Alexander Vostroknutov*

Social Norms in Experimental Economics: Towards a Unified Theory of Normative Decision Making

https://doi.org/10.1515/auk-2020-0002

Abstract: Even though standard economic theory traditionally ignored any motives that may drive incentivized social decision making except for the maximization of personal consumption utility, the idea that 'preferences for fairness' (following social norms) might have an economically tangible impact appeared relatively early. I trace the evolution of these ideas from the first experiments on bargaining to the tests of the hypothesis that pro-sociality in general is driven by the desire to adhere to social norms. I show how a recent synthesis of economics approach with psychology, sociology, and evolutionary human biology can give rise to a mathematically rigorous, psychologically plausible, and falsifiable theory of social norms. Such a theory can predict which norms should emerge in each specific (social) context and is capable of organizing diverse observations in economics and other disciplines. It provides the first glimpse at how a unified theory of normative decision making might look like.

Keywords: social norms, economic experiments, behavioural economics, game theory

1 Introduction

In the past several decades social scientists have accumulated a vast body of experimental and empirical evidence that makes it hard to deny that social norms, customs, conventions, moral rules, fashions, etc. are major factors that drive human social decision making (Fehr/Schurtenberger 2018). Today, even economists start to admit that narrow self-interest cannot account for widely spread human tendencies to cooperate with others, share resources, trust, reciprocate, reward, and punish. The reason for this is not that economists have finally become convinced that these behaviours are simply too common and automatic to be driven by some intricate plans to increase personal consumption utility in the future, but

***Corresponding author: Alexander Vostroknutov,** Department of Economics (MPE), Maastricht University, The Netherlands, e-mail: a.vostroknutov@maastrichtuniversity.nl

rather the emerging realisation that the proper functioning of economies and economic growth in general are largely impossible without basic human propensity to be pro-social (Knack 2000).

The economists' reluctance to study pro-sociality may have left economics lagging behind other social sciences in studying the effects of social norms on human behaviour. Yet I strongly believe that being a late comer to study of social norms turns out an asset rather than a liability. Economics can contribute essential insights to understanding pro-sociality precisely because its general framework of rationality (or its 'bounded' versions), game theory, empirical methods of behavioural and experimental economics puts the burden of proof squarely on those who intend to explain 'pro-social' behaviour by a proliferation of propensities to behave in prosocial ways. Even the a priori methodologically contested principle that choice has to be understood as an end-result of some underlying optimization process has, I believe, proved heuristically fruitful in inducing economists to ask questions beyond the scope of other disciplines (except, of course, biology).

The precedingly invoked characteristics of an economics approach explain why economics could in recent years contribute so much to a better understanding of why people follow norms, why specifically these and not some other norms, etc. In this article I will try to demonstrate in some more detail how economics accomplished this: how an economics approach can be used to study norm-driven behaviour, how it can generate new knowledge in this field, and how the emerging economic conceptualisation of social norms can serve as a unifying framework in which the questions mentioned above can be asked in a meaningful and mathematically disciplined way. In the first part of this article (*section 2*), I will overview the evolution of economic thinking about pro-sociality in experimental economics. In the second part (*section 3*) I will focus on recent models and experiments that explicitly theorize about social norms. And in the third part (*section 4*) I will provide some ideas about how the future economic account of social norms might look like.

Before I get to this though I need to clarify two important points. First, not all social norms, customs, or traditions are suitable for 'utilitarian' economic analysis. For example, there might be no specific economic reason why we dress up a Christmas tree and not, say, a birch or an oak. The decision to go with a coniferous species probably had little to do with specific material benefits enjoyed by

the people who started this tradition.[1] Similarly, the emergence of traditional costumes or superstitious beliefs might not have any economic reason either. Thus, I would like to emphasize that when I talk about studying social norms, traditions, or conventions in economics I mean norms and rules of social conduct that have *direct material consequences* for parties involved. Such 'economic' norms guide the behaviour in redistribution, bargaining, coordination, cooperation problems, and any other types of human strategic interactions that can have tangible economic costs and benefits, or can bring different levels of 'utility'. Therefore, I will be talking about norms that guide behaviour in situations that can be represented as *game forms*, mathematical objects in non-cooperative game theory that specify players' strategies and material payoffs that are obtained by each player in each strategy profile.

Second, it is not the purpose of this article to deliberate upon pros and cons of different terminologies, definitions, and classifications pertaining to social norms that have been proposed (e.g., Sugden 2004; Bicchieri 2006). In constructing my arguments, I will rely on *functional* social-norms-related hypotheses coming from evolutionary human biology (Boyd/Richerson 1988; De Waal et al. 2006; Henrich 2015; Laland 2018). According to this view, social norms, customs, and other rules of social behaviour emerged in the co-evolutionary process of 'genes, mind, and culture' (Lumsden/Wilson 1981) as devices that *simplify* ingroup cooperation and by doing that increase the survival chances for people who follow them (as compared to completely selfish individuals who do not follow any norms). I find this view particularly useful for the analysis of economic norms because its basic concepts are similar in nature (optimization) and most importantly because it provides a *raison d'être* for the norm-driven behaviour that we observe in reality. This makes it possible not only to understand *why* some norms exist, but also to predict *which* norms can emerge under some new circumstances.[2] My focus on the evolutionary origins of norms instead of their classification does not mean however that I will ignore the mechanisms through which norms are maintained. To give an example, washing your hands is a good way to prevent the spread of diseases. Given its benefits to the community, we can expect that such a norm can become common. However, this norm can be *descriptive* (I wash my hands because others

1 Of course, once the Christmas tradition has been established, one can study for example the elasticity of demand for Christmas trees and other economic consequences of having this tradition. However, this is different from understanding why we use a coniferous species.

2 For instance, Demsetz 1974 gives gives an example how expanding fur trade in America led to the emergence of property rights on hunting lands, because of the pressure on the population of animals that this expansion created. In this case new norms (property rights) have evolved for purely economic reasons.

do) or *injunctive* (I wash my hands because not doing so will harm the community). Therefore, even though the norm has a clear benefit, the way that it is supported (through public shaming in case of descriptive norm or internalised guilt in case of injunctive norm) can have important economic policy implications if one wanted to promote this norm among a specific group of people, who can be more responsive to 'descriptive' or 'injunctive' incentives (e.g., among children or adults). I will come back to my definitions of injunctive and descriptive norms in *section 4*.

2 Evidence of Normative Decision Making

2.1 Social Preferences in the Dictator and Ultimatum Games

It is interesting to note that explanations of social behaviour generally related to some form of following social norms had been proposed in experimental economics since its inception. Specifically, such hypotheses had emerged once it was noticed that the behaviour of experimental subjects in bargaining games deviates significantly from the predictions of standard game theory where it is assumed that players are exclusively driven by the maximization of their own material payoffs (I will call such players 'selfish'). To my knowledge, the study of 'ultimatum bargaining' by Güth et al. (1982) is the earliest paper that can be firmly categorised (ex post) as experimental economics and where the authors explain the observed deviations from selfishness by appealing to the concept of 'fair' allocation of resources and the concept of 'punishment' of unfair offers. Even more to the point, Hoffman and Spitzer (1985) used bargaining games to explicitly test three theories of 'distributive justice' way ahead of their time.

Even though the authors of these studies have pioneered the field, the game forms that they used are rather complex from the contemporary point of view, given our current understanding of a wide variety of norm-related incentives that people may respond to. Therefore, for expositional purposes it is easier to start with a much simpler game form analysed by Forsythe et al. (1994), further FHSS. This is the study where the well-known Dictator game was first used in an economic experiment.[3] In the Dictator game, one player (a dictator) is provided with some amount of money by the experimenters (e.g., $4). The task of the dictator is to propose a division of this amount between herself and another player (a re-

3 Some versions of the Dictator game were used in earlier, but mostly psychological experiments (e.g., Kahneman et al. 1986).

ceiver). Once the dictator has proposed a division, it is implemented and the game ends (for example, the dictator can keep $3 and give the receiver $1). The receiver in this game is passive and does not make any choices, which is crucial since this guarantees that the choice of the dictator is not influenced by any possible 'retributive' motives of the receiver. Therefore, the Dictator game was and still is considered a good instrument to elicit what will later be coined 'social preferences'. Indeed, since the choice of the dictator is anonymous and subjects just receive the proposed division of money without any additional material consequences, a choice of the dictator to give some amount to the receiver should reveal her 'taste for fairness' (in the parlance of FHSS), which is not obfuscated by other possible motives (e.g., reputation concerns or a possibility of retribution after the experiment). It is also not inconceivable that subjects may find themselves playing something very similar to the Dictator game in reality. Volunteering your time to work at an NGO or donating money to charity are almost perfect examples of dictator giving.[4]

FHSS find a typical pattern of dictator giving that has been replicated in hundreds of other studies since then (Engel 2011). Around 20% of dictators share the money equally (each player gets half of the 'pie'), around 30% keep all the money leaving the receiver with nothing, the rest 50% choose something in between. FHSS have also run a separate version of the same Dictator game, only without paying subjects anything (they divide hypothetical amounts of money). In this case, around 50% of dictators choose equal split and only 10% keep all the hypothetical money for themselves. FHSS did not propose any formal model that explains the behaviour in the Dictator game, however there are two important observations that we can make in view of the experimental results. First, we can see that there is a considerable *heterogeneity* in 'tastes for fairness' in the population ranging from subjects who seem to be completely selfish (give nothing to the receiver) to very fair subjects who share $4 equally (and also subjects with intermediate taste for fairness who give less than half but more than zero). Second, we can observe that when giving is costly many subjects choose to give less than when it is free. This tells us that there is a *trade-off* between monetary and 'fairness' incentives, which implies that subjects try to balance personal material payoff and how 'fair' their choice is.

Once the fact that many people are ready to trade-off personal consumption utility for fair allocation of resources has been firmly established, the next generation of researchers took on a challenge to develop a formal mathematical model

4 Franzen/Pointner 2013 show that the behaviour in the Dictator game does have external validity.

that would account for this behaviour in the Dictator and other experimental game forms. The main issue here was to understand which allocations are considered 'fair' or 'unfair' and how to build a model that explicitly incorporates the money/fairness trade-off and the observed heterogeneity in tastes for fairness. At that time economists already had a well-established tradition of modelling trade-offs between consuming different goods with utility functions. Given some budget and two goods (e.g., apples and bananas) a utility function that takes into account the amounts of both goods can be written that would produce some optimal choice of the amounts of apples and bananas that depends on the parameters of the utility function. For example, given a budget of \$10, some people optimally choose to buy 1 apple and 5 bananas, some others buy 5 apples and 1 banana, yet others buy only apples, etc. Therefore, the utility function automatically trades-off apples for bananas, while its parameters determine the optimal ratio of apples to bananas that can be different across individuals (heterogeneity). The same principle was applied to the money/fairness trade-off. Specifically, it was proposed that people are not just selfish (maximize only their own material payoff) but that their utility depends also on the amount of money received by the other player.

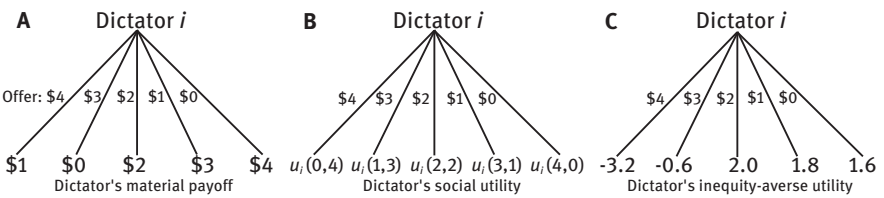

Fig. 1: *A.* The Dictator game form: a dictator is choosing how much out of \$4 to offer to the receiver. If the dictator maximizes material payoff, she chooses to offer \$0; *B.* The Dictator game: instead of material payoffs, the dictator might be maximizing social utility that depends on the amounts of money allocated to both players; *C.* The Dictator game where the dictator has inequity-averse utility function with $\alpha_i = 0.8$ and $\beta_i = 0.6$. The optimal choice is to offer \$2.

To illustrate, *figure 1A* shows the Dictator game form with the material payoffs of the dictator defined by the amounts of money she can receive. This game form represents the choice faced by a selfish dictator, who of course will maximize own material payoff and choose to give the receiver nothing. In Figure 1B the Dictator game form is turned into a *game* where the monetary payoffs are transformed by an *outcome-based social utility function* $u_i(x, y)$, or 'social utility', which determines the utility that a dictator *i* 'actually' receives if she cares about some properties of the material payoff distribution among the players in a given outcome.

Here x is the material payoff of the dictator and y is the material payoff of the receiver. The preferences represented by such a utility function that depend on the distribution of income in each outcome are called *social preferences*. Notice that when I said earlier that I believe that the concept of optimization of utility can be very useful to model social behaviour, I meant models of this kind, where players act *as if* they maximize *some* utility function, so that all the powerful machinery of rational choice theory can be used as before, it is just that this utility can depend in principle on any aspect of the strategic situation.

The task of the researchers, who first started to think about social behaviour in this way, was to determine the shape of $u_i(x, y)$ that would best fit the observed behaviour in the Dictator and other game forms popular at the time. Several models were proposed (e.g., Fehr/Schmidt 1999; Bolton/Ockenfels 2000). The most popular one is the *inequity aversion* utility proposed by Fehr and Schmidt (1999). The idea of these authors was that players care about their material payoffs, however, they also dislike unequal material payoffs. The specification that Fehr and Schmidt (1999) proposed is given by

$$u_i(x, y) = x - \alpha_i \max\{y - x, 0\} - \beta_i \max\{x - y, 0\}. \quad (1)$$

According to this social utility, player i receives the material utility from her money, x, minus some utility proportional to the distance between her (x) and other player's material payoff (y). If the other player gets more money, or $y > x$, we get $u_i(x, y) = x - \alpha_i(y - x)$. If $y < x$ then $u_i(x, y) = x - \beta_i(x - y)$. The individual parameters α_i and β_i determine the "taste for fairness" of player i. For example, if $\alpha_i = \beta_i = 0$ we get a standard selfish player. As the values of these parameters grow, player i starts to care more and more about inequality as compared to personal material payoff. The idea that the parameters differ across individuals represents the heterogeneity observed in the experiments. It is also assumed that $\alpha_i \geq \beta_i \geq 0$, or that player i dislikes disadvantageous inequality more than she dislikes advantageous inequality. *Figure 1C* shows the Dictator game with an inequity-averse dictator i with $\alpha_i = 0.8$ and $\beta_i = 0.6$. Notice that the equal split (both players get \$2) has the highest social utility. So, a dictator with such social preferences would optimally choose it. As the values of α_i and β_i decrease the optimal choice moves in the direction of selfish action 'give \$0 to the receiver'. This accounts for the heterogeneity of behaviour observed in the experiment of Forsythe et al. (1994).[5]

Overall, it was found that inequity aversion provides a potential explanation of the behaviour in the Dictator game if we assume that there is enough hetero-

5 To achieve this last effect we technically need to assume concave utility of money, or replace the material utility x in equation (1) with $f(x)$, where f is some concave function.

geneity in the parameters α_i, β_i, and in the curvature of the utility of money in the population. More importantly, it also can describe the behaviour in the Ultimatum game form studied by Güth et al. (1982) and also FHSS. The Ultimatum game form is the same as the Dictator game form except that after the allocating actor (now called proposer) proposes a division of \$4, the receiver (who is now called responder) can accept this division, in which case it is implemented, or reject it, in which case both players receive nothing. Notice that if both players receive nothing the inequity-averse utility of the *responder* is zero, which is higher than his utility from accepting low amounts (divisions (0, 4) and (1, 3), first material payoff to the responder). If the responder is also inequity-averse, then his utility from accepting the division is the same as that of the proposer only with the 'money axis' reversed. So, a responder with $\alpha_i = 0.8$ and $\beta_i = 0.6$ gets utility $-3.2 < 0$ from the allocation (0, 4) and $-0.6 < 0$ from (1, 3), see *figure 1C*. This means that such a responder will reject low offers of 0 and 1 and will prefer both players to get nothing instead. Many studies (see Oosterbeek et al. 2004, for meta-analysis) found that the observed behaviour of the responders in the Ultimatum game does actually follow such a 'rejection threshold' pattern: they reject low offers below some threshold and accept high ones. Most importantly, it was observed that the proposers in the Ultimatum game make much more equal offers than in the Dictator game (Forsythe et al. 1994). According to the inequity aversion model this happens out of fear of rejection if we assume that inequity-averse preferences of the players are common knowledge. Inequity aversion therefore can be counted as the first model of *normative strategic behaviour* that succeeded at accounting for actual human choices in the Dictator and Ultimatum games. It is very important to emphasize here, that this explanation relies on *both* social preferences *and* strategic thinking inherent to game theoretic models. Thus, it suggests that people not only maximize social utility, but also strategically take into account the optimization of social utility by others (proposers in the Ultimatum game do not make very low offers because they expect them to be rejected). This was an important achievement of game theory as a predictive model of human behaviour.

2.2 Problems with Inequity Aversion

Given the initial success of theories of inequity aversion at describing choices in the Dictator and Ultimatum games, the next wave of researchers started to run more experiments that were designed to test the model's ability to account for behaviour in a large variety of other games. Unfortunately, theories based on inequity aversion did not fare too well. In an experiment by Engelmann and Strobel (2004), subjects acting as dictators were choosing a material payoff allocation for

three players. For example, one of the choices they made was between allocations (21, 12, 3) and (13, 12, 5), where the middle material payoff is for the decision maker (measured in experimental monetary units). Notice that here the material payoffs for the three players do not sum up to a constant as was the case in Dictator and Ultimatum game forms, which were framed as divisions of some fixed sum of money. Engelmann and Strobel (2004) found that inequity aversion explained only a very small percentage of choices in their experiment, and that instead *preference for material-payoff efficiency* or *preference for maximin* worked best at reconciling the data. The preference for material-payoff efficiency can be expressed by the following social utility function, defined over the material payoffs of three players:

$$u_i(x, y, z) = x + \gamma_i(x + y + z). \tag{2}$$

A decision maker with preferences represented by this type of utility enjoys her material payoff x, but at the same time likes situations where all players together get more money, which is expressed by the second term ($\gamma_i \geq 0$ is an individual parameter). Such a player might choose a more unequal allocation over a less unequal one, if it provides a higher sum of material payoffs to all players.

The preference for maximin can be expressed as

$$u_i(x, y, z) = x + \delta_i \min\{x, y, z\}. \tag{3}$$

Here, a decision maker ranks allocations according to how she and the relatively worst-off player fares. This type of social preference is different from both inequity aversion and material-payoff efficiency, since technically it focuses only on the material payoff of the least well-off player.[6]

The findings of Engelmann and Strobel (2004) were interesting for several reasons. First, their experiment demonstrated that inequity aversion is not the only 'social preference' out there, and that there are other possible normative criteria that people might choose to follow. The second and the more important problem was that in a world with diverse social preferences we—and here I mean both humans and social scientists—need to know in which situations which social preference applies. However, the social utility functions mentioned above obviously do not specify under which conditions they are most likely to produce a good fit. Therefore, by finding other possible social preferences that explain behaviour, Engelmann and Strobel (2004) uncovered a new deeper problem, namely that of finding a mapping between each possible choice situation and the type of social

6 Baader/Vostroknutov 2017, who replicated the results of Engelmann/Strobel 2004, find that students who study economics are mostly prone to maximize material-payoff efficiency, whereas many students from humanities followed maximin.

preference that is most likely to apply to it. This problem was not easy to solve. For example, the experiment by Galeotti et al. (2018) shows that a slight change in material payoffs can bring a radical shift from choice being seemingly guided by inequity aversion to one apparently driven by material-payoff efficiency. At this point, it was unclear what exactly can bring about such a dramatic shift in preferences.

2.3 Problems with Consequentialist Preferences

In principle, it is not inconceivable that people may use different normative criteria in different strategic situations or that some more complex social utility function $u_i(x, y)$ that amalgamates inequity aversion and preference for efficiency/maximin can account for behaviour in a wider variety of games. Several attempts along these lines have been made (e.g., Charness/Rabin 2002). However, a simple experiment by McCabe et al. (2003) has brought up a more serious problem that has left this research strategy hanging in the air. These authors considered two game forms shown in *figure 2*.

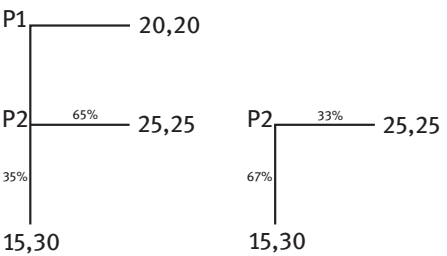

Fig. 2: The two game forms studied by McCabe et al. (2003); the payoffs are expressed in experimental monetary units that are exchanged for real money after the experiment

In the left game form (a mini-Trust game) player P1 first decides whether to choose an allocation (20, 20) and finish the game or pass the move to player P2, who in her turn can choose between allocations (25, 25) and (15, 30), where the first monetary payoff goes to P1 and the second to P2. The right game form is the same as the left except that P1 does not choose anything and only P2 decides which allocation shall be implemented. Therefore, comparing choices of P2's in the two game forms can tell us if they are influenced by the presence of the original move of P1 or not. Notice that if subjects are driven exclusively by *some outcome-based (consequentialist) preference* of a type considered above, where the social utility

of an outcome depends *only* on the material payoffs received by players in this outcome, then the distribution of P2's choices should not change since the allocations between which P2 is choosing are the same in both game forms. Thus, inequity aversion, preference for efficiency or maximin, or any other social preference for that matter, will make the same prediction in the two game forms irrespective of the presence or absence of P1's move.

The percentages shown on the actions of P2 in *figure 2* reflect the number of subjects who chose (25, 25) and (15, 30) in the two game forms (statistically significantly different). It is clear that the percentage of P2's who go for the pro-social choice (25, 25), which is both more equal and more material-payoff efficient than (15, 30), is much higher when P2's choose after P1 (the left game form) as compared to when they choose first (the right game form).[7] This difference immediately refutes consequentialist models of social behaviour where social utility of a given outcome is a function exclusively of the material payoffs received by players in this outcome.

These results have demonstrated that no model of social preferences of the type considered above can explain this and other similar shifts in choices. Therefore, McCabe et al. (2003) proposed that *intentions* of P1 matter for the choice made by P2. Specifically, by not choosing (20, 20), P1 signals to P2 that he trusts her to make the choice (25, 25) and many P2's flattered by such kind treatment reciprocate and go for (25, 25) instead of following their selfish urges by choosing (15, 30). The same mechanism is not present in the right game form since P1 cannot signal anything, which explains why fewer P2's choose (25, 25).

These findings have sparked interest in reciprocal behaviour that goes beyond usual social preferences. Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), and Cox et al. (2007) proposed general models of reciprocity in games that were using arguments similar to that of McCabe et al. (2003). For example, in Falk and Fischbacher (2006) a second mover evaluates how 'kind' the action of the first mover was by comparing the possible material payoffs that she can obtain as a result of this action to the material payoffs that she could have obtained had the first mover chosen differently (a kind action brings the second mover more material payoff than an unkind action). After that, the second mover 'matches' this kindness by choosing how kind her action towards the first mover should be.

The reciprocity models can in principle account for the behaviour in McCabe et al. (2003) and similar experiments. However, they do not resolve the multiplic-

7 This effect has been replicated many times (e.g., Goeree/Holt 2001; Charness/Rabin 2002; Cox/Deck 2005).

ity of social preferences problem mentioned in the previous section, since we observe different social preferences in single-move game forms where reciprocity has no bite (like the Dictator game or game forms used in Engelmann and Strobel (2004)). Another problem with reciprocity models arises from their conceptualisation of kindness. Isoni and Sugden (2018) point out that it is not very clear why the choice of P1 to pass the move to P2 in the left game form in *figure 2* should be considered kind when all that P1 might really want is to obtain the highest possible material payoff from having P2 choose (25, 25). In other words, a deeply selfish but cunning P1 might pretend to be nice by passing the move to P2 who then chooses (25, 25) and thus helps P1 to reach his ultimately selfish goals. If this is the case, then P2 should not really think that P1 is demonstrating kindness by passing the move, and thus should not reciprocate. From an argument similar to this one, Isoni and Sugden (2018) conclude that in order for reciprocity to work properly some 'joint action' by both players is necessary, or that the choice of P1 to pass the move to P2 should be considered a part of some joint plan to reach (25, 25).[8]

2.4 Problems with Context

Incorporating reciprocity along with the apparent multiplicity of social preferences leads to increasing context-dependence—and decreasing empirical testability—of economic models based on these approaches. This problem is further compounded by another type of behavioural 'irregularity' that shows itself in Dictator game experiments and thus on the original home-turf of the social preference approach. List (2007) and Bardsley (2008) considered 'give-take' Dictator game forms in which the dictator can choose to either *give* money to the receiver, as in the standard Dictator game, or to *take* money from her. In the experiment by List (2007) both dictators and receivers had initial endowments, and a dictator could either give the receiver up to $5 or take up to $5. In this game a large majority of subjects chose to give or take nothing, which was different from the standard Dictator game where many subjects chose to give $2.5 (another experimental treatment in List 2007).

This result cannot be accounted for by any 'regular' social preference. Suppose that you are inequity-averse and you choose to give $2.5 in the standard

8 This observation connects the reciprocity problem described here with the literature on collective intentionality or we-intentions (e.g., Tuomela/Miller 1985; Sugden 1993). Interestingly, in case of reciprocity common social norms can be regarded as the source of such collective intentionality (see *section 3*).

Dictator game. This means that you would never choose to take money from the receiver, should such an opportunity arise, simply because taking money would make the allocation even more unequal than it already was if you kept all the money (which you optimally did not do). Material-payoff efficiency (2) does not predict anything interesting in this context because the sum of material payoffs is constant in all outcomes, so it just turns into selfishness. This means that no efficiency maximizing subject would choose $2.5 in the standard Dictator game. Maximin also predicts that if someone has chosen to give $2.5 in the standard Dictator game, then this individual should not take money from the receiver given the opportunity, because this makes the minimal material payoff even lower than in the allocation where nothing is given or taken. Therefore, a simple addition of seemingly irrelevant actions to the choice set has a dramatic influence on behaviour without any good explanation. These experiments emphasized that the *context* of the choice matters for normative decision making, alas in a not very obvious way.

2.5 Problems with Social Context

Finally, all the problems with understanding social behaviour described above were detected in standard laboratory settings where subjects are co-equal strangers of similar social standing, where roles of the players are chosen randomly (no role entitlements), and where money is windfall, in the sense that subjects do not have any specific ownership claims to it prior to the experiment. It is not surprising therefore that many researchers aimed at testing other possibilities. What happens when subjects feel entitled to the role of the dictator or the money that they have to divide? What happens if subjects believe that they play with someone from the outgroup instead of the ingroup?

In a series of papers, Elizabeth Hoffman and co-authors (Hoffman et al. 1994; 1996; 2000) investigated some of these questions by modulating the perceived entitlement to the role of the dictator and the anonymity of the decisions. They corroborated that a higher degree of anonymity makes people more selfish, and that perceived entitlements to the role of a dictator induce players to offer less money. This was later complemented by studies where dictators or receivers earned the money to be divided in the Dictator game (e.g., List 2007; Oxoby/Spraggon 2008), as well as by studies that changed the perception of in/outgroup among subjects (e.g., Chen/Li 2009). Overall, these experiments supported 'folk' intuitions associated with these phenomena: subjects were less willing to share money that they considered their own (or take money from a receiver who earned it) and subjects behaved more selfishly towards outgroup than towards ingroup. Finally, experiments on social learning (e.g., Bicchieri/Xiao 2009; Panizza et al. 2020)

demonstrated that observing others share money in some way made subjects more likely to do the same. Of course, none of these effects created by *social context* were explicitly modeled in social preference or reciprocity models (though, see Akerlof/Kranton 2000).

3 Social Norms in Experimental Economics

The previous section sketched the trajectory of experimental research up to the emergence of the new literature that explicitly emphasizes the role of social norms in economic decision making. It is important to note here that this new thinking started to gain track exactly because neither social preferences nor models of reciprocity could provide good answers to how decisions in a specific strategic situation should be modeled. First, it was unclear which social preference out of many is applicable in a given game form. Second, reciprocity models often produced strange results like mixed Nash equilibria and had other conceptual problems. And third, the knowledge that context and social context matter rendered the applicability of all these models precarious. The 'social norms paradigm' is gaining more attention in economics because it promises, at least in principle, to resolve all these issues. Indeed, we know that social norms are context-dependent, that they incorporate social context (ownership, entitlements, in/outgroup), and that people have different propensities to follow them. All this together suggests that a single framework, where social norms are considered to be the main driving force of pro-sociality, can encompass all the behavioural phenomena mentioned above.

One of the first studies that followed this path was Kessler and Leider (2012), where a *norm-dependent utility function* has been proposed as a conceptually new device that models the human tendency to follow social norms.[9] Specifically, the authors assumed that subjects maximize the following utility function (in the Dictator game):

$$u_i(x) = x - \phi_i|\hat{x} - x|. \tag{4}$$

Here x is dictator i's material payoff; $\hat{x}$ is the *norm*, or the amount of money kept ($4 minus the offer to the receiver in terms of *figure 1*) that is considered the *socially most appropriate*; $\phi_i \geq 0$ is the individual parameter that measures i's propensity to follow norms ($\phi_i = 0$ gives us the standard selfish decision maker); the distance between the norm and the actual amount kept $|\hat{x} - x|$ is thought to measure the

9 Earlier studies by Cappelen et al. 2007 and López-Pérez 2008 proposed similar formulations.

disutility of deviating from the norm.[10] Two things should be noted here. First, this is *not* a social preference model. A player with norm-dependent utility does not really care what material payoff the other player gets (at least not directly), he only cares about his own material payoff x and the extent to which his action conforms to the norm $\hat{x}$. So, if for some reason the norm is to keep all the money ($\hat{x}$ = \$4; for example if the dictator owns it), then even the most norm-following dictator will keep all the money, as the norm prescribes. From the perspective of social preferences this would look like selfish behaviour. At the same time, if we consider the usual Dictator game with windfall money, then the norm might be to divide money equally ($\hat{x}$ = \$2), in which case the same dictator—who kept all the money in the Dictator game with ownership—will share it equally. This behaviour would look like inequity aversion from the perspective of social preferences. From the point of view of social preferences the different choices of a norm-following dictator in the two versions of the Dictator game look inconsistent: she acts selfishly in one context, but pro-socially in the other. However, if we believe that the dictator is adhering to social norms, then her behaviour is consistent since the norms in the two social contexts are different and the dictator chooses as the norm prescribes.

It should be reiterated that social preferences models do not use information concerning the different nature of entitlement in the two Dictator games (which is why the behaviour looks inconsistent from this perspective). We could modify the social preferences framework casuistically, e.g. by declaring that people act selfishly whenever they play with their own money and only exhibit inequity aversion when dealing with windfall money. Following this logic, in principle we could specify which social preferences people are using in each specific game form and context by just enumerating all possibilities. However, this boils down to stating that social preferences may be different in each context, without generating testable hypotheses concerning how pro-social behaviour differs contingent on context. Conversely, the social norms paradigm does provide us with testable hypotheses: if we know what the norms are in some game forms or contexts, then we should expect that norm-following individuals with higher values of ϕ_i will follow the norms in *more* contexts, and norm-breaking individuals (ϕ_i close to zero) in fewer contexts.

This brings me to the second—in a sense—more fundamental point. The norm-dependent utility model in Kessler and Leider (2012) does not explicitly specify how $\hat{x}$ comes about. It is assumed that it is known for the game form in

10 Kessler/Leider 2012 consider a more general norm-dependent utility where the disutility of deviation from the norm can be any increasing function of $|\hat{x} - x|$. I do not consider this formulation here for expositional purposes.

question. This can be regarded as a weakness of this theory. However, I believe that the important achievement of this study was the mathematical conceptualisation of the norm-dependent utility per se, and the demonstration that this type of utility specification can account for observations that previous models could not. Even though the question of how to determine the norm is left open, the paper showed how norm-dependent utility can be conceptualized for purposes of proper experimental economic enquiry.

The problem of indicating general empirical ways and means of specifying norm-dependent utility—namely the lack of clarity with regard to what the norm actually is in a given context—has been addressed by Krupka and Weber (2013), further KW, probably the most prominent paper in this literature. KW did not come up with a theoretical way to determine the norm $\hat{x}$, but instead proposed an experimental task that would allow to empirically 'measure' norms in any game context. Their idea was exactly as I argued above that it is possible to test the predictive power of theories relying on norm-dependent utility, if we know a way to measure the norm relevant in a given context for which the theory implies behavioral predictions.

KW proposed a *norm elicitation task* in which subjects rate *social appropriateness* of different actions in a game form. For example, for a Dictator game the task is formulated as follows. Consider the choice of some dictator to offer 30% of the money to the receiver and to keep 70%. How socially appropriate do you find this choice? (Rated on a 4-item Likert-scale from 'very inappropriate' to 'very appropriate'.) But KW went beyond this rather conventional way of letting subjects rate the social appropriateness of actions. Subjects were going to win a prize (e.g., $10) if they rate social appropriateness coincident with the ratings provided by the majority of other subjects in the session. A subject intending to win $10 is thereby induced to report the level of social appropriateness that she believes the majority will indicate, instead of reporting her own personal opinion about the appropriateness of such action if the task was not incentivized (or incentivized differently). By using this method, we can elicit beliefs concerning social appropriateness of actions in the Dictator game (or any other game form), take averages across subjects, and thereby determine which action is considered the norm: it should be the action that is rated as the most appropriate on average.

In the Dictator game, this elicitation task produces the estimates of social appropriateness shown in *figure 3*. Notice that equal split is considered the most appropriate action on average. This yields the value of $\hat{x}$, and allows us to test whether norm-dependent utility captures the behaviour in the Dictator game. In their analysis, KW estimated the parameters of the norm-dependent utility function similar to (4) and came to the conclusion that the actual amounts offered (by a different group of dictators who did not participate in the norm elicitation task)

are correlated with average social appropriateness (as measured by the task). This result constituted the first direct experimental confirmation that beliefs about social appropriateness of actions (normative beliefs) exert an influence on pro-social behaviour.

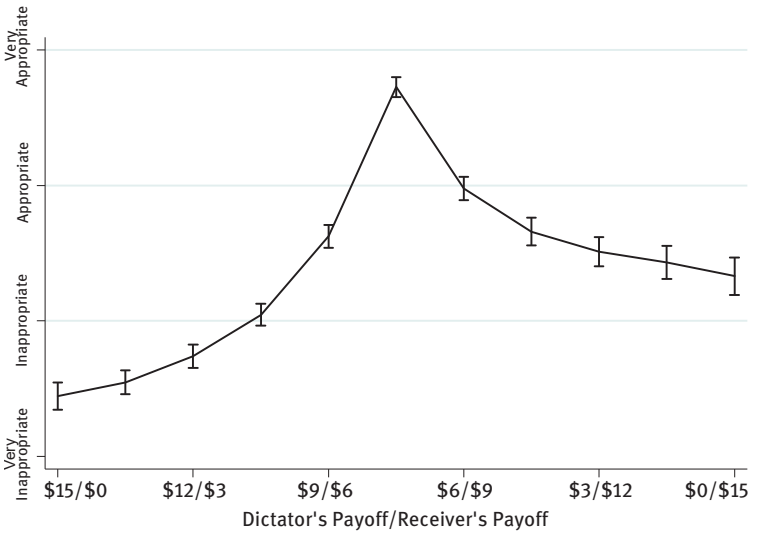

Fig. 3: Average normative valences elicited in the norm elicitation task by Krupka and Weber (2013). The data shown are from Kimbrough/Vostroknutov (2018). The error bars are ±1 SE.

In search for further evidence concerning how social norms drive pro-social behaviour, Kimbrough and Vostroknutov (2016)—subsequently KV or 'we'—made an additional distinction. While KW merely collected information about what subjects believe to be prevailing attitudes and expectations of others, we explicitly took into account the decision of complying or deviating from these expectations. After all, forming the belief that something is expected to be done and actually deciding on doing or not doing it are categorically distinct. KW pinned down $\hat{x}$ in order to show that pro-sociality is driven by norms, we decided to estimate ϕ_i in a norm-dependent utility specification as in (4). As I mentioned above, the main kind of falsifiable predictions that models of normative preferences can offer is that norm-following individuals should be more inclined to choose as the norm prescribes, whereas norm-breaking individuals should tend to choose in more selfish ways. Yet, this in itself is the result of two separate factors: 1) the beliefs concerning what is expected and 2) the disposition to actually comply with what is

expected.[11] Accordingly, we formulated our task as that of empirically measuring individual proxies for ϕ_i and to test predictions that arise from their interaction with normative expectations in several contexts: the Dictator, Ultimatum, Trust (Berg et al. 1995), and Public Goods (Isaac et al. 1994) game forms. Under our hypothesis, we expected to observe that individuals with high estimate of ϕ_i behave more according to the norms of social appropriateness (as elicited in line with KW) in all these game forms, while individuals with low ϕ_i behave more selfishly.

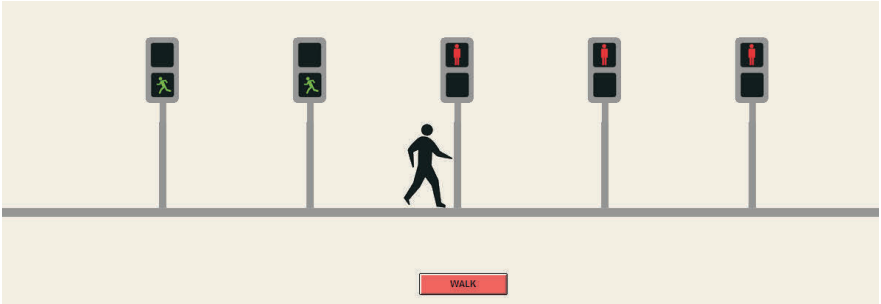

Fig. 4: The rule-following task in Kimbrough and Vostroknutov (2016)

In designing a task (instrument) for measuring ϕ_i, our intuition was that general individual proclivities to act compliant with expectations of what is deemed pro-social in particular contexts would express themselves even when material consequences for other individuals are absent. Therefore, we proposed an individual *rule-following task* that measures how much money subjects are willing to forgo if following a costly rule defined by the experimenter that has only self- and no other-regarding consequences. In the task a figurine walks as an avatar of the participant across the screen and stops in front of each of several red traffic lights (see *figure 4*). Subjects choose whether to walk on at a red light or to wait until the light turns green. The choice to wait is costly, each second that subjects wait decreases their earnings in the task by 8 cents. Moreover, the instructions explic-

11 Strictly speaking, this hypothesis can only hold in simple strategic situations where it is reasonable to assume that there is not much heterogeneity in normative expectations that can influence the behaviour of norm-following individuals. The problem of heterogeneity in normative expectations can be dealt with if we use the KW task to elicit these expectations for the subjects who are also making choices in the game forms of interest (e.g., the Dictator game). This way individual normative expectations can be directly connected to the behaviour of the individual holding them. Thomsson/Vostroknutov 2017, Panizza et al. 2018, and Merguei et al. 2020 show how this technique can be used in experiments.

itly state that '*The rule is to wait for each light to turn green*'. Subjects are facing a trade-off between following the rule and earning more money. We hypothesised that subjects who follow this artificial rule and voluntarily lose money by doing so (high ϕ_i) should also comply with social norms that express particular other-regarding expectations. Subjects who do not follow this rule (low ϕ_i) and earn more money as a result behave more selfishly vis-à-vis norms of social appropriateness.

Using the number of seconds that a subject chooses to wait at all five traffic lights as a proxy for the value of ϕ_i, we found that the number of seconds spent waiting at the traffic lights indeed correlated significantly with how close the choices of our subjects were to the norm of equal split in the Dictator game (we determined the norm by means of the KW's norm elicitation task). Specifically, subjects who followed the rule and waited at all traffic lights (rule-followers) were more likely to choose equal split, than subjects who did not wait at the lights (rule-breakers). These subjects were mostly going for the selfish option of keeping all the money. Subjects with intermediate estimates of ϕ_i chose offers between \$0 and equal split. This result based on 67 observations—which guarantees that it is not a random coincidence—provided evidence, complementary to the findings of KW, that pro-social choices in the Dictator game are related to the propensity to follow rules (or norms), but differentiated between factors in action and belief or opinion space.

Our experiments have established similar results for the other social dilemmas that we studied. For example, in the Ultimatum game we found that rule-followers have significantly higher rejection thresholds than rule-breakers (69 observations). In the social norms framework, this means that rule-followers were more willing to *punish* the proposers for not making an equal-split offer than rule-breakers. This is consistent with the idea that people who are more prone to follow norms are also more prone to costly punish norm violators, which follows naturally from the evolutionary account of norms that I advocated earlier.[12] Similarly, we found that rule-followers reciprocate significantly more in the Trust game (96 observations), and that groups of assortatively matched rule-followers are able to sustain cooperation in the repeated Public Goods game, as compared to the groups of rule-breakers and mixed groups that failed to do so (72 observations).[13] This last result is especially remarkable, since we know from many Public Goods

12 Cooperative norms need to be supported by various mechanisms that prevent norm violations: punishment, rewards, reputation, etc. (Henrich 2015).

13 The way Public Goods game (or voluntary contribution mechanism, VCM) works is explained in Kliemt 2020.

experiments that cooperation decays in almost all conceivable conditions (Zelmer 2003).[14]

At this point it is reasonable to question the motives of subjects who wait at the traffic lights in our task. It can be argued that there are non-normative reasons why people may do that. For example, they might wait (or pass through) out of habit because this is what they normally do. In order to eliminate any specific effects of the traffic-light framing we have developed and tested a new task that does not evoke any associations with past experiences (Kimbrough/Vostroknutov 2018). In this task subjects should individually allocate one hundred balls into two baskets, yellow and blue. For each ball put in the yellow basket a subject receives 10 cents and for each ball put in the blue basket 5 cents is received. Subjects are told in the instructions that *'The rule is to put all balls into the blue basket'*. Thus, following the rule decreases subjects' material payoffs because they get twice as much money from putting the balls into the yellow basket. Our results show that the correlation between the choices in this task and dictator giving is even stronger than in Kimbrough and Vostroknutov (2016) (based on 180 observations in three countries). Therefore, we can rule out the effects of the traffic-light framing on our results.

Nevertheless, other possible confounds remain. For example, it can be argued that subjects conform to the so-called experimenter demand (Zizzo 2010): they do what they think the experimenter expects of them (wait at the lights or put the balls into the blue basket). Under this view, the correlation between the performance in a rule-following task and dictator giving arises because subjects presumably follow experimenter demand in the Dictator game too. This may as well be. However, estimates from Fleming and Zizzo (2015) and Panizza et al. (2020) show that only around 20% of subjects are susceptible to the experimenter demand effect. Moreover, such conformity or 'obedience' in our rule-following tasks cannot be the result of fear of repercussions since subjects understand perfectly well that their choices are anonymous and nothing will happen to them if they 'disobey' (except for earning more money as described in the instructions). At the same time, compliance with authority can also be normative in nature. Therefore, even if experimenter demand plays a role in our results, it does not necessarily contradict our hypothesis that the propensity to follow norms is driving this behaviour.

14 It should be noted that we do not equate high propensity to follow rules or norms with prosociality per se, but rather with adherence to existing social norms. This means that in a society where 'bad' anti-social norms are prevalent, people who wait at traffic lights might demonstrate higher degrees of anti-social behaviour. This conjecture is yet to be tested.

I believe that there are at least two main normative motives that are responsible for the correlation between the choices in the rule-following tasks and dictator giving. First, some people, who are prone to conform with the actions of others, might believe that the majority is choosing to wait at the lights and also to divide the money equally in the Dictator game. So, they wait and divide the money equally if they have strong propensity to follow norms in general (high ϕ_i). Second, the rule in a rule-following task can be seen as a part of a 'promise-keeping contract' between a subject and the experimenter: by consenting to participate in the experiment, the subject is implicitly promising to follow the experimental instructions including the rule to wait at the lights. Thus, if we see promises as something that 'ought' to be kept and splitting the money equally in the Dictator game as something 'ought' to be done, then individuals who follow norms in general (high ϕ_i) will keep the promise and split the money equally, and individuals who do not follow norms (low ϕ_i) will not keep their promise and give the receiver nothing.

Together, the experiments of KW and KV provide, in my view, convincing evidence that pro-sociality in social dilemmas is determined to a large extent by an adherence to norms that is driven by more general 'desires' to behave in rule-compliant ways and more specific 'beliefs' concerning particular normative expectations in a particular context. In recent years both the norm elicitation task and the rule-following task were used by many researchers interested in decision making in various strategic settings (e.g., Kimbrough/Vostroknutov 2015; Barr et al. 2017; Kassas/Palma 2018; Panizza et al. 2018; 2020; Thomsson/Vostroknutov 2017; Gächter et al. 2017; Hoeft et al. 2018; Gürdal et al. 2018; Chang et al. 2019). There are several methodological contributions and replications. Merguei et al. (2020) tested a new version of the KW task with continuous appropriateness scale instead of a discrete 4-item one. D'Adda et al. (2016) showed that the norm elicitation task is robust to certain order effects.

4 Towards a Unified Theory of Normative Decision Making

4.1 A Need for a Theory of Social Norms

As I have suggested above, in principle we can measure norms in any settings and measure rule-following propensities in the population. Conceivably, this might provide sufficient information to predict what behaviour should be expected in a particular context. Nevertheless, the proposed methods do not really generalise

from one game context to another. Let me illustrate with an exaggerated example. Suppose that we have elicited the norm in the $4-Dictator game using the KW task and found that it is $\hat{x} = \$2$. How do we know that in a $6-Dictator game we will find $\hat{x} = \$3$? This sounds obvious, however for more complex generalisations it becomes much harder to give a definite answer.[15] Unless the costs of obtaining an estimate of $\hat{x}$ are zero, we will never be able to measure norms in all conceivable Dictator games, let alone in all game forms that are interesting to economists.

One way to solve this problem is to develop a theory that would tell us what norm should be present in each game form and context. In this case, measurements obtained from the KW task can be treated as experimental tests of a more general theory. If the theory predicts $\hat{x} = \$2$ in the $4-Dictator game and $\hat{x} = \$3$ in the $6-Dictator game, then after observing $\hat{x} = \$2$ in an experiment we can say that this corroboration makes $\hat{x} = \$3$ in the $6-Dictator game more likely. I believe that the creation of such a theory is the main challenge faced by the current generation of behavioural and experimental economists who are interested in social behaviour. Luckily, it seems that several groups of researchers are working in this direction (Sontuoso 2013; Cox et al. 2018; Ellingsen/Mohlin 2019; d'Adda et al. 2019). My co-author Erik Kimbrough and myself have also been busy developing such a theory. So, in the remainder of this paper I would like to present our theoretical results and to speculate about the future of the social norms paradigm in economics.

In our thinking about a theory of norms, we had to go back to the basic principles that we believed are at root of normative decision making. We followed a well-established tradition in philosophy that grounds 'moral sense' in the *emotions* that are aroused by both attained and forgone material payoffs, as well as *empathy* that allows us, humans, to understand what others might feel about their actual or counterfactual material payoffs (Hume 2003[1740]; Smith 1982[1759]; Prinz 2007). The idea that empathy plays an important role in the emergence of morality also dovetails nicely with findings of evolutionary human biology (e.g., Henrich 2015). Imagine that we live in a pre-historic tribe and let's say that each other week we go hunting bison. This is a collective undertaking, so the problem of distribution of meat comes up after each hunt. Without social norms guiding meat redistribution that everyone agrees to follow, each attempt at redistribution can easily turn into a brawl, since each tribe member obviously wants more meat. This

15 In the aforementioned example, List 2007 used a Dictator game where dictators first received $5 and then were given another $5 that they could share with receivers. Thomsson and Vostroknutov 2016 had the same set-up except that they gave dictators $10 and said that they could give away no more than $5. The behaviour in the two experiments was very different, which probably means that the game form was perceived differently from the normative perspective as well.

is not conducive to evolutionary success. So, how should the redistribution proceed? One possibility—that increases the survival chances of the tribe as a whole feeding back on individual gene level survival—is to observe how upset each tribe member will be with the prospect of getting no meat and try to distribute it so that the overall level of dissatisfaction with the resulting distribution is reduced. This would decrease the chances of a fight, plus take into account that some tribe members might need sustenance more than others, because they are sick for example. The internalization of such redistributive procedures based on empathy can in principle lead to the emergence of *injunctive norms*, which I define here as rules of social conduct that define how the redistribution 'ought' to be done regardless of any other factors (e.g., how others are redistributing).[16]

In Kimbrough and Vostroknutov (2020c) we analyse the norms that a 'dissatisfaction-minimizing' procedure can give rise to. We conjecture that in a given game form each allocation of material payoffs (e.g., (x, y) in case of two players) has an *injunctive normative valence*, which is a number in the interval $[-1, 1]$ representing how socially appropriate this allocation is in the context of all other possible allocations in the game form (-1 for the least appropriate and 1 for the most appropriate).[17] The normative valence of (x, y) is determined by the *dissatisfaction* that players feel in this allocation. Player i feels dissatisfaction at (x, y) when the material payoff that she has received (x) is less than some other material payoff that she *could have received* in some other allocation possible in the game form (e.g., some material payoff $z > x$). We assume that the dissatisfaction that i feels about (x, y) because of z is equal to $z - x$.[18] Following this logic, we can compute *aggregate dissatisfaction* at (x, y) by summing up dissatisfactions of all players because of all higher material payoffs that each of them could have received in the game form.[19] Once aggregate dissatisfactions are computed for each allocation in the game form, we multiply them by -1 and normalize the resulting numbers to the interval $[-1, 1]$, obtaining the injunctive normative valences. Thus, we postulate that the normative valence of an allocation is inversely proportional to the

16 Gavrilets/Richerson 2017 analyse an evolutionary model where internalized norm-following emerges. Their model is based on similar assumptions.

17 There is an implicit assumption here that all allocations that can happen in the game form are common knowledge among the players.

18 We use the additive functional form $z - x$ for simplicity and tractability. In general, dissatisfaction can be defined as a function $g(z, x)$ increasing in the first and weakly decreasing in the second argument (see Kimbrough/Vostroknutov 2020a).

19 In the left game form in *figure 2* the dissatisfaction of P1 at (20, 20) is $5 = 25 - 20$ (no dissatisfaction due to material payoff 15, because $15 < 20$); the dissatisfaction of P2 at (20, 20) is $15 = (25 - 20) + (30 - 20)$; and aggregate dissatisfaction at (20, 20) is $20 = 5 + 15$.

aggregate dissatisfaction that all players feel in it. In other words, the most socially appropriate allocation (the norm) is the allocation with the smallest aggregate dissatisfaction.[20] In fact, normative valences define a more complex object, an *injunctive norm function* $\eta(x, y)$, that maps each allocation (x, y) attainable in a game form into its normative valence in $[-1, 1]$. 'The norm' can then be equivalently defined as an allocation where the maximum of $\eta(x, y)$ is attained.[21] We assume that each player i maximizes a norm-dependent utility defined as

$$u_i(x, y) = x + \phi_i \eta(x, y). \tag{5}$$

Here as before, x is the material payoff of player i and $\phi_i \geq 0$ is the norm-following propensity. This utility function trades-off material payoffs to self and the desire to act in a way that decreases aggregate dissatisfaction. Since $\eta(x, y)$ is unambiguously determined by the material payoffs in the game form, what this utility specification gives us is a theory of normative behaviour that generates norms endogenously from the set of all material-payoff allocations in the strategic situation. The theory takes a game form as an input and produces a collection of norm-dependent utilities (one for each player) defining how they will behave.[22]

20 Some preliminary theoretical results that I obtained with my colleague Hannes Rusch suggest that this specific way of computing normative valences gives an evolutionary advantage to the group that adheres to them as compared to following normative valences obtained from other forms of aggregation. For example, aggregating *rejoice* from having a higher material payoff than what could have been received (the opposite of dissatisfaction) creates a completely different set of normative incentives that do not favour cooperation as much as normative valences based on dissatisfactions, which leads to poorer relative performance of norms based on rejoice. We conjecture that dissatisfaction-based norms exist because they are better at promoting cooperation than norms based on other sentiments.

21 There is a slight discrepancy between our definition of a norm function and similar objects in other models. In the earlier studies (e.g., Kessler/Leider 2012; Krupka/Weber 2013) 'the norm' was defined as an *action*, whereas we attach normative valences to game outcomes (allocations) instead. This difference plays no role in simple one-move game forms like the Dictator game where actions are equivalent to allocations that they entail. However, in more complex multi-move game forms our definition makes more sense since actions in such game forms are just means to achieving certain allocations. So, actions can 'acquire' normative valence due to the allocations that they lead to.

22 I would like to note here that our theory applies to 'small' strategic interactions of several players without past history of choices and where normative expectations are assumed to be driven by the theory (all players believe that others' utilities are described by (5)). This is the kind of interactions that is usually studied in microeconomics and that can be tested in the lab. When we talk about 'large' social norms or institutions with many participants and long history and traditions, many other factors influence the behaviour: the 'descriptive' component pertaining to the usage of the institution in the past (I introduce descriptive norms into the model in *section*

In Kimbrough and Vostroknutov (2020c) we we show that this theory predicts that equal split is the norm in the Dictator and Ultimatum games and that the norm in game forms studied by Engelmann and Strobel (2004) is the most material-payoff-efficient allocation or the allocation with the highest minimal material payoff (if we additionally assume a concave utility of money instead of the linear one as in (5)). Thus, our theory can generate different 'social preferences' in different game forms, which potentially resolves the multiplicity of social preferences problem that I discussed in *section 2.2*. Since the normative valence of each allocation in a game form depends on all other allocations, our theory also can resolve problems mentioned in *section 2.3*. The presence of an allocation (20, 20) in the left game of *figure 2* changes the relative normative valences of allocations (15, 30) and (25, 25) as compared to the right game form where (20, 20) is not available. This change qualitatively explains the different percentages of subjects choosing (15, 30) and (25, 25). Thus, our theory seems to incorporate 'reciprocal' behaviour in dynamic games, at least in those that we checked (Trust games, repeated Dictator games). The same goes for the problems with context described in *section 2.4*. Adding 'taking' options to the Dictator game changes the norm from offering \$2.5 to offering \$0, exactly the change that List (2007) observes in his experiment.[23]

When we thought about norms in a social context discussed in *section 2.5* (ownership claims, in- or outgroup, social status) we have realised that these phenomena can be easily incorporated in our model if we think that people attach different weights to dissatisfactions of others when computing aggregate dissatisfaction. For example, the dissatisfaction of a person with low social status (or outgroup) can count less than the same dissatisfaction coming from a high-status individual (or ingroup). This makes it socially appropriate to give larger portions of the pie in the Dictator game to high-status individuals or the ingroup than to

4.3); empirical and normative expectations (Bicchieri 2006); trust in the authority who is charged with maintaining the institution; etc. In such environments the 'moral' norms that our theory describes should still influence the decisions, though their influence can be limited by these other factors.

23 Notice as well that our theory can produce complex strategic behaviours in games with norm-dependent utility (5). For example, a selfish proposer ($\phi_i = 0$) in the Ultimatum game, who believes that the responder is norm-following, will not offer zero or some small amount because she expects low offers to be rejected by the norm-following player (punishment of the move not consistent with the norm of equal split that our theory predicts). So, selfish players in some contexts might behave in accordance with the norm, but for purely selfish reasons (avoiding punishment). At the same time, a very norm-following proposer (high ϕ_i) will offer equal split not because he is trying to avoid punishment, but because offering equal split increases his norm-dependent utility, even if the responder is selfish and will accept any division.

low-status individuals or the outgroup. In general, the lower is the *social weight* of a player in the aggregation, the less others will care about his dissatisfaction and correspondingly his material payoffs.[24] When we attach player-specific weights to their dissatisfactions we obtain norm functions influenced by social context. We find that the model with such 'social' weights does account for behavioural changes between ingroup and outgroup in social context experiments mentioned in *section 2.5*, for example in Chen and Li (2009). Ownership claims for material payoffs can be similarly introduced: the owner of some amount of money will feel much more dissatisfied after losing it than a player who lost the same amount of windfall money or the player who lost someone else's money (e.g., if previously stolen). Introducing appropriate weights on dissatisfactions in this manner creates a general class of norms according to which it is not inappropriate for the money owner to not share it with others (Oxoby/Spraggon 2008).

Finally, in Kimbrough and Vostroknutov (2020c) we present a model of punishment in dynamic game forms that seems necessary for maintaining norm compliance (Mackie 1982; Henrich 2015). In our view such retributive norms are responsible for rejections in the Ultimatum and other similar game forms. Having injunctive normative valences of each allocation already defined, it becomes relatively simple to determine by how much a player violates the norm when she chooses an action that makes the norm unreachable. In such cases, we postulate that the size of punishment is proportional to the size of norm violation that can be computed as the difference in normative valences of the most appropriate outcome and the outcome that a norm-violator intended to reach by deviating. We show how punishment can account for rejections in the Ultimatum game and for the connection that we found in KV between the rejection thresholds and rule-following propensity (see *section 3*). In addition, our assumption that punishment of norm-violators is a norm in itself sheds some light on *third-party punishment* when people punish someone for a norm violation that does not harm them directly (Fehr/Fischbacher 2004). Since punishment can be seen as a norm in its own right, any norm-following individual feels obliged to get involved in it even if the norm violation had no direct material consequences for her. This idea is corroborated by a wide-spread practice of punishing individuals who refuse to punish others (e.g., Axelrod 1986).

24 Social weight of a player can also be negative, which would normatively justify outright hostility towards him.

4.2 Social Norms and Bounded Rationality

The fact that our theory can account for observations from a wide variety of experiments suggests that the models in this class can be promising candidates for developing a unified theory of normative decision making in games. Even though our model is just a first imperfect example of this new class of models, we can nevertheless explore some of its implications concerning how dissatisfaction-based norms can fare in reality. In Kimbrough and Vostroknutov (2020a) we make the observation that in practice $\eta(x, y)$ is *hard to compute*. Indeed, to calculate aggregate dissatisfaction we need to know the dissatisfactions of all players in all possible allocations in a game. This can be a daunting task even for a person who possesses all the necessary information. Economists tend to assume that a fully rational economic agent is not facing any factual constraints of memory and reasoning capacity and will therefore be up to the task (Simon 1990). Typically though, real agents are subject to cognitive constraints and limited control over emotional influences on decision making. Therefore, it is not inconceivable that people might rely on *moral rules* when making normative judgements instead of computing the norm function $\eta(x, y)$ in each new situation. In our terminology, moral rules are simple heuristics that are supposed to approximate the norm function in some class of game forms.[25] For example, the rule to divide a 'pie' equally among co-equal strangers in Dictator-game-like situations may be one of such heuristics that produces results close enough to $\eta(x, y)$ in most cases. If the actual norm is very costly to compute, then moral rules can be optimally chosen as a crude but computationally cheap replacement.

In Kimbrough and Vostroknutov (2020a) we provide a method of determining which moral rules are likely to emerge in specific classes of game forms. The idea is simple: take some class of game forms (e.g., all Dictator games), define some moral rule (e.g., the pie should be divided equally), and check how often the norm generated by the norm function $\eta(x, y)$ coincides with the prescription of the moral rule. If a moral rule predicts the same norm as $\eta(x, y)$ in, say, 95% of game forms from the chosen class then we can conclude that this rule is likely to emerge in this class of game forms simply because it is cheaper for the decision makers to use the moral rule instead of the computationally complex $\eta(x, y)$. This technique allows to understand why certain moral rules arise in specific contexts and even how costly (in terms of dissatisfaction) it is to use them. Another important result that we prove analytically is that there is no moral rule that can fully

25 The term 'moral rule' is used in many literatures with somewhat different meaning. In what follows I will use it explicitly as just defined without making any allusions to the other possible usages of the term.

capture the complexity of $\eta(x, y)$. The logic behind this result is that $\eta(x, y)$ is *extremely context-dependent* (the normative valence at (x, y) depends on *all* other allocations), whereas a moral rule—at least if it is constructed in accordance with very general axioms that we propose—can only be context-dependent to a certain lower degree. This result says that any rule codified in law, for instance, can never fully capture the normative complexity that exists in reality. In other words, situations will always arise in which the law is not going to be perfectly 'just'.

I believe that in reality moral rules should exert a rather large influence on normative decision making. In the world where people have problems computing injunctive norms we should observe many of them seeking *advice* on how they should behave. People who give such advice are well-known to all of us. They are elderly, shamans, priests, kings, philosophers, politicians, psychoanalysts, etc. Ancient myths, legends, fairy tales, religious texts, movies are packed with moral lessons and from the bounded rationality perspective can be considered as 'moral textbooks' that teach people using simplified examples of what is right and what is wrong. The ubiquity of such sources strongly suggests that people have difficulties navigating moral conundra and need guidance in the form of moral rules.

4.3 Descriptive Norms

The idea that injunctive norms that I defined above are hard to compute, and that the reduction of computational costs is the reason behind certain types of normative behaviour (e.g., adoption of moral rules), can shed some light on other norm-related phenomena. In particular, some people attach positive normative valences in the sense of 'ought' mentioned above to outcomes in a game that happened more often in the past than other outcomes (a form of individual descriptive ethics). For example, people who wash their hands because others do it (but not because not washing hands harms the community) might still reprimand an individual who does not wash her hands because they find this behaviour morally wrong (we ought to do what others in the community do). The reason why people might think that something is 'right' simply because it happened often in the past (and 'wrong' if it did not happen often) might be that they cannot afford to compute $\eta(x, y)$ and use observations of past behaviour as an approximation of $\eta(x, y)$. Sometimes it might even be the only possible way to learn it. For example, it is a well-established fact that children have a very strong tendency to copy the behaviour of prestigious adults (Henrich/Gil-White 2001; Laland 2018). When considered from the complexity of norms perspective, it makes sense given that children do not have a fully developed capacity for empathy and 'moral calculus' in general (Wellman et al. 2001). So, it is possible that copying others is the only

way children can learn to behave in a socially appropriate manner expected of them.[26] In what follows I will call normative beliefs about what is right and what is wrong obtained from observations of others *descriptive norms*. This definition is not completely in line with other definitions of descriptive norms (e.g., Bicchieri 2006), however it is more natural when we talk about copying others as a way to learn what ought to be done.

Descriptive norms of this type play an important role in normative decision making. To give an example, Nishi et al. (2016) find that Americans make cooperative choices in a social dilemma faster than they make selfish choices (and more of them as well). Conversely, Indian subjects make selfish choices faster than cooperative ones (and more selfish choices). If everyone in both populations was computing injunctive norms, for example $\eta(x, y)$, then we would not see any difference in reaction times or the proportions of cooperative choices. Therefore, the existence of this difference may indicate that faster decisions correspond to the 'common' choice in the respective countries: cooperation in the US and selfish behaviour in India. This suggests that many subjects in these experiments cooperated or defected because this is simply what others around them are doing all the time. If this argument is correct, then individuals from cultures where selfishness is a descriptive norm (like in India presumably) can get involved in a so-called 'antisocial punishment', a puzzling phenomenon reported in Herrmann et al. (2008) when cooperators in a Public Goods game are punished by defectors. Indeed, if one comes from a culture where selfish behaviour is common, this person might punish cooperators simply because they violate the 'selfishness norm'.[27]

In our third paper (Kimbrough/Vostroknutov 2020b) we we incorporate the aforementioned kind of descriptive norms into our theory of injunctive norms (2020c). In particular, we assume that before a game is played players can observe some history of previous choices by others. These previous choices form a distribution over final material-payoff allocations that can be normalized to $[-1, 1]$ and treated as a *descriptive norm function* $\delta(x, y)$ that produces *descriptive normative valences* for each allocation in the game form. Specifically, the allocations that were never chosen in the past are assigned a descriptive normative valence of -1 (very inappropriate) and the allocations that are chosen all the time are assigned

26 Of course, copying the behaviour of others can have other uses, apart from being a cheap substitute for injunctive norms. Many individual skills, like food-processing techniques for instance (Henrich 2015), are learned explicitly through observation.

27 Herrmann et al. 2008 do not find antisocial punishment in the US or Northern European countries. They find that it is mostly prevalent in Oman, Greece, Russia, Saudi Arabia, and Belarus.

a high descriptive normative valence close to 1. Next, we postulate the following norm-dependent utility function:

$$u_i(x, y) = x + \phi_i \left[\psi_i \eta(x, y) + (1 - \psi_i)\delta(x, y)\right]. \quad (6)$$

This utility specification is the same as (5) except that instead of the injunctive norm function $\eta(x, y)$ we have $\psi_i \eta(x, y) + (1 - \psi_i)\delta(x, y)$, which is a convex combination of the injunctive and descriptive norm functions. To which extent a particular individual i is relying on injunctive or descriptive norms is defined by the parameter $\psi_i \in [0, 1]$, that can be viewed as a fixed individual preference. There are many reasons why ψ_i can be different across people and populations. One possibility, as I mentioned above, is that people might not possess cognitive capacities to compute $\eta(x, y)$, so they rely on descriptive norms instead (ψ_i close to 0). Or it can be that in some cultures people mostly rely on injunctive norms (ψ_i close to 1), whereas in other cultures on descriptive norms, which is not directly linked to computational complexity. At this point there is no evidence that would allow us to make any conclusions about the nature of heterogeneity in ψ_i.[28]

Despite our lack of knowledge with regard to ψ_i, the model in (6) can explain certain behaviours that cannot be interpreted in a purely injunctive theory of norms. The most important phenomenon is social learning. Bicchieri and Xiao (2009) and many other studies (e.g., Panizza et al. 2020) find that people's choices in social dilemmas are influenced by the observation of others' actions. Under our theory, people who are most prone to do that are those with low ψ_i, or people who rely a lot on descriptive norms. Antisocial punishment that I have mentioned above is another type of behaviour that can be understood in terms of our theory. 'Descriptive people' (with low ψ_i) who have observed many defections in the past will defect themselves and punish others who cooperate (because they break the descriptive norm). However, 'injunctive people' (with high ψ_i) will cooperate and punish others who defect, because for them the descriptive norm is irrelevant. I believe that this is a plausible explanation, because in countries where antisocial punishment is observed there is also a substantial degree of normal 'social' punishment of the defectors. The presence of both types of punishment in one population can therefore be explained by the heterogeneity in ψ_i and by the long past history of selfish behaviour.

28 I plan to change this situation by conducting experiments specifically designed to estimate ψ_i and its influence on behaviour in social dilemmas.

5 Future Developments

Throughout the text I used the term 'normative decision making' to define the scope of applicability of behavioural theories in economics. This choice was not random. The 'decision making' part emphasizes that the theories of norms are applicable to any environment where there is a choice among multiple alternatives. The 'normative' part means that this choice is usually between having more consumption utility and adhering to a norm. I can hardly imagine any choice situation to which this definition does not apply, except probably individual decision making (choices under uncertainty, individual learning, etc.). Therefore, I believe that the behavioural theories of norms outlined above and those yet to come can be of tremendous help to all social scientists. It is not only that these theories provide a framework in which hypotheses about social norms can be experimentally tested, they also give us a mathematically precise language in which we can discuss and develop new theories.

Assuming that our theory of norms presented in Kimbrough and Vostroknutov (2020c; a; b) got at least the broad picture right, I would like to speculate about the types of questions that can be tackled with it. The normative approach in behavioural economics gives us a new toolbox of experimental tasks that can be used to measure norms and norm-following behaviour. I talked already about the norm elicitation task (Krupka/Weber 2013) and the rule-following tasks (Kimbrough/Vostroknutov 2016; 2018) that can be used to identify norms and preferences 'in the wild' (Bicchieri 2016). In an ongoing study in Iraq we have also used *third-party Dictator games* (Chen/Li 2009) to identify in/outgroup social weights that serve as an input to our model (see *section 4.1*). We (Erik Kimbrough, Vera Mironova, and myself) ran a survey where people on the streets of Baghdad and Mosul were asked to divide $2 between two anonymous others who are followers of the same/different religion (Shia or Sunni) and belong to the same/different tribe (extended family). The model tells us that the proportions in which people divide the money—for example, give $1.5 to a person of the same religion and tribe and $0.5 to a person of different religion and different tribe—uncover their social weights on the two outgroups (different religion, different tribe) and can be used to predict how they are going to behave when matched with individuals from these outgroups in any other game form. Our preliminary results indicate for example that Iraqi people care about others from the same religion (but different tribe) more than they care about others from the same tribe but different religion. By 'more' I mean higher social weight, which in our theory implies more cooperation, more sharing, more reciprocity, which collectively can be called 'more trust'. This technique can be used to create maps of 'trust relationships' (social

weights) among different social groups in a given country or community and provide valuable predictions of individual behaviour in games involving people from these groups.

An experimental task to estimate ψ_i, an important parameter that defines to which extent an individual follows descriptive or injunctive norms, is currently under construction. I plan to use another third-party Dictator game in which subjects are asked to divide some amount of money between two others (say, a Green and a Yellow person with randomly assigned labels) knowing that previous 10 participants gave all the money to the Yellow person and nothing to the Green one (these 10 people divided money between different Yellow and Green persons, so the money is not 'accumulated' by the same Yellow/Green individuals). If a subject mostly follows injunctive norms (high ψ_i), she will divide money equally ignoring the information about the choices of previous subjects as well as the randomly assigned colour. If however she follows descriptive norms (low ψ_i) then she should adhere to the previous behaviour and give all the money to the Yellow person. I conjecture that the choices in this task will give us an estimate of ψ_i, which can be used to identify this 'normative preference' in different social groups.

For example, some indirect evidence provided in Thomsson and Vostroknutov (2017) suggests that people who self-identify as conservatives are more prone to follow descriptive rather than injunctive norms, whereas people who self-identify as liberals are the opposite. This assertion will need thorough testing. However, if there is any truth to it, then heterogeneity in ψ_i can be the factor that defines *social polarisation*. To understand why, imagine that we have two geographically separate communities, a village and a town. The village is populated mostly by individuals who follow a descriptive norm $\delta(x, y)$ that does not coincide with the injunctive norm $\eta(x, y)$: e.g., they prosecute gay people. In the town there is a mixture of people from different places, so descriptive norms are not very prominent (you can observe any sorts of behaviour) which leads to the prevalence of injunctive norms $\eta(x, y)$ that dictate that prosecution of gay people is wrong. If you are a person who is motivated by injunctive norms $\eta(x, y)$ and you can choose where to live, you will prefer the town because it is more consistent with your normative views described by $\eta(x, y)$. However, if you are a person inclined to follow descriptive norms, you may choose the village because it has well-defined (descriptive) norms of behaviour, unlike the town where descriptive norms are 'eclectic'. This is a possible mechanism of social polarisation that can be tested with the task for measuring ψ_i.

6 Conclusion

In this article I have presented an account of how theorizing about normative behaviour in experimental economics developed over the years. As the empirical evidence was gradually accumulating, the theories of normative decision making evolved as well: from theories of social preferences (inequity aversion, preference for material-payoff efficiency or maximin) and reciprocity models to the explicit introduction of rule-following propensity in the norm-dependent utility functions. Today, the social norms paradigm in behavioural economics can offer a set of new experimental tools, like the norm elicitation task (Krupka/Weber 2013) or the rule-following task (Kimbrough/Vostroknutov 2016; 2018), that provide means to investigate the influence of social norms on behaviour in practically any context that involves choice. The estimates from these tasks—as well as measures of 'trust relationships' among different social groups obtained from third-party Dictator games and the history of previous choices—can help to create 'models' of existing institutions. A norm-dependent utility function like (6), calibrated with the estimates of relevant parameters and embedded in a game, can be used then to predict behaviour or to hypothesize about consequences of policies that change the institution. Though, it is important to emphasise that the theory of norms that I sketched above is a positive theory not intended to make normative statements about which norms should or should not prevail in a given society.

On a final note, I hope that this overview has demonstrated that the machinery of rational choice theory commonly used in economics does not have to be at odds with other approaches to studying social norms in social sciences and philosophy. The models sketched in this article are all based on rational choice theory, however they show enough flexibility to incorporate a wide range of behavioural phenomena related to norm-following. These models are also capable of generating new predictions like social polarisation that—even though not conceptually new to philosophy—nevertheless constitute an important step forward because they are expressed in a strict mathematical language that makes them amenable to experimental tests. This paves the way to the development of a unified theory of normative decision making that could become the one theory that philosophers and many social scientists, including economists, can actually agree on.

Acknowledgment: I would like to thank Michael Baurmann, Marat Charlaganov, Erik Kimbrough, Hartmut Kliemt, Anton Leist, Hannes Rusch, and James Tremewan for invaluable comments. All mistakes are my own.

References

Akerlof, G. A./R. E. Kranton (2000), Economics and Identity, in: *Quarterly Journal of Economics* 115, 715–753

Axelrod, R. (1986), An Evolutionary Approach to norms, in: *American Political Science Review* 80, 1095–1111

Baader, M./A. Vostroknutov (2017), Interaction of Reasoning Ability and Distributional Preferences in a Social Dilemma, in: *Journal of Economic Behavior & Organization* 142, 79–91

Bardsley, N. (2008), Dictator Game Giving: Altruism or Artefact?, in: *Experimental Economics* 11, 122–133

Barr, A./T. Lane/D. Nosenzo (2017), *On the Social Appropriateness of Discrimination*, technical report, CeDEx Discussion Paper Series

Berg, J./J. Dickhaut/K. McCabe (1995), Trust, Reciprocity, and Social History, in: *Games and Economic Behavior* 10, 122–142

Bicchieri, C. (2006), *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge

— (2016), *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*, Oxford

—/E. Xiao (2009), Do the Right Thing: But Only If Others Do So, in: *Journal of Behavioral Decision Making* 22, 191–208

Bolton, G. E./A. Ockenfels (2000), ERC: A Theory of Equity, Reciprocity, and Competition, in: *American Economic Review* 90, 166–193

Boyd, R./P. J. Richerson (1988), An Evolutionary Model of Social Learning: The Effects of Spatial and Temporal Variation, in: *Social Learning: Psychological and Biological Perspectives*, 29–48

Cappelen, A. W./A. D. Hole/E. Ø. Sørensen/B. Tungodden (2007), The Pluralism of Fairness Ideals: An Experimental Approach, in: *American Economic Review* 97, 818–827

Chang, D./R. Chen/E. Krupka (2019), Rhetoric Matters: A Social Norms Explanation for the Anomaly of Framing, in: *Games and Economic Behavior* 116, 158–178

Charness, G./M. Rabin (2002), Understanding Social Preferences with Simple Tests, in: *Quarterly Journal of Economics* 117, 817–869

Chen, Y./S. X. Li (2009), Group Identity and Social Preferences, in: *American Economic Review* 99, 431–457

Cox, J. C./C. A. Deck (2005), On the Nature of Reciprocal Motives, in: *Economic Inquiry* 43, 623–635

—/D. Friedman/S. Gjerstad (2007), A Tractable Model of Reciprocity and Fairness, in: *Games and Economic Behavior* 59, 17–45

—/J. A. List/M. Price/V. Sadiraj/A. Samek (2018), *Moral Costs and Rational Choice: Theory and Experimental Evidence*, mimeo, Georgia State University, University of Chicago, University of Alabama, University of Southern California

d'Adda, G./M. Drouvelis/D. Nosenzo (2016), Norm Elicitation in Within-subject Designs: Testing for Order Effects, in: *Journal of Behavioral and Experimental Economics* 62, 1–7

—/M. Dufwenberg/F. Passarelli/G. Tabellini (2019), *Partial Norms*, Available at SSRN 3353192

De Waal, F./S. Macedo/J. Ober (2006), *Primates and Philosophers: How Morality Evolved*, Princeton

Demsetz, H. (1974), Toward a Theory of Property Rights, in: *Classic Papers in Natural Resource Economics*, 163–177

Dufwenberg, M./G. Kirchsteiger (2004), A Theory of Sequential Reciprocity, in: *Games and Economic Behavior* 47, 268–298
Ellingsen, T./E. Mohlin (2019), *Decency*, Department of Economics School of Economics and Management, Lund University, 3
Engel, C. (2011), Dictator Games: A Meta Study, in: *Experimental Economics* 14, 583–610
Engelmann, D./M. Strobel (2004), Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution, in: *American Economic Review* 94, 857–869
Falk, A./U. Fischbacher (2006), A Theory of Reciprocity, in: *Games and Economic Behavior* 54, 293–315
Fehr, E./K. M. Schmidt (1999), A Theory of Fairness, Competition, and Cooperation, in: *Quarterly Journal of Economics* 114, 817–868
—/U. Fischbacher (2004), Third-party Punishment and Social Norms, in: *Evolution and Human Behavior* 25, 63–87
—/I. Schurtenberger (2018), Normative Foundations of Human Cooperation, in: *Nature Human Behaviour* 2, 458–468
Fleming, P./D. J. Zizzo (2015), A Simple Stress Test of Experimenter Demand effects, in: *Theory and Decision* 78, 219–231
Forsythe, R./J. L. Horowitz/N. E. Savin/M. Sefton (1994), Fairness in Simple Bargaining Experiments, in: *Games and Economic Behavior* 6, 347–369
Franzen, A./S. Pointner (2013), The External Validity of Giving in the Dictator game, in: *Experimental Economics* 16, 155–169
Gächter, S./L. Gerhards/D. Nosenzo (2017), The Importance of Peers for Compliance with Norms of Fair Sharing, in: *European Economic Review* 97, 72–86
Galeotti, F./M. Montero/A. Poulsen (2018), Efficiency versus Equality in Bargaining, in: *Journal of European Economic Association*, forthcoming
Gavrilets, S./P. J. Richerson (2017), Collective Action and the Evolution of Social Norm Internalization, in: *Proceedings of the National Academy of Sciences* 114, 6068–6073
Goeree, J./C. Holt (2001), Ten Tittle Treasures of Game Theory and Ten Intuitive Contradictions, in: *American Economic Review* 91, 1402–1422
Gürdal, M. Y./O. Torul/A. Vostroknutov (2018), *Norm Compliance, Enforcement, and the Survival of Redistributive Institutions*, mimeo, Bogazici University and University of Trento
Güth, W./R. Schmittberger/B. Schwarze (1982), An Experimental Analysis of Ultimatum Bargaining, in: *Journal of Economic Behavior and Organization* 3, 367–388
Henrich, J. (2015), *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*, Princeton
—/F. J. Gil-White (2001), The Evolution of Prestige: Freely Conferred Deference As a Mechanism for Enhancing the Benefits of Cultural Transmission, in: *Evolution and Human Behavior* 22, 165–196
Herrmann, B./C. Thöni/S. Gächter (2008), Antisocial Punishment across Societies, in: *Science* 319, 1362–1367
Hoeft, L./W. Mill/A. Vostroknutov (2018), *Normative Perception of Power Abuse*, mimeo, University of Mannheim, MPI Bonn, University of Trento
Hoffman, E./M. L. Spitzer (1985), Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice, in: *Journal of Legal Studies* 14, 259–298
—/K. McCabe/K. Shachat/V. Smith (1994), Preferences, Property Rights, and Anonymity in Bargaining Games, in: *Games and Economic Behavior* 7, 346–380

—/—/V. Smith (1996), Social Distance and Other-regarding Behavior in Dictator Games, in: *American Economic Review* 86, 653–660
—/—/— (2000), The Impact of Exchange Context on the Activation of Equity in Ultimatum Games, in: *Experimental Economics* 3, 5–9
Hume, D. (2003[1740]), *A Treatise of Human Nature*, Oxford
Isaac, R. M./J. M. Walker/A. W. Williams (1994), Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Large Groups, in: *Journal of public Economics* 54, 1–36
Isoni, A./R. Sugden (2018), Reciprocity and the Paradox of Trust in Psychological Game Theory, in: *Journal of Economic Behavior & Organization* 167, 219–227
Kahneman, D./J. L. Knetsch/R. H. Thaler (1986), Fairness and the Assumptions of Economics, in: *Journal of business* 59, S285–S300
Kassas, B./M. A. Palma (2018), *Social Norms: First Impulse or Last Resort?*, mimeo, Texas
Kessler, J. B./S. Leider (2012), Norms and Contracting, in: *Management Science* 58, 62–77
Kimbrough, E. O./A. Vostroknutov (2015), The Social and Ecological Determinants of Common Pool Resource Sustainability, in: *Journal of Environmental Economics and Management* 72, 38–53
—/— (2016), Norms Make Preferences Social, in: *Journal of European Economic Association* 14, 608–638
—/— (2018), A Portable Method of Eliciting Respect for Social Norms, in: *Economics Letters* 168, 147–150
—/— (2020a), *Injunctive Norms and Moral Rules*, mimeo, Chapman University and Maastricht University
—/— (2020b), *A Theory of Descriptive Norms*, mimeo, Chapman University and Maastricht University
—/— (2020c), *A Theory of Injunctive Norms*, mimeo, Chapman University and Maastricht University
Kliemt, H. (2020), Economic and Sociological Accounts of Social Norms, in: *Analyse & Kritik* 42, 41–96
Knack, S. (2000), *Trust, Associational Life and Economic Performance in the OECD*, manuscript, The World Bank
Krupka, E. L./R. A. Weber (2013), Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?, in: *Journal of European Economic Association* 11, 495–524
Laland, K. N. (2018), *Darwin's Unfinished Symphony: How Culture Made the Human Mind*, Princeton
List, J. A. (2007), On the Interpretation of Giving in Dictator Games, in: *Journal of Political Economy* 115, 482–493
López-Pérez, R. (2008), Aversion to Norm-breaking: A Model, in: *Games and Economic behavior* 64, 237–267
Lumsden, C. J./E. O. Wilson (1981), *Genes, Mind, and Culture: The Coevolutionary Process*, Harvard
Mackie, J. L. (1982), Morality and the Retributive Emotions, in: *Criminal Justice Ethics* 1, 3–10
McCabe, K. A./M. L. Rigdon/V. L. Smith (2003), Positive Reciprocity and Intentions in Trust Games, in: *Journal of Economic Behavior and Organization* 52, 267–275
Merguei, N./M. Strobel/A. Vostroknutov (2020), *The Effect of Heterogeneous Norms on the Decision to Punish Norm Violators*, mimeo, Maastricht University

Nishi, A./N. A. Christakis/D. G. Rand (2016), *Cooperation, Decision Time, and Culture: Online Experiments with American and Indian Participants*, mimeo, Yale University

Oosterbeek, H./R. Sloof/G. Van De Kuilen (2004), Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-analysis, in: *Experimental Economics* 7, 171–188

Oxoby, R. J./J. Spraggon (2008), Mine and Yours: Property Rights in Dictator Games, in: *Journal of Economic Behavior & Organization* 65, 703–713

Panizza, F./A. Vostroknutov/G. Coricelli (2018), *Meta-context and Choice-set Effects in Mini-dictator Games*, mimeo, University of Trento and University of Southern California

—/—/— (2020), *Norm Conformity Leads to Extreme Social Behavior*, mimeo, University of Trento, Maastricht University, University of Southern California

Prinz, J. (2007), *The Emotional Construction of Morals*, Oxford

Simon, H. (1990), A Mechanism for Social Selection and Successful Altruism, in: *Science* 250, 1665–1668

Smith, A. (1982[1759]), *The Theory of Moral Sentiments*, Indianapolis

Sontuoso, A. (2013), *A Dynamic Model of Belief-dependent Conformity to Social Norms*, MRTA paper 53234

Sugden, R. (1993), Thinking As a Team: Towards an Explanation of Nonselfish Behavior, in: *Social Philosophy and Policy* 10, 69–89

— (2004), *The Economics of Rights, Co-operation and Welfare*, Springer

Thomsson, K./A. Vostroknutov (2017), Small-world Conservatives and Rigid Liberals: Attitudes towards Sharing in Self-proclaimed Left and Right, in: *Journal of Economic Behavior and Organization* 135, 181–192

Tuomela, R./K. Miller (1985), We-intentions and Social Action, in: *Analyse & Kritik* 7, 26–43

Wellman, H. M./D. Cross/J. Watson (2001), Meta-analysis of Theory-of-mind Development: The Truth about False Belief, in: *Child Development* 72, 655–684

Zelmer, J. (2003), Linear Public Goods Experiments: A Meta-analysis, in: *Experimental Economics* 6, 299–310

Zizzo, D. J. (2010), Experimenter Demand Effects in Economic Experiments, in: *Experimental Economics* 13, 75–98