohen. 2003. "The Neural Basis of Economic Decision-Making in the Ultimatum Game." *Science*, 300(5626), 1755-58.
Schoemaker, Paul J.H. and Philip E. Tetlock. 2012. "Taboo Scenarios: How to Think About the Unthinkable." *California Management Review*, 54(2), 5-24.
Sen, Amartya K. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy & Public Affairs*, 6(4), 317-44.
Shenhav, Amitai and Joshua D. Greene. 2014. "Integrative Moral Judgment: Dissociating the Roles of the Amygdala and Ventromedial Prefrontal Cortex." *The Journal of Neuroscience*, 34(13), 4741-49.
Shweder, Richard A.; Jonathan Haidt; Randall Horton and Craig Joseph. 2008. "The Cultural Psychology of the Emotions: Ancient and Renewed," *Handbook of Emotions, 3rd Ed.* New York, NY, US: The Guilford Press, 409-27.
Simon, Herbert A. 1955. "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics*, 69(1), 99-118.
Singer, Tania and Claus Lamm. 2009. "The Social Neuroscience of Empathy." *Annals of the New York Academy of Sciences*, 1156(1), 81-96.
Skitka, Linda J. 2010. "The Psychology of Moral Conviction." *Social and Personality Psychology Compass*, 4(4), 267-81.
Skitka, Linda J.; Christopher W. Bauman and Edward G. Sargis. 2005. "Moral Conviction: Another Contributor to Attitude Strength or Something More?" *Journal of Personality and Social Psychology*, 88(6), 895-917.
Smith, Adam. 1982. *The Theory of Moral Sentiments*. Indiana: Liberty Fund.
Smith, Alexander. 2011. "Group Composition and Conditional Cooperation." *The Journal of Socio-Economics*, 40(5), 616-22.
Smith, Vernon L. and Bart J. Wilson. 2014. "Fair and Impartial Spectators in Experimental Economic Behavior." *Review of Behavioral Economics*, 1(1–2), 1-26.

____. 2019. *Humanomics: Moral Sentiments and the Wealth of Nations for the Twenty-First Century*. Cambridge: Cambridge University Press.
____. 2017. "Sentiments, Conduct, and Trust in the Laboratory." *Social Philosophy and Policy*, 34(1), 25-55.
Sobel, Joel. 2005. "Interdependent Preferences and Reciprocity." *Journal of Economic Literature*, 43(2), 392-436.
Song, Myung-ja; Judith G. Smetana and Sang Yoon Kim. 1987. "Korean Children's Conceptions of Moral and Conventional Transgressions." *Developmental Psychology*, 23(4), 577-82.
Stevens, Francis and Katherine Taber. 2021. "The Neuroscience of Empathy and Compassion in Pro-Social Behavior." *Neuropsychologia*, 159, 107925.
Sugden, Robert. 1984. "Reciprocity: The Supply of Public Goods through Voluntary Contributions." *The Economic Journal*, 94(376), 772-87.
Takahashi, Hidehiko; Noriaki Yahata; Michihiko Koeda; Tetsuya Matsuda; Kunihiko Asai and Yoshiro Okubo. 2004. "Brain Activation Associated with Evaluative Processes of Guilt and Embarrassment: An Fmri Study." *NeuroImage*, 23(3), 967-74.
Tangney, June Price; Jeff Stuewig and Debra J. Mashek. 2007. "Moral Emotions and Moral Behavior." *Annual Review of Psychology*, 58(1), 345-72.
Tetlock, Philip E. 2003. "Thinking the Unthinkable: Sacred Values and Taboo Cognitions." *Trends in Cognitive Sciences*, 7(7), 320-24.
Tetlock, Philip E.; Barbara A. Mellers and J. Peter Scoblic. 2017. "Sacred Versus Pseudo-Sacred Values: How People Cope with Taboo Trade-Offs." *American Economic Review*, 107(5), 96-99.
Thöni, Christian and Stefan Volk. 2018. "Conditional Cooperation: Review and Refinement." *Economics Letters*, 171, 37-40.
Tomasello, Michael and Amrisha Vaish. 2013. "Origins of Human Cooperation and Morality." *Annual Review of Psychology*, 64(1), 231-55.
Tooby, John and Leda Cosmides. 1990. "The Past Explains the Present: Emotional Adaptations and the Structure of Ancestral Environments." *Ethology and Sociobiology*, 11(4), 375-424.
Tungodden, Bertil. 2004. "Some Reflections on the Role of Moral Reasoning in Economics," NHH,
van Wolkenten, Megan; Sarah F. Brosnan and Frans B. M. de Waal. 2007. "Inequity Responses of Monkeys Modified by Effort." *Proceedings of the National Academy of Sciences*, 104(47), 18854-59.
Vanberg, V. J. 2008. "On the Economics of Moral Preferences." *American Journal of Economics and Sociology*, 67(4), 605-28.
Warneken, Felix; Brian Hare; Alicia P. Melis; Daniel Hanus and Michael Tomasello. 2007. "Spontaneous Altruism by Chimpanzees and Young Children." *PLOS Biology*, 5(7), e184.
Warneken, Felix and Michael Tomasello. 2007. "Helping and Cooperation at 14 Months of Age." *Infancy*, 11(3), 271-94.
____. 2009. "The Roots of Human Altruism." *British Journal of Psychology*, 100(3), 455-71.

Weimann, Joachim. 1994. "Individual Behaviour in a Free Riding Experiment." *Journal of Public Economics*, 54(2), 185-200.

Wiech, Katja; Guy Kahane; Nicholas Shackel; Miguel Farias; Julian Savulescu and Irene Tracey. 2013. "Cold or Calculating? Reduced Activity in the Subgenual Cingulate Cortex Reflects Decreased Emotional Aversion to Harming in Counterintuitive Utilitarian Judgment." *Cognition*, 126(3), 364-72.

Wittig, Martina; Keith Jensen and Michael Tomasello. 2013. "Five-Year-Olds Understand Fair as Equal in a Mini-Ultimatum Game." *Journal of Experimental Child Psychology*, 116(2), 324-37.

Wittmann, Marco K.; Patricia L. Lockwood and Matthew F.S. Rushworth. 2018. "Neural Mechanisms of Social Cognition in Primates." *Annual Review of Neuroscience*, 41(1), 99-118.

Wynne, Clive D. L. 2004. "Fair Refusal by Capuchin Monkeys." *Nature*, 428(6979), 140-40.

Young, Liane and James Dungan. 2012. "Where in the Brain Is Morality? Everywhere and Maybe Nowhere." *Social Neuroscience*, 7(1), 1-10.

Yu, Hongbo; Jie Hu; Li Hu and Xiaolin Zhou. 2013. "The Voice of Conscience: Neural Bases of Interpersonal Guilt and Compensation." *Social Cognitive and Affective Neuroscience*, 9(8), 1150-58.

Zahn, Roland; Ricardo de Oliveira-Souza; Ivanei Bramati; Griselda Garrido and Jorge Moll. 2009. "Subgenual Cingulate Activity Reflects Individual Differences in Empathic Concern." *Neuroscience Letters*, 457(2), 107-10.

Zahn, Roland; Ricardo de Oliveira-Souza and Jorge Moll. 2020. "Moral Motivation and the Basal Forebrain." *Neuroscience & Biobehavioral Reviews*, 108, 207-17.

Zahn, Roland; Ricardo de Oliveira-Souza and Jorge Moll. 2013. "Moral Emotions," J. Armony and P. Vuilleumer, *The Cambridge Handbook of Human Affective Neuroscience.* New York, United States of America: Cambridge University Press, 491-508.

Zajonc, Robert B. and Hazel Markus. 1982. "Affective and Cognitive Factors in Preferences." *Journal of Consumer Research*, 9(2), 123-31.

Zizzo, Daniel John. 2003. "Money Burning and Rank Egalitarianism with Random Dictators." *Economics Letters*, 81(2), 263-66.

Zizzo, Daniel John and Andrew J. Oswald. 2001. "Are People Willing to Pay to Reduce Others' Incomes?" *Annales d'Économie et de Statistique*, (63/64), 39-65.