Mathematical foundations of moral preferences

Valerio Capraro[1,*] and Matjaž Perc[2,3,4,†]

[1]*Department of Economics, Middlesex University, The Burroughs, London NW4 4BT, U.K.*

[2]*Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, 2000 Maribor, Slovenia*

[3]*Department of Medical Research, China Medical University Hospital, China Medical University, Taichung 404, Taiwan*

[4]*Complexity Science Hub Vienna, Josefstädterstraße 39, 1080 Vienna, Austria*

(Dated: December 16, 2020)

Abstract

One-shot anonymous unselfishness in economic games is commonly explained by social preferences, which assume that people care about the monetary payoffs of others. However, during the last ten years, research has shown that different types of unselfish behaviour, including cooperation, altruism, truth-telling, altruistic punishment, and trustworthiness are in fact better explained by preferences for following one's own personal norms – internal standards about what is right or wrong in a given situation. Beyond better organising various forms of unselfish behaviour, this moral preference hypothesis has recently also been used to increase charitable donations, simply by means of interventions that make the morality of an action salient. Here we review experimental and theoretical work dedicated to this rapidly growing field of research, and in doing so we outline mathematical foundations for moral preferences that can be used in future models to better understand selfless human actions and to adjust policies accordingly. These foundations can also be used by artificial intelligence to better navigate the complex landscape of human morality.

*Electronic address: v.capraro@mdx.ac.uk

†Electronic address: matjaz.perc@gmail.com

I. INTRODUCTION

Most people are not completely selfish. Given the right circumstances, they are happy to give up a part of their benefit to help other people or the society as a whole. Psychologists and economists have long observed that some people act unselfishly even in one-shot anonymous interactions, when there are no direct or indirect benefits for doing so [1, 2]. The question is why? Understanding what motivates people to act unselfishly in one-shot, anonymous interactions is of great theoretical and practical importance. Theoretically, it may lead to a more complete and precise mathematical framework to formalise human decision-making, while practically, it may suggest policies and interventions to promote unselfish behaviour, with the ultimate goal of improving our societies.

To study one-shot unselfishness, behavioural scientists usually turn to laboratory experiments using economic games, in which experimental subjects have to make monetary decisions that involve various forms of other-regarding behaviour. In this context, and throughout this review, selfishness and other-regarding behaviour is defined with respect to monetary payoffs. Clearly, a behaviour that is unselfish from the point of view of monetary outcomes may turn out to be selfish from a more general perspective that takes into account also psychological benefits and costs. For example, some people may engage in unselfish behaviour to decrease negative mood [3] or increase positive feelings [4]. Therefore, in the last decades, behavioural scientists have been trying to mathematically explain unselfish behaviour by means of a utility function that depends on factors other than solely the monetary payoff of the decision maker. Based on empirical data scholars have initially advanced the explanation that human unselfishness in one-shot anonymous interactions is primarily driven by people not caring only about their own monetary payoff, but caring, at least to some extent, also about the monetary payoffs of the other people involved in the interaction [5–10].

However, about fifteen years ago, this *social preference hypothesis* came under critique because some experiments showed that two particular forms of unselfish behaviour, altruistic punishment and altruism, could not be entirely explained by preferences defined solely over monetary outcomes. In 2010, building on work on the effect of social norms on people's behaviour [11–20], Bicchieri and Chavez [21] proposed to explain altruistic punishment assuming that people have preferences for following their "personal norms" (what they personally believe to be the right thing to do) beyond the monetary consequences that this action brings about. Subsequently, Krupka and

Weber [22] proposed to explain altruism using "injunctive norms" (what one believes others would approve/disapprove); however, in their analysis, they did not consider a potential role of personal norms. In the last five years, numerous other experiments challenged social preference models in several behavioural domains, other than altruistic punishment and altruism [23–30]; moreover, the best interpretation of these results turns out to be in terms of personal norms, rather than other types of norms. Namely, the best way to organise these results is through the moral preference hypothesis, according to which people have preferences for following their personal norms, beyond the economic consequences that these actions bring about. This framework organises several forms of one-shot, anonymous unselfish behaviour, including cooperation, altruism, altruistic punishment, trustworthiness, honesty, and the equality-efficiency trade-off. We note at this stage that personal norms are not universally given. They certainly depend on the culture; for example, they can come from the internalisation of cultural values [15]. But they can also depend on the individual; anecdotal evidence suggests that, even within the same family, there might be people with different beliefs about what is right or wrong in a given situations. We will discuss this in more details in Section VII F.

The moral preference hypothesis also holds promise of being very useful in practice. The idea is simple. If people care about doing the right thing, then just providing cues that make the rightness of an action salient should work just fine in promoting desirable behaviour. In fact, research has already demonstrated the applicability of this approach outside of the laboratory, showing in particular that nudges towards doing the right thing can increase charitable donations [31].

In the light of ample empirical research supporting the moral preference hypothesis, theoretical research aiming to formalise human decision-making by means of a mathematical framework is also at a crossroads. On the one hand, the traditional approach involving monetary payoffs has worked well in explaining many challenging aspects of pro-social behaviour. But on the other, experiments indicate that there are likely hard boundaries to this simplistic approach, which will thus have to be amended by more avant-garde concepts, including formalising the intangibles of moral psychology and philosophy.

Here we review this rapidly growing field of research within the following sections. Section II reviews the main economic games that have been developed to study one-shot unselfishness. Section III reviews social preference models, as earlier attempts to explain unselfishness in one-shot economic games within a unified theoretical framework. This section also describes a number of experiments that violate social preference models. Section IV shows how these experiments

can be organised by general moral preferences for doing what one believes to be the right thing. Section V focuses on practical applications of the moral preference hypothesis. Section VI reviews the models of moral preferences that have been introduced so far and proposes a new model that explicitly takes into account the importance of personal norms. Lastly, Section VII outlines a number of key questions for future work, while Section VIII summarises the main conclusions.

Taken together, this review thus outlines a mathematical formalism for morality, which shall inform future models aimed at better understanding selfless actions as well as artificial intelligence that strives to emulate counterintuitive human decision-making.

II. MEASURES OF UNSELFISH BEHAVIOUR

There are various forms of unselfish behaviour. For example, giving money to a homeless person on the street is, in principle, quite different from collaborating with a colleague on a common project, or from telling the truth when one is tempted to lie. To take this source of heterogeneity into account, scholars have developed a series of simple games and decision problems that are meant to prototypically represent different types of unselfish behaviour. These are simple scenarios in which experimental subjects can make decisions that have real consequences. To incentivise these decisions, behavioural scientists usually use monetary payoffs (at least among adult subjects, whereas other forms of remuneration, such as stickers, might be more effective among children).

In this review, we will be mainly focused on one-shot decisions that are *purely* unselfish, meaning that they bring no monetary benefit to the decision maker (and possibly bring a cost), no matter the beliefs of the decision maker regarding the behaviour of other people involved in the interaction. Specifically, we measure altruistic behaviour using the dictator game (see Table 1 for all the definitions), cooperative behaviour in pairwise interactions using the prisoner's dilemma, truth-telling using the sender-receiver game, the tradeoff between equality and efficiency using the trade-off game, trustworthiness using player 2 in the trust game, and altruistic punishment using Player 2 in the ultimatum game. In the last section we will also briefly consider decisions that are *strategically* unselfish, such as trust (player 1 in the trust game) and strategic fairness (player 1 in the ultimatum game), which might actually maximise the payoff of the decision maker, depending on their beliefs about the behaviour of the second player. The distinction between pure unselfishness and strategic unselfishness generalises the distinction between pure cooperation and strategic cooperation, introduced by Rand in his meta-analysis [32].

III. SOCIAL PREFERENCES AND THEIR LIMITATIONS

Behavioural scientists have long recognised that some people do act unselfishly even in one-shot anonymous interactions. For example, the first comprehensive empirical work on the one-shot prisoner's dilemma dates back to 1965 [1]. Formal frameworks to explain one-shot unselfishness have a more recent history, starting in 1994, when Ledyard observed that cooperation, altruism, and altruistic punishment could be explained by assuming that people maximise a utility function that depends not only on their own monetary payoff, but also on the total monetary payoff of the other people that are involved in the interaction [5]. See Table 2 for the exact mathematical definition. Since then, several models have been introduced. In 1998, Levine [6] proposed a utility function in which the level of altruism depends on the level of altruism of the other players. Subsequently, in 1999, Fehr and Schmidt [7] proposed a framework according to which players care about minimising inequities. In 2000, Bolton and Ockenfels [8] followed a similar idea and introduced a general inequity aversion model, in which the utility of an action depends negatively on the distance between the amount of money the decision maker gets if that action is implemented and the amount of money the decision maker would get if the equal allocation were implemented. The authors proposed an explicit mathematical formula only for the case of $n = 2$ players. In 2002, Andreoni and Miller [9] estimated the behaviour of experimental subjects in a number of dictator game choices using a specific utility function taking into account altruistic tendencies as well as potential convexity in the preferences. In the same year, Charness and Rabin [10] introduced a general utility function which, depending on the relative relationship between its two parameters, can cover several cases, including competitive preferences, inequity aversion preferences, and social efficiency preferences. We refer to Table 2 for the exact functional forms. (Besides these models, scholars have also studied models that can be applied to specific subsets of one-shot anonymous interactions (e.g., [4]). In this review, we focus on models that can be applied to any one-shot anonymous interaction involving unselfish behaviour).

While differing in many details, all social preference models share one common property: they assume that the utility of a decision maker is a function of the monetary payoffs of the available actions. This assumption came under considerable criticism for the first time in 2003 when Falk, Fehr and Fischbacher [33] showed that rejection rates in the ultimatum game depend on the choice set available to the proposer. Specifically, the split (8,2) — 8 to the proposer and 2 to the responder — is more likely to be accepted in ultimatum games in which the only other choice available to

the proposer is (10,0), compared to ultimatum games in which the only other choice available to the proposer is (5,5). Therefore, responders prefer accepting (8,2) over rejecting it in the former case, but they prefer rejecting (8,2) over accepting it in the latter one, despite the fact that these choices have the same monetary consequences in the two cases. Clearly, this cannot be explained by any model of social preferences. See [21, 26] for conceptual replications.

Shortly after, in 2005, Uri Gneezy introduced the sender-receiver game [34]. In his experiments, decision makers were less likely to implement an allocation of money when implementing this allocation also required misreporting a private information. Also this finding cannot be explained by any model of social preferences and, more generally, also not by any utility function that depends only on the monetary payoffs that are associated with the available actions. This thus indicates that (some) people have an intrinsic cost of lying, which goes beyond their preferences toward monetary outcomes. To further support this interpretation, several scholars have independently studied the sender-receiver game in contexts in which lying would benefit both the sender and the receiver to the same extent. This case is particularly important because, when the benefit for the sender is equal to the benefit for the receiver, all social preference models predict that the totality of people would lie. However, this prediction turned out to be violated in experiments, which showed that a significant proportion of people tell the truth [23, 35, 36].

Subsequently social preference models came under critique also in one of the behavioural domains in which they had been most successful, namely in research involving the dictator game. Dana, Cain and Dawes [37] and Lazear, Malmendier and Weber [38] observed that some dictator game givers would prefer to altogether avoid the dictator game interaction if given the chance. These people thus preferred giving over keeping in a context in which they were forced to play the dictator game, but preferred keeping over giving in a context in which they could choose whether to play the dictator game or not. This finding, as in the preceding examples, cannot be explained by any utility function that is based solely on monetary outcomes.

For the same game, and along similar lines, List [39], Bardsley [40], and Cappelen et al. [41] found that extending the choice set of the dictator by adding the possibility to take money from the recipient has the effect to make some dictators less likely to give. Therefore, these dictators preferred giving over keeping, when the taking option was not available, but preferred keeping over giving, when the taking option was available. This finding likewise cannot be explained by any preference over monetary payoffs. A conceptually similar point was also made by Krupka and Weber [22] and Capraro and Vanzo [29], who found that even minor changes in the instructions of

the dictator game can notably impact people's behaviour.

In the years after 2013, the inability of purely monetary-based models to explain empirically observed behaviour engulfed many other games and decision problems, whose experimental regularities had been previously thought to be explainable in terms of social preferences. Examples included the prisoner's dilemma [25, 27], the trust game [25], as well as different variants of the trade-off game [27, 28, 30], thus resulting in a crisis of the social preference hypothesis.

IV. THE RISE OF THE MORAL PREFERENCE HYPOTHESIS

To solve a crisis, one needs a paradigm shift. The shift started in 2010, when Bicchieri and Chavez [21] proposed an elegant solution for one of the aforementioned empirical observations. This solution builds on classic work suggesting that, in everyday life, people's behaviour is partly determined by what they believe to be the norms in a given context [11–20]. This observation led behavioural scientists to propose several classifications of norms. Particularly relevant for the thesis of this review is the distinction between personal and social norms [15]. And moreover, among the social norms, the distinction between injunctive and descriptive norms [17]. Personal norms refer to internal standards about what is right or wrong in a given situation; injunctive norms refer to what other people approve or disapprove of in that situation; descriptive norms refer to what other people actually do. In one-shot anonymous games, like the games considered in this review, the distinction among personal, descriptive, and injunctive norms roughly corresponds to Bicchieri's personal normative beliefs, empirical expectations, and normative expectations [18]. See Table 3 for precise definitions.

The groundbreaking intuition of Bicchieri and Chavez [21] was to apply the theory of norms to deviations from monetary-based social preferences in the ultimatum game. Specifically, Bicchieri and Chavez showed that the ultimatum game offer that is consider to be fair by responders depends on the choice set available to the proposer; moreover, responders tend to reject offers that they consider unfair. This suggests that altruistic punishment is driven by responders following their personal norms, beyond the monetary consequences that these actions bring about. In particular, this explains the aforementioned results of Falk, Fehr, and Fischbacher [33], that responders reject the same offer at different rates depending on the other offers available to the proposer.

Shortly after, in 2013, Krupka and Weber [22] applied a similar approach to several variants of the dictator game. However, instead of focusing on personal norms, they focused on injunctive

norms. For each of the available actions, subjects were asked to declare whether they found the corresponding action to be "very socially inappropriate", "somewhat socially inappropriate", "somewhat socially appropriate", or "very socially appropriate". Subjects were given a monetary prize if they matched the modal choice made by other participants. Observe that, in this way, Krupka and Weber incentivised the elicitation of the injunctive norm. (The elicitation of personal norms cannot be incentivised.) In doing so, Krupka and Weber found that people believe that others think that avoiding a dictator game interaction is far less socially inappropriate than keeping the whole amount of money in a dictator game that one is obliged to play. Therefore, the empirical results summarised above regarding dictator games with an exit option [37, 38] can be explained simply by a change in the perception of what is the injunctive norm in that context. Similarly, Krupka and Weber found that people believe that others think that keeping the money in a dictator game with a taking option is far less socially inappropriate than keeping the money in the dictator game without the taking option. In this way, they could explain also the results of List [39], Bardsley [40], and Cappelen et al. [41] in terms of a change in the perception of the injunctive norm. Finally, Krupka and Weber presented a novel experiment in which subjects played the dictator game in either of two variants: in the Standard variant, dictators started with \$10 and had to decide how much of it, if any, to give to the recipient; in the Bully variant, the money was initially split equally among the dictator and the recipient, and the dictator could either give, take, or do nothing. The authors found that people were more altruistic in the Bully variant compared to the Standard variant, and this was driven by the fact that people rated "taking from the recipient" far less socially appropriate than "not giving to the recipient".

The work of Krupka and Weber suggests that taking into account injunctive norms is important to explain deviations from social preference models in the dictator game. But are the injunctive norms really the main force behind the observed behavioural changes, or are there also other norms playing more primary roles? In the last five years, a set of empirical studies tried to address this question. Schram and Charness [24] analysed the behaviour of dictators who were given an advice from third parties about the injunctive norm. They observed that dictators became more pro-social only when their choices were made public. By contrast, when their choices remained private, they found no significant increase in pro-sociality, compared to the case in which they did not receive any information about the injunctive norm. These results indicate that, although injunctive norms might correlationally explain behavioural changes in anonymous (and thus private) dictator game experiments, they are unlikely to be the primary motivation. In fact, being that these games were

played anonymously, in front of the screen of a computer, the intuition suggests that the norms primarily at play are the personal norms. Two recent works provide evidence for this hypothesis. Capraro and Vanzo [29] found that framing effects in the dictator game generated by morally loaded instructions can be explained by changes in the perception of what people "personally think to be the right thing" in the given context (i.e., their personal norms). Capraro et al. [31] showed that making personal norms salient prior to playing the dictator game (by asking subjects to state what they personally think to be the morally right thing to do) has a strong effect on subsequent dictator game donations, even persisting to a second-stage prisoner's dilemma interaction.

This set of works thus suggests that dictator game giving is driven by personal norms. Putting this together with the results of Bicchieri and Chavez, we obtain that both altruism and altruistic punishment can be explained by people following their personal norms.

More recently, this finding has been not only replicated, but, more importantly, also extended to explain several other forms of unselfish behaviour. In 2016, Kimbrough and Vostroknutov [25] introduced a task "that measures subjects' preferences for following rules and norms, in a context that has nothing to do with social interaction or distributional concerns". They found that this measure of norm-sensitivity predicts dictator game altruism, trust game trustworthiness (but not trust), and ultimatum game rejection thresholds (but not offers). Taken together, this indicates that altruism, trustworthiness, and altruistic punishment are driven by a common desire to adhere to a personal norm. In 2017, Eriksson et al. [26] conducted an ultimatum game experiment under two different conditions. The difference, however, was only in the labels that were used to describe the action of refusing the proposer's offer. In one treatment, this action was labeled "rejecting the proposer's offer", while in the other treatment, the same action was labeled "reducing the proposer's payoff". Since these two options are monetarily equivalent, any utility function depending only on the monetary payoffs of the available actions predict that responders should behave the same way in both cases. But contrary to this prediction, Eriksson et al. found that responders displayed higher rejection thresholds in the "rejection frame" than in the "reduction frame". Moreover, they showed that the observed framing effect could be explained by a change in what people think to be the right thing to do. Specifically, subjects tended to rate the action of reducing the proposer's offer to be morally worse than the action of rejecting the proposer's offer, in spite of the fact that these two actions had the same monetary consequences. In 2018, Capraro and Rand [27] showed that behaviour in the trade-off game is highly sensitive to the labels used to describe the available actions. In line with Eriksson et al. [26], Capraro and Rand also found that their framing effects

could be explained by a change in what people think to be the right thing to do. Notably, framing effects in the trade-off game have been replicated several times [28, 30, 42–44] and a recent work has shown that these moral framings tap into relatively internalised moral preferences [44]. Moreover, Capraro and Rand also considered a situation in which the personal norm conflicted with the descriptive norm, and found that people tend to follow the personal norm, rather than the descriptive norm. The same research also revealed a correlation between the framing effect in the trade-off game and giving in the dictator game and cooperation in the prisoner's dilemma, thus indicating that not only trade-off decisions are driven by personal norms, but that altruism and cooperation are also subject to that same facilitator. Cooperative behaviour is also typically correlated to altruistic behaviour [45–47], suggesting that they are driven by a common underlying motivation.

To the best of our knowledge, there are no works directly exploring the role of personal norms on truth-telling in the sender-receiver game. However, Biziou-van-Pol et al. [23] have shown that there is a positive correlation between truth-telling in the sender-receiver game (in the Pareto white lie condition), giving in the dictator game, and cooperation in the prisoner's dilemma, suggesting that these types of behaviours are driven by a common motivation. Since the aforementioned research suggests that altruism and cooperation are driven by personal norms, this correlation suggests that lying aversion is so too.

In sum, research accumulated in the last ten years suggests that several forms of one-shot, anonymous unselfishness, including altruism, altruistic punishment, truth-telling, cooperation, trustworthiness, and the equality-efficiency trade-off, can be explained using a unified theoretical framework, whereby people have moral preferences for following their personal norms, beyond the monetary payoff that these actions bring about. Of course, this is not meant to imply that monetary payoffs do not play any role in explaining one-shot unselfishness, but simply that something else, in addition to monetary payoffs, should be taken into account. The thesis is that this 'something else' are the personal norms, which gives rise to the moral preference hypothesis as described in Table 4. Also, this is not meant to imply that other types of norms play no role in these forms of one-shot selfless behaviour. For example, nudging the injunctive norm in the prisoner's dilemma [31] and in the trade-off game [48] has a similar effect as nudging the personal norm. Moreover, it is possible that social norms ultimately drive personal norms, because they allow to enhance or preserve one's sense of self-worth and avoid self-concept distress, resulting in a self-reinforcing behaviour that eventually benefits one's own self-image [15]. However, the aforementioned liter-

ature suggests that, at a proximate level, personal norms have a greater explanatory power, in the sense that they consistently explain people's behaviour also in games where injunctive norms have been shown to play a limited role (e.g., dictator game) or where descriptive norms play a limited role (e.g., the trade-off game).

V. PRACTICAL APPLICATIONS

Behavioural scientists and policy makers have been using norm-based interventions to foster pro-sociality in real life for decades [49–60]. Although these paternalistic interventions have been criticised because they subtly violate people's freedom of choice [61] and can be exploited by malicious institutions [62] (see [63] for a response to these critiques), they are well-studied because, compared to standard procedures to foster pro-sociality (punishment and rewards), they allow to save the monitoring cost that the institution needs to pay in order to know who to punish or reward.

Norm-based interventions typically manipulate the descriptive or the injunctive norm in a given context, and show that this has an effect on people's behaviour in that same context. The more recent works reviewed in the previous section, showing that unselfish behaviour in one-shot, anonymous economic games is primarily driven by a desire to follow the personal norms, suggest that a more effective mechanism to increase pro-sociality might be to use norm-based interventions that target personal norms, rather than social norms. The interest in targeting personal norms, compared to other mechanisms to promote pro-sociality, is also that targeting personal norms is potentially cheaper than other mechanisms. Clearly, it is cheaper than punishment and rewards because it avoids the monitoring cost. Additionally, it saves the cost of collecting information about the behaviour or the moral judgments of other people, which forms the basis of interventions targeting social norms.

In recent years, there has been a growing body of research exploring the effect of nudging personal norms on various forms of unselfish behaviour. Some works using economic games found that making personal norms salient increases donations in the dictator game [31, 64], cooperation in the prisoner's dilemma [31, 65], as well as decreases in-group favouritism, at least on average [66]. This suggests that nudging personal norms might be effective to increase pro-sociality in one-shot anonymous decisions that have consequences outside the laboratory. Along these lines, Capraro et al. [31] found that asking people to report what they personally think is the morally right thing to do increases crowdsourced charitable donations by 44%.

VI. MODELS OF MORAL PREFERENCES

We have thus seen that several forms of unselfish behaviour can be organised by moral preferences for following the personal norms. The question is, can we model this using a formal utility function?

There have been some attempts to formalise people's tendency to follow a norm [22, 25, 67–75]. Most of these models, however, are either very specific in the sense that they can be applied only to certain games, or do not distinguish among different types of norms. Three models can be applied to every game of interest in this review (and, more generally, to every one-shot game) and distinguish among different types of norms.

Levitt and List [68] introduced a model where the utility of an action a depends on the monetary payoff associated to that action, $v_i(\pi_i(a))$, as well as on the moral cost (or benefit), $m(a)$, associated to that action:

$$u_i(a) = v_i(\pi_i(a)) + m(a).$$

Levitt and List assumed that the moral cost (or benefit) depends primarily on three factors: whether the action is recorded or performed in the presence of an observer, whether the action has negative consequences on other players, and whether the action is in line with "social norms or legal rules that govern behavior in a particular society". Therefore, Levitt and List's model, although useful in many circumstances, it does only mention social norms, while ignoring the effect of personal norms.

A similar model was considered by Krupka and Weber [22], with the key difference that they focused on injunctive norms specifically. Krupka and Weber introduced a function N defined over the set of available actions that, given an action a, returns a number $N(a)$ representing the extent to which society views a as socially appropriate. They also assumed that people are heterogeneous in the extent to which they care about doing what society considers to be appropriate. In doing so, they obtain the utility function:

$$u_i(a) = v_i(\pi_i(a)) + \gamma_i N(a).$$

As mentioned above, one of the main contributions of Krupka and Weber was to introduce an experimental technique to elicit the injunctive norm. To this end, they asked participants to rate

each of the available actions in terms of their social appropriateness. Participants were incentivised to match the modal choice of the other participants.

Very recently, in 2020, Kimbrough and Vostroknutov presented a different approach, but still based on injunctive norms [75]. Specifically, they introduced the utility function

$$u_i(a) = v_i(\pi_i(a)) + \phi_i \eta(a),$$

where ϕ_i represents the extent to which i cares about following the injunctive norm, and $\eta(a)$ represents a measure of whether the society thinks that a is socially appropriate. Although this utility function looks very similar to the one proposed by Krupka and Weber, it differs from it in one important dimension. While Krupka and Weber's social appropriateness, $N(a)$, is computed by asking participants what they think others would approve or disapprove (and therefore it need not depend only on the monetary consequences of the available actions), Kimbrough and Vostroknutov's injunctive norm, η, is built axiomatically from the game and it is assumed to be inversely proportional to the overall dissatisfaction of the players, defined as the difference between what they get in a given scenario and what they could have gotten in others. This implies that one limitation of this approach is that people always prefer Pareto dominant allocations over Pareto dominated ones. But, in experiments, this property is not always satisfied. For example, when lying is Pareto dominant, some people still tell the truth, and these people tend to cooperate in a subsequent prisoner's dilemma and give in a subsequent dictator game [23]. Moreover, in trade-off games framed in such a way that the Pareto dominant allocation is presented as morally wrong, people tend to choose the Pareto dominated option [27, 28].

In sum, previous formal models consider only social norms or, more specifically, injunctive norms. But, as we have seen in the previous sections, unselfish behaviour in one-shot anonymous interactions is often driven by personal norms, rather than by social norms. Taking inspiration from the above models, one can formalise this using the utility function:

$$u_i(a) = v_i(\pi_i(a)) + \mu_i P_i(a),$$

where μ_i represents the extent to which player i cares about doing what s/he personally thinks to be the morally right thing to do and $P_i(a)$ represents the extent to which i personally thinks that a is morally right. This functional form might superficially seem similar to the ones discussed earlier, but it differs from those in two important points. One point is that the personal norm $P_i(a)$

typically depends on the individual i, whereas the injunctive norm depends only on the society in which the individual lives. The second point is the very fact that P_i represents the extent to which i thinks that a is the morally right thing to do, whereas $m(a), N(a)$, and $\eta(a)$ represent social norms. In general, the personal norm might not be aligned with the social norms. In practice, $P_i(a)$ can be estimated using a suitable experiment, whereas μ_i and v_i can be estimated, on average, using statistical techniques, following a similar method as the one developed by Krupka and Weber for injunctive norms [22]. Specifically, one can estimate $P_i(a)$ by asking subjects to self-report the extent to which they personally think that action a is the morally right thing to do. Then one can use these ratings to predict the behaviour, using a simple regression. The coefficient of this regression will give the average of the μ_i's. Also, putting the monetary payoffs in the regression, one can also get an estimation for the average of the v_i's.

This utility function based on personal norms has a greater predictive power than its counterparts based only on social norms, in the sense that it explains behaviour in a larger set of games, compared to their counterparts based on social norms. We have seen earlier that Schram and Charness [24] found that making the injunctive norm salient does *not* increases altruistic behaviour in the anonymous dictator game. D'Adda et al. [53] found that making the descriptive norm salient has only a marginally significant effect on anonymous dictator game giving; this effect also vanishes in a second interaction, played immediately after. Along the same lines, Dimant, van Kleef and Shalvi [76] found that promoting the injunctive norm and promoting the descriptive norm does *not* have any effect on people's honesty in a deception game in which subjects can lie for their benefit. On the other hand, numerous works have shown that nudging personal norms impacts several forms of unselfish behaviour, ranging from altruism [31, 64], altruistic punishment [26], cooperation [31, 65], and the equality-efficiency trade-off [27]. Moreover, the effect typically persists for at least another interaction and even spills across contexts [31]. All these results are consistent with a utility function based on personal norms and are not consistent with a utility function based only on social norms.

We present a summary of all above-discussed moral preference models in Table 5.

VII. FUTURE WORK

This is an exciting field of research, which provides a unified view of human choices in several contexts of decision-making, while having, at the same time, significant practical implications.

Nonetheless, there are several questions that need to be explored in future research, as detailed in what follows and summarised in Table 6.

A. The utility function

From a mechanistic perspective, the moral preference hypothesis raises the question of how can we express the utility function of a decision maker. Scholars have tried to give mathematical sense to people's morality since the foundation of mathematical economics [77, 78]. About two centuries later, the question is still open, even in the simple setting of one-shot anonymous interactions. One simple way to do so is to assume that people are torn between maximising their monetary payoff and doing what they personally think to be the morally right thing. This can be done with a utility function of the shape $u_i(a) = v_i(\pi_i(a)) + \mu_i P_i(a)$. Although this utility function outperforms their counterparts based on social norms, as well as social preferences, it undoubtedly represents just a first candidate. Future work should explore other ways to formalise moral preferences, through finer experiments with the power to detect small variations in how people weight their personal norm against monetary incentives. Future work should also find ways to estimate what people think to be the right thing in a given context, without asking it to the participants in a separate experiment. The literature reviewed above shows that, in many cases, it is enough to change only one word in the instructions of a decision problem to change people's perception of what is the right thing to do in a given context. This suggests that $P_i(a)$ partly depends on the language in which the action a is presented. Exploring this dependence can greatly improve the predictive power of the utility function. How can one do so? Recent work shows that emotional content in messages increases their diffusion in social media [79–81]. Translating this finding in the context of one-shot games, it suggests that the emotions carried by the instructions of the decision problem might contribute to the computation of P_i. Along these lines, it is possible that one can use sentiment analysis to better estimate P_i. Sentiment analysis is a technique developed by computational linguists that allows to assign a polarity to any given piece of text [82]. In principle, this polarity could enter the utility function of a decision maker and work as an additional motivation or obstacle for choosing an action, beyond its monetary consequences. In any case, mathematically describing or at least quantifying the seemingly intangible moral preferences, and in doing so building bridges between computational linguistics, behavioural economics, and moral psychology, is a fascinating direction for future work.

B. Evolution of norms

Where do personal norms come from? One explanation is that they come from the internalisation of behaviours that, although not individually optimal in the short term, they are optimal in the long run. It is therefore important to understand which unselfish behaviours can be selected in the long term, and under which conditions. A promising line of research uses evolutionary game theory and statistical physics to find the conditions that promote the evolution of cooperation on networks [83]. More recently, scholars have started applying similar techniques also to study the evolution of other forms of unselfish behaviour [84], such as truth-telling in the sender-receiver game [85, 86] and trustworthiness in the trust game [87]. Some works along this line have also looked at the evolution of choices in the ultimatum game [88–91]. Future work should extend the same techniques to other forms of unselfish behaviour.

C. Personal norms versus social norms

The experimental literature reviewed in the previous sections suggests that several forms of one-shot, anonymous unselfishness can be unified under a framework according to which people have preferences for following their personal norms. Moreover, preliminary evidence suggests that nudging personal norms can be an effective tool for fostering pro-sociality: making personal norms salient affects altruism, cooperation, altruistic punishment, and trade-off decisions between equality and efficiency [26, 31, 64, 65].

This, of course, does not mean that the social norms play no role at all. For example, nudging injunctive norms has a significant effect on the one-shot, anonymous, prisoner's dilemma [31] and the trade-off game [48]. One question that is still open, however, is whether these effects are fundamentally distinct from the effect of nudging personal norms. It is indeed possible that nudging injunctive norms in these games also nudge personal norms, and this is what makes people change their behaviour. A working paper suggests that people who follow injunctive norms in the trade-off game are different from those who follow personal norms [48]. Therefore, it is possible that a larger model taking into account both personal and injunctive norms might have an even greater predictive power, at least in some contexts, than a model based exclusively on personal norms. Further experiments comparing the effect of nudging different norms are needed to clarify this point. The evidence in this case is indeed still lacunar. One study compared the relative effect

of the descriptive and the injunctive norms in the dictator game, and found that people tend to follow the descriptive norm [49]. Another study compared the relative effect of nudging personal norms and the descriptive norms in the trade-off game, and found that people tend to follow the personal norms [27]. The aforementioned working paper compared the effect of nudging the personal and the injunctive norm in the trade-off game and found that they have a similar effect; moreover, when the two norms are in conflict, some people follow the personal norm and other follow the injunctive norm [48]. This suggests that people's behaviour depends on their focus of attention within an interconnected matrix of norms. Therefore, future work should explore norm salience, also in cases where more than one type of norm is simultaneously made salient.

Research should also go beyond anonymous decisions and investigate what happens when choices are observable. The intuition suggests that when choices are observable, social norms may play a bigger role compared to when they remain private; in line with this intuition, Schram and Charness [24] showed that nudging the injunctive norms impacts public but not private dictator game giving. However, no studies compared the relative effectiveness of targeting different norms in public decisions.

D. Boundary conditions of interventions based on personal norms

Having in mind potential practical applications, another important question concerns the boundary conditions of interventions based on personal norms. From a temporal perspective, previous research suggests that interventions targeting personal norms can last for several interactions within the same experiment [31, 65]. However, it seems unrealistic to expect that their effect will last indefinitely. For example, a recent field experiment targeting injunctive norms found an effect that diminishes after repeated interventions, although it can be restored after waiting a sufficient amount of time between interventions [92]. From the decisional context point of view, there will certainly be behavioural domains in which targeting personal norms might not be as effective. For example, a recent work suggests that risky cooperation in the stag-hunt game is primarily driven by preferences for efficiency, rather than by preferences for following personal norms [42].

E. External validity of interventions based on personal norms

Given the potential relevance of this line of work for the society at large, future studies should explore the external validity of interventions based on personal norms. At the time of this writing, only one study investigated the effect of nudging personal norms in contexts in which decisions have consequences outside the laboratory. This study found that nudging personal norms increases crowdsourced charitable donations to real humanitarian organisations by 44% [31].

F. The moral phenotype and its topology

We have seen that different forms of unselfish behaviour can be explained by a general tendency to do the right thing. We are tempted to call this tendency "moral phenotype", extending the notion of "cooperative phenotype" introduced by Peysakhovich, Nowak, and Rand [46]. See also [47]. In their work, Peysakhovich and colleagues observed that pro-social behaviours in the dictator game, the public goods game (a variant of the prisoner's dilemma with more than two players), and the trust game (both players) were all correlated; and they termed this general pro-social tendency cooperative phenotype. Therefore, the cooperative phenotype is uni-dimensional. On the other hand, the moral phenotype is likely to be multi-dimensional. For example, we have seen earlier that both altruistic punishment and altruistic giving are driven by preferences for doing the right thing. However, Peysakhovich, Nowak, and Rand [46] found that they are not correlated. It is possible that they are not correlated because they come from different personal norms. The multi-dimensionality of morality is not a new idea, and several authors have come to suggest it in the last decades from different routes. For example, Haidt and colleagues argue that differences in people's moral concerns can be explained by individual differences across six "foundations" [93–95]. Kahane, Everett and colleagues have shown that (act) utilitarianism decomposes itself in at least two dimensions [96, 97]. Curry, Mullins, and Whitehouse [98] have reported that seven moral rules are universal across societies, but societies vary on how they rank them. However, we are not aware of any work exploring how different personal norms link to different forms of one-shot unselfishness.

Another topological property of the moral phenotype that deserves further scrutiny is the boundary. Does, for example, the moral phenotype include decisions that are strategically unselfish, such as strategic fairness (ultimatum game offers) and trust (trust game transfers), both of

which maximise the decision maker's payoff depending on the decision maker's beliefs about the behaviour of the other player? Previous evidence is limited and mixed. Bicchieri and Chavez [21] showed that ultimatum game offers are partly driven by normative beliefs; Peysakhovich, Nowak, and Rand [46] found that trustees' decisions correlate with dictator game and public goods game decisions. By contrast, Kimbrough and Vostroknutov [25] found that trustees' and proposers' decisions are not correlated to their measure of norm-sensitivity.

G. A dual-process approach to personal norms

Do personal norms come out automatically, or do they require deliberation? Research recently explored the cognitive basis of unselfish behaviour, by using cognitive process manipulation, such as time pressure and cognitive load, in order to favour instinctive responses [99–108]. It has been shown that promoting intuition favours cooperation [32] and altruistic punishment [109]. The evidence regarding altruism is instead more mixed [110, 111]. Instead, a meta-analysis suggests that intuition decreases truth-telling, when lying harms abstract others, while leaving it unaffected when it harms concrete others [112]. Furthermore, results are inconclusive in the context of trustworthiness and the equality-efficiency trade-off (see [113] for a review). This line of work suggests that whether personal norms come out automatically or require deliberation may not have a general answer, but might depend on the specific behavioural context, and possibly also on the individual characteristics of the decision maker. More work is needed to understand which personal norms, in which context, and for which people, become internalised as automatic reactions.

VIII. CONCLUSIONS

The moral preference hypothesis is emerging as a unified framework to explain a wide range of one-shot, anonymous unselfish behaviours, including cooperation, altruism, altruistic punishment, truth-telling, trustworthiness, and the equality-efficiency trade-off. This framework has promising practical implications, given that interventions making personal norms salient have been shown to be effective at increasing charitable donations. Future work should explore further mathematical formalisations of moral preferences in terms of a utility function, investigate the evolution and internalisation of personal norms, study the external validity and the boundary conditions of policy interventions based on personal norms, compare the relative effectiveness of targeting dif-

ferent types of norms, examine the topology of the moral phenotype, and analyse the cognitive foundations of morality, possibly using a dual-process perspective.

Overall, the goal of this line of research should be to build bridges between different scientific disciplines to arrive at a better, perhaps more mechanistic, explanation of human decision-making. The outlined mathematical formalism for morality should be used to inform future models aimed at better understanding selfless actions, and it should also be used in artificial intelligence to better navigate the complex landscape of human morality and to better emulate human decision-making. Ultimately, the goal is to use the obtained insights to develop more efficient policies and interventions to increase good virtues and decrease bad ones, and to collectively strive towards better human societies.

The past century has seen strict compartmentalisation of different scientific disciplines leading to groundbreaking and important discoveries that might had been impossible without it. But while technology and industry might fare well on idiosyncratic breakthroughs, human societies do not. The grandest challenges of today remind us that sustainable social welfare and organisation require a wholesome interdisciplinary and cross-disciplinary approach, and we hope this review will be an inspiration towards this goal.

Acknowledgments

This work was supported by the Slovenian Research Agency (Grant Nos. P1-0403, J1-2457, J4-9302, and J1-9112).

---

[1] Rapoport, A., Chammah, A. M., and Orwant, C. J. *Prisoner's dilemma: A study in conflict and cooperation*, volume 165. University of Michigan press, (1965).

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

[3] Cialdini, R. B., Darby, B. L., and Vincent, J. E. Transgression and altruism: A case for hedonism. *Journal of Experimental Social Psychology* **9**(6), 502–516 (1973).

[4] Andreoni, J. Impure altruism and donations to public goods: A theory of warm-glow giving. *The eEonomic Journal* **100**(401), 464–477 (1990).

[5] Ledyard, J. O. Public goods: A survey of experimental research. In Handbook of Experimental Economics, J., K. and A., R., editors. Princeton Univ. Press, Princeton, NJ (1995).

[6] Levine, D. K. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics* **1**(3), 593–622 (1998).

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

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

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

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

[11] Smith, A. *The theory of moral sentiments*. Penguin, (2010).

[12] Durkheim, É. *Les règles de la méthode sociologique*. Flammarion, (2017).

[13] Parsons, T. *The Structure of social action: A Study in Social Theory with Special Reference to a Group of Recent European Writers*. Free Press, New York: London, (1937).

[14] Geertz, C. *The interpretation of cultures*, volume 5019. Basic books, (1973).

[15] Schwartz, S. H. Normative influences on altruism. In *Advances in experimental social psychology*, volume 10, 221–279. Elsevier (1977).

[16] Elster, J. Social norms and economic theory. *Journal of Economic Perspectives* **3**(4), 99–117 (1989).

[17] Cialdini, R. B., Reno, R. R., and Kallgren, C. A. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology* **58**(6), 1015–1026 (1990).

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

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

[20] Hawkins, R. X., Goodman, N. D., and Goldstone, R. L. The emergence of social norms and conventions. *Trends in cognitive sciences* **23**(2), 158–169 (2019).

[21] Bicchieri, C. and Chavez, A. Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making* **23**(2), 161–178 (2010).

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

[23] Biziou-van Pol, L., Haenen, J., Novaro, A., Occhipinti Liberman, A., and Capraro, V. Does telling white lies signal pro-social preferences? *Judgment and Decision Making* **10**, 538–548 (2015).

[24] Schram, A. and Charness, G. Inducing social norms in laboratory allocation choices. *Management Science* **61**(7), 1531–1546 (2015).

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

[26] Eriksson, K., Strimling, P., Andersson, P. A., and Lindholm, T. Costly punishment in the ultimatum game evokes moral concern, in particular when framed as payoff reduction. *Journal of Experimental Social Psychology* **69**, 59–64 (2017).

[27] Capraro, V. and Rand, D. G. Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making* **13**, 99–111 (2018).

[28] Tappin, B. M. and Capraro, V. Doing good vs. avoiding bad in prosocial choice: A refined test and extension of the morality preference hypothesis. *Journal of Experimental Social Psychology* **79**, 64–70 (2018).

[29] Capraro, V. and Vanzo, A. The power of moral words: Loaded language generates framing effects in the extreme dictator game. *Judgment and Decision Making* **14**, 309–317 (2019).

[30] Huang, L., Lei, W., Xu, F., Yu, L., and Shi, F. Choosing an equitable or efficient option: A distribution dilemma. *Social Behavior and Personality: An international journal* **47**(10), 1–10 (2019).

[31] Capraro, V., Jagfeld, G., Klein, R., Mul, M., and van de Pol, I. Increasing altruistic and cooperative behaviour with simple moral nudges. *Scientific Reports* **9**(1), 1–11 (2019).

[32] Rand, D. G. Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science* **27**(9), 1192–1206 (2016).

[33] Falk, A., Fehr, E., and Fischbacher, U. On the nature of fair behavior. *Economic inquiry* **41**(1), 20–26 (2003).

[34] Gneezy, U. Deception: The role of consequences. *American Economic Review* **95**(1), 384–394 (2005).

[35] Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. When do we lie? *Journal of Economic Behavior & Organization* **93**, 258–265 (2013).

[36] Erat, S. and Gneezy, U. White lies. *Management Science* **58**(4), 723–733 (2012).

[37] Dana, J., Cain, D. M., and Dawes, R. M. What you don't know won't hurt me: Costly (but quiet) exit

in dictator games. *Organizational Behavior and human decision Processes* **100**(2), 193–201 (2006).

[38] Lazear, E. P., Malmendier, U., and Weber, R. A. Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics* **4**(1), 136–63 (2012).

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

[40] Bardsley, N. Dictator game giving: altruism or artefact? *Experimental Economics* **11**(2), 122–133 (2008).

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

[42] Capraro, V., Rodriguez-Lara, I., and Ruiz-Martos, M. J. Preferences for efficiency, rather than preferences for morality, drive cooperation in the one-shot stag-hunt game. *Journal of Behavioral and Experimental Economics* (2020).

[43] Capraro, V. Gender differences in the trade-off between objective equality and efficiency. *Judgment and Decision Making* **15**(4), 534–544 (2020).

[44] Capraro, V., Jordan, J. J., and Tappin, B. M. Does observability amplify sensitivity to moral frames? evaluating a reputation-based account of moral preferences. *arXiv preprint arXiv:2004.04408* (2020).

[45] Capraro, V., Jordan, J. J., and Rand, D. G. Heuristics guide the implementation of social preferences in one-shot prisoner's dilemma experiments. *Scientific Reports* **4**, 6790 (2014).

[46] Peysakhovich, A., Nowak, M. A., and Rand, D. G. Humans display a 'cooperative phenotype'that is domain general and temporally stable. *Nature communications* **5**(1), 1–8 (2014).

[47] Reigstad, A. G., Strømland, E. A., and Tinghög, G. Extending the cooperative phenotype: Assessing the stability of cooperation across countries. *Frontiers in psychology* **8**, 1990 (2017).

[48] Human, S. J. and Capraro, V. The effect of nudging personal and injunctive norms on the trade-off between objective equality and efficiency. *Available at https://psyarxiv.com/mx27g/* (2020).

[49] Bicchieri, C. and Xiao, E. Do the right thing: but only if others do so. *Journal of Behavioral Decision Making* **22**(2), 191–208 (2009).

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

[51] Zafar, B. An experimental investigation of why individuals conform. *European Economic Review* **55**(6), 774–798 (2011).

[52] Raihani, N. J. and McAuliffe, K. Dictator game giving: The importance of descriptive versus injunctive norms. *PloS ONE* **9**(12), e113826 (2014).

[53] d'Adda, G., Capraro, V., and Tavoni, M. Push, don't nudge: Behavioral spillovers and policy instruments. *Economics Letters* **154**, 92–95 (2017).

[54] Frey, B. S. and Meier, S. Social comparisons and pro-social behavior: Testing" conditional cooperation" in a field experiment. *American Economic Review* **94**(5), 1717–1722 (2004).

[55] Croson, R. T., Handy, F., and Shang, J. Gendered giving: The influence of social norms on the donation behavior of men and women. *International Journal of Nonprofit and Voluntary Sector Marketing* **15**(2), 199–213 (2010).

[56] Cialdini, R. B., Kallgren, C. A., and Reno, R. R. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology*, volume 24, 201–234. Elsevier (1991).

[57] Ferraro, P. J. and Price, M. K. Using nonpecuniary strategies to influence behavior: evidence from a large-scale field experiment. *Review of Economics and Statistics* **95**(1), 64–73 (2013).

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

[59] Goldstein, N. J., Cialdini, R. B., and Griskevicius, V. A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research* **35**(3), 472–482 (2008).

[60] Hallsworth, M., List, J. A., Metcalfe, R. D., and Vlaev, I. The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics* **148**, 14–31 (2017).

[61] Hausman, D. M. and Welch, B. Debate: To nudge or not to nudge. *Journal of Political Philosophy* **18**(1), 123–136 (2010).

[62] Glaeser, E. L. Paternalism and psychology. Technical report, National Bureau of Economic Research, (2005).

[63] Sunstein, C. R. *Why nudge?: The politics of libertarian paternalism*. Yale University Press, (2014).

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

[65] Dal Bó, E. and Dal Bó, P. "Do the right thing:" the effects of moral suasion on cooperation. *Journal*

of Public Economics **117**, 28–38 (2014).

[66] Bilancini, E., Boncinelli, L., Capraro, V., Celadin, T., and Di Paolo, R. "Do the right thing" for whom? an experiment on ingroup favouritism, group assortativity and moral suasion. *Judgment and Decision Making* (2020).

[67] Bénabou, R. and Tirole, J. Incentives and prosocial behavior. *American Economic Review* **96**(5), 1652–1678 (2006).

[68] Levitt, S. D. and List, J. A. What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives* **21**(2), 153–174 (2007).

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

[70] Andreoni, J. and Bernheim, B. D. Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica* **77**(5), 1607–1636 (2009).

[71] Della Vigna, S., List, J. A., and Malmendier, U. Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics* **127**(1), 1–56 (2012).

[72] Kessler, J. B. and Leider, S. Norms and contracting. *Management Science* **58**(1), 62–77 (2012).

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

[74] Kimbrough, E. and Vostroknutov, A. Injunctive norms and moral rules. Technical report, mimeo, Chapman University and Maastricht University, (2020).

[75] Kimbrough, E. O. and Vostroknutov, A. A theory of injunctive norms. Technical report, mimeo, Chapman University and Maastricht University, (2020).

[76] Dimant, E., van Kleef, G. A., and Shalvi, S. Requiem for a nudge: Framing effects in nudging honesty. *Journal of Economic Behavior and Organization* **172**, 247–266 (2020).

[77] Jevons, W. S. *The theory of political economy*. Macmillan, (1879).

[78] Bentham, J. *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Clarendon Press, (1996).

[79] Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., and Van Bavel, J. J. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences* **114**(28), 7313–7318 (2017).

[80] Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., and Van Bavel, J. J. An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental*

Psychology: General **148**(10), 1802 (2019).

[81] Brady, W. J., Crockett, M., and Van Bavel, J. J. The mad model of moral contagion: The role of motivation, attention and design in the spread of moralized content online. *Available at: https://psyarxiv.com/pz9g6/download?format=pdf* (2019).

[82] Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 79–86. Association for Computational Linguistics, (2002).

[83] Perc, M., Jordan, J. J., Rand, D. G., Wang, Z., Boccaletti, S., and Szolnoki, A. Statistical physics of human cooperation. *Phys. Rep.* **687**, 1–51 (2017).

[84] Capraro, V. and Perc, M. Grand challenges in social physics: In pursuit of moral behavior. *Front. Phys.* **6**, 107 (2018).

[85] Capraro, V., Perc, M., and Vilone, D. The evolution of lying in well-mixed populations. *Journal of the Royal Society Interface* **16**(156), 20190211 (2019).

[86] Capraro, V., Perc, M., and Vilone, D. Lying on networks: The role of structure and topology in promoting honesty. *Phys. Rev. E* **101**, 032305 (2020).

[87] Kumar, A., Capraro, V., and Perc, M. The evolution of trust and trustworthiness. *Journal of the Royal Society Interface* **17**(169), 20200491 (2020).

[88] Page, K. M., Nowak, M. A., and Sigmund, K. The spatial ultimatum game. *Proc. R. Soc. Lond. B* **267**, 2177–2182 (2000).

[89] Killingback, T. and Studer, E. Spatial ultimatum games, collaborations and the evolution of fairness. *Proc. R. Soc. Lond. B* **268**, 1797–1801 (2001).

[90] Iranzo, J., Floría, L., Moreno, Y., and Sánchez, A. Empathy emerges spontaneously in the ultimatum game: Small groups and networks. *PLoS ONE* **7**, e43781 (2011).

[91] Szolnoki, A., Perc, M., and Szabó, G. Defense mechanisms of empathetic players in the spatial ultimatum game. *Phys. Rev. Lett.* **109**, 078701 (2012).

[92] Ito, K., Ida, T., and Tanaka, M. Moral suasion and economic incentives: Field experimental evidence from energy demand. *American Economic Journal: Economic Policy* **10**(1), 240–67 (2018).

[93] Haidt, J. and Joseph, C. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus* **133**(4), 55–66 (2004).

[94] Graham, J., Haidt, J., and Nosek, B. A. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* **96**(5), 1029–1046 (2009).

[95] Haidt, J. *The righteous mind: Why good people are divided by politics and religion*. Vintage, (2012).

[96] Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., and Savulescu, J. Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review* **125**(2), 131–164 (2018).

[97] Everett, J. A. and Kahane, G. Switching tracks? Towards a multidimensional model of utilitarian psychology. *Trends in Cognitive Sciences* **24**(2), 124–134 (2020).

[98] Curry, O. S., Mullins, D. A., and Whitehouse, H. Is it good to cooperate. *Current Anthropology* **60**(1), 47–69 (2019).

[99] Rand, D., Greene, J., and Nowak, M. Spontaneous giving and calculated greed. *Nature* **489**, 427–430 (2012).

[100] Andersen, S., Gneezy, U., Kajackaite, A., and Marx, J. Allowing for reflection time does not change behavior in dictator and cheating games. *Journal of Economic Behavior & Organization* **145**, 24–33 (2018).

[101] Bereby-Meyer, Y., Hayakawa, S., Shalvi, S., Corey, J. D., Costa, A., and Keysar, B. Honesty speaks a second language. *Topics in cognitive science* , 1–12 (2018).

[102] Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., Chmura, T. G., Cornelissen, G., Døssing, F. S., Espín, A. M., et al. Registered replication report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science* **12**(3), 527–542 (2017).

[103] Capraro, V., Schulz, J., and Rand, D. G. Time pressure and honesty in a deception game. *Journal of Behavioral and Experimental Economics* **79**, 93–99 (2019).

[104] Capraro, V., Corgnet, B., Espín, A. M., and Hernán-González, R. Deliberation favours social efficiency by making people disregard their relative shares: evidence from usa and india. *Royal Society open science* **4**(2), 160605 (2017).

[105] Chen, F. and Fischbacher, U. Cognitive processes underlying distributional preferences: A response time study. *Experimental Economics* , 1–26 (2019).

[106] Chuan, A., Kessler, J. B., and Milkman, K. L. Field study of charitable giving reveals that reciprocity decays over time. *Proceedings of the National Academy of Sciences* **115**(8), 1766–1771 (2018).

[107] Everett, J. A., Ingbretsen, Z., Cushman, F., and Cikara, M. Deliberation erodes cooperative behavior—even towards competitive out-groups, even when using a control condition, and even when eliminating selection bias. *Journal of Experimental Social Psychology* **73**, 76–81 (2017).

[108] Holbein, J. B., Schafer, J. P., and Dickinson, D. L. Insufficient sleep reduces voting and other

prosocial behaviours. *Nature human behaviour* **3**(5), 492 (2019).

[109] Hallsson, B. G., Siebner, H. R., and Hulme, O. J. Fairness, fast and slow: A review of dual process models of fairness. *Neuroscience & Biobehavioral Reviews* **89**, 49–60 (2018).

[110] Rand, D. G., Brescoll, V. L., Everett, J. A., Capraro, V., and Barcelo, H. Social heuristics and social roles: Intuition favors altruism for women but not for men. *Journal of Experimental Psychology: General* **145**(4), 389 (2016).

[111] Fromell, H., Nosenzo, D., and Owens, T. Altruism, fast and slow? evidence from a meta-analysis and a new experiment. Technical report, (2020).

[112] Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., and Shalvi, S. Intuitive honesty versus dishonesty: Meta-analytic evidence. *Perspectives on Psychological Science* **14**(5), 778–796 (2019).

[113] Capraro, V. The dual-process approach to human sociality: A review. *Available at SSRN 3409146* (2019).

[114] Gneezy, U., Rockenbach, B., and Serra-Garcia, M. Measuring lying aversion. *Journal of Economic Behavior & Organization* **93**, 293–300 (2013).

Table 1: Glossary of games and unselfish behaviours

Dictator game: We measure altruistic behaviour using the dictator game. The *dictator* is given a certain amount of money and has to decide how much of it, if any, to give to the *recipient*, who starts with nothing. The recipient is passive.

Prisoner's dilemma: We measure cooperative behaviour using the prisoner's dilemma. Two players simultaneously decide whether to cooperate or to defect. Cooperating means paying a cost c to give a benefit $b > c$ to the other player; defecting means doing nothing.

Sender-Receiver game: We measure lying aversion using the sender-receiver game. The *sender* is given a private information and has to report it to the *receiver*. In some experiments the receiver is passive [23, 114], in others is active [34, 36]. Here we focus on the case in which the receiver is passive. In this case, if the sender reports the truthful information, then the sender and the receiver are paid according to Option A; if the sender reports an untruthful information, then the sender and the receiver are paid according to Option B. Only the sender knows the exact payoffs associated to the two options. Depending on these payoffs, one can classify lies into four main classes: black lies are those that benefit the sender at a cost to the receiver; altruistic white lies are those that benefit the receiver at a cost to the sender; Pareto white lies are those that benefit both the sender and the receiver; spiteful lies are those that harm both the sender and the receiver.

Trade-Off game: We measure the trade-off between equality and efficiency using the trade-off game. A decision-maker has to decide between two possible allocations of money that affect people other than the decision-maker. One decision is equal (i.e., all people involved in the interaction receive the same monetary payoff), the other decision is efficient (i.e., the sum of the monetary payoffs of all people is greater than it is in the equal allocation).

Trust game: We measure trustworthiness using the second player in the trust game. The *truster* is given a certain amount of money and has to decide how much of it, if any, to transfer to the *trustee*. The amount sent to the trustee is multiplied by a constant (usually equal to 3) and given to the trustee. The trustee decides how much of the amount s/he received to return to the truster.

Ultimatum game: We measure altruistic punishment using the second player in the ultimatum game. The *proposer* makes an offer about how to split a sum of money between him/herself and the *responder*. The responder decides whether to accept or reject the offer. If the offer is accepted, the proposer and the responder get paid according to the agreed offer; if the offer is rejected neither the proposer nor the responder get any money. Rejecting a low offer is considered to be a measure of altruistic punishment.

Table 2: Social preference models

Let x_i be the monetary payoff of player i. Social preference models assume that the utility function of player i, u_i, is defined over the monetary payoffs that are associated with the available actions. The main functional forms that have been proposed are the following.

Ledyard (1994): $u_i(x_1, \ldots, x_n) = x_i + \alpha_i \sum_{j \neq i} x_j$, where α_i is an individual parameter representing i's level of altruism. People with $\alpha_i = 0$ maximise their monetary payoff; people with $\alpha_i > 0$ are altruistic; people with $\alpha_i < 0$ are spiteful.

Levine (1998): $u_i(x_1, \ldots, x_n) = x_i + \sum_{j \neq i} \frac{\alpha_i + \lambda \alpha_j}{1+\lambda} x_j$, where α_i is an individual parameter representing i's level of altruism, whereas $\lambda \in [0, 1]$ is a parameter representing how sensitive players are to the level of altruism of the other players.

Fehr and Schmidt (1999): $u_i(x_1, \ldots, x_n) = x_i - \frac{\alpha_i}{n-1} \sum_{j \neq i} \max(x_j - x_i, 0) - \frac{\beta_i}{n-1} \sum_{j \neq i} \max(x_i - x_j, 0)$, where α_i, β_i are individual parameters representing the extent to which player i cares about disadvantageous and advantageous inequities, respectively

Bolton and Ockenfels (2000): $u_i(x_1, x_2) = \alpha_i x_i - \frac{\beta_i}{2} \left(\sigma_i - \frac{1}{2}\right)^2$, where $\sigma_i = \frac{x_i}{x_1 + x_2}$, with $\sigma_i = \frac{1}{2}$ if $x_1 + x_2 = 0$, $\alpha_i > 0$ is an individual parameter representing the extent to which player i cares about their own monetary payoff, and $\beta_i > 0$ is an individual parameter representing the extent to which player i cares about minimising the distance between their share and the fair share.

Andreoni and Miller (2002): $u_1(x_1, x_2) = (\alpha_1 x_1^{\rho_1} + (1 - \alpha_1) x_2^{\rho_1})^{1/\rho_1}$, where α_1 represents the extent to which the dictator cares about their own payoff, whereas ρ_1 takes into account a potential convexity in the preferences.

Charness and Rabin (2002): $u_2(x_1, x_2) = (\rho_2 r + \sigma_2 s)x_1 + (1 - \rho_2 r - \sigma_2 s)x_2$. Depending on the relative relationship between ρ_2 and σ_2, this utility function can cover several cases, including competitive preferences, inequity aversion preferences, and social efficiency preferences.

Table 3: The classification of norms

Behavioural scientists have long been aware of the fact that people's behaviour in a given context is influenced by what are perceived to be the norms in that context. In the same context, multiple norms might be at play. Scholars have proposed several norm classifications. In this review, we will be mainly concerned with the following three.

Schwartz [15] classified norms into two main categories, namely *personal norms* and *social norms*. Personal norms refer to internal standards about what is right and what is wrong in a given context. Social norms refer to rules and standards of behaviour that affect the choices of individuals without the force of law. Social norms are typically externally motivated.

Cialdini, Reno and Kallgren [17] focused on social norms and classified them into two main categories, namely *injunctive norms* and *descriptive norms*. Injunctive norms refer to what people think others would approve or disapprove. Descriptive norms refer to what others actually do.

Bicchieri [18] proposed a classification in three main categories, namely *personal normative beliefs*, *empirical expectations*, and *normative expectations*. Personal normative beliefs refer to personal beliefs about what should happen in a given situation. Empirical expectations refer to personal beliefs about how others would behave in a given situation. Normative expectations refer to personal beliefs about what others think one should do.

Therefore, to the extent to which people believe that what should (or should not) happen in a given situation corresponds to their internal standards about what is right (or wrong), then Bicchieri's personal normative beliefs correspond to Schwartz's personal norms. In one-shot anonymous games (where decision makers receive no information about the behaviour of other people playing in the same role), descriptive norms correspond to empirical expectations (we replace the actual behaviour of others with the beliefs). Finally, normative expectations correspond to injunctive norms. Therefore, at least for the games and decision problems considered in this review, Bicchieri's classification can be interpreted as a synthesis of the previous two classifications.

Table 4: The moral preference hypothesis

Previous work explained unselfish behaviour in one-shot, anonymous economic games using social preferences defined over monetary outcomes. According to this "social preference hypothesis", some people act unselfishly because they do not only care about their own monetary payoff, but they also care about the monetary payoffs of other people. However, especially in the last five years, numerous experiments challenged social preference models. The best way to organise these results is through the moral preference hypothesis, according to which people have preferences for following their own personal norms – what they think to be the right thing to do – beyond the monetary consequences that these actions bring about. This framework outperforms the social preference hypothesis at organising cooperation in the prisoner's dilemma, altruism in the dictator game, altruistic punishment in the ultimatum game, trustworthiness in the trust game, truth-telling in the sender-receiver game, and trade-off decisions between equality and efficiency in the trade-off game.

Table 5: Moral preference models

Let a be an action for player i. Moral preference models assume that the utility function of player i, u_i, describes a tension between the material payoff associated to a, $v_i(\pi_i(a))$, and the moral utility. The main functional forms that have been proposed are the following.

Levitt and List (2007): $u_i(a) = v_i(\pi_i(a)) + m(a)$. The moral cost or benefit associated to a, $m(a)$, is assumed to depend on whether the action is observable, on the material consequences of that action, and on the set of *social norms* and rules in place in the society where the decision maker lives.

Krupka and Weber (2013): $u_i(a) = v_i(\pi_i(a)) + \gamma_i N(a)$, where γ_i is the extent to which i cares about following the *injunctive norm* and $N(a)$ represents the extent to which society views a as socially appropriate.

Kimbrough and Vostroknutov (2020): $u_i(a) = v_i(\pi_i(a)) + \phi_i \eta(a)$, where ϕ_i is the extent to which i cares about following the *injunctive norm* and $\eta(a)$ represents the extent to which society views a as socially appropriate. (The main difference between $\eta(a)$ and $N(a)$ regards the way they are computed.)

Our proposal: $u_i(a) = v_i(\pi_i(a)) + \mu_i P_i(a)$, where μ_i represents the extent to which i cares about following their own *personal norms* and $P_i(a)$ represents the extent to which i personally thinks that a is the right thing to do.

Table 6: Outstanding challenges

- Exploring in which contexts interventions targeting personal norms are more effective at promoting one-shot unselfish behaviour than interventions targeting social norms.
- Finding the boundary conditions of interventions targeting personal norms.
- Investigating the dimension and the boundary of the "moral phenotype", to understand how different personal norms can drive different forms of unselfish behaviour and whether the moral phenotype includes behaviours that are strategically unselfish, such as strategic fairness and trust.
- Building bridges between computational linguistics, moral psychology, and behavioural economics, with the goal of understanding how to express people's utility function also in terms of the instructions of a decision problem.
- Using techniques from evolutionary game theory, applied mathematics, network science, and statistical physics to explore which types of unselfish behaviour are more likely to evolve in order to understand which personal norms are more likely to be internalised.
- Exploring the cognitive basis of personal norms using a dual-process perspective.