Game-Theoretic Accounts of Social Norms: The Role of Normative Expectations

Cristina Bicchieri *and* Alessandro Sontuoso

November 3, 2017

<a>1.INTRODUCTION

The everyday notion of social norms consists of informal rules of behavior followed by a specific group or society. Social norms constitute a particular instance of the class of informal institutions, which have long been studied by sociologists, psychologists and anthropologists. Economists, too, have become interested in norms, as empirical work on cultural traits has shown that informal institutions (such as social norms) may affect economic outcomes (North 1991; for an extensive review of the literature on cultural variables and economic choices, see Alesina and Giuliano 2015). While neoclassical economics traditionally conceived of institutions as exogenous constraints, research in political economy has generated new insights into the study of *endogenous* institutions. Specifically, endogenous norms have been shown to restrict the individual's action set and drive preferences over action profiles (Bowles 1998; Ostrom 2000).

As a consequence, the standard economic framework positing exogenous (and in particular self-centered) preferences has come under scrutiny. Widely documented deviations from the predictions of models with self-centered agents have informed alternative accounts of individual choice (for one of the first models of interdependent preferences, see Stigler and Becker 1977). Everyday examples of these deviations may be brought about by norms that informally prescribe how people ought to behave in the household or workplace. For instance, Arrow's (1972) models of job discrimination show that entrepreneurs who could turn a profit on hiring labor cheaply from a racially discriminated group were restrained from doing so, owing to

Contact: Philosophy, Politics and Economics, University of Pennsylvania, 249 S. 36th St., Philadelphia, Pennsylvania, 19104. E-mail: cb36@sas.upenn.edu; sontuoso@sas.upenn.edu.

social norms involving discriminatory principles. Similarly, Akerlof's (1980) analysis shows that if a social norm prohibited an employer from hiring labor at a reduced wage, employees would not cooperate in training new workers who undercut existing wages, as they would suffer a loss of reputation for partaking in the norm violation. Other practices that have been explained by the enforcement of informal norms include the voluntary supply of public goods (Sugden 1984), as well as altruistic or reciprocity-based transactions such as gift-giving (for reviews, see Kolm and Mercier Ythier 2006).

Some of the above accounts have helped reconcile insights about norm-driven behavior with instrumental rationality. They have contributed to informing the design of laboratory experiments on non-standard preferences (for a survey of early experiments, see Ledyard 1995; more recent experiments are reviewed by Fehr and Schmidt 2006 and Kagel and Roth 2016). In turn, experimental findings have inspired the formulation of a wide range of models aiming to rationalize the behavior observed in the laboratory (for reviews, see Camerer 2003; Dhami 2016).

In this connection, it has been argued that the upholding of social norms could simply be modeled as the optimization of a utility function that includes the others' welfare as an argument. For instance, consider some of the early social preference theories, such as Bolton and Ockenfels's (2000) or Fehr and Schmidt's (1999) models of inequity aversion. These frameworks can explain a wealth of evidence on preferences for equitable payoff distributions; they cannot, however, account for (belief-dependent) conditional preferences, such as those reflecting principles of reciprocity (for example, I will be nice to others, if I believe others will be nice to me).

The approach to social norms taken by philosophically inclined scholars has highlighted the importance of conditional preferences in supporting social norms (Sugden 2000; Bicchieri 2002). In particular, according to Bicchieri's (2006) account, preferences for conformity to social norms are conditional on empirical beliefs (that is, first-order beliefs that a certain behavior will be followed) as well as normative expectations (that is, second-order beliefs that a certain behavior ought to be followed).

We note that some of the social preference theories can account for motivations conditional on *empirical* beliefs, whereby a player upholds a principle of fair behavior if he or she believes co-players will uphold it too (Rabin 1993; Charness and Rabin 2002; Dufwenberg and Kirchsteiger 2004; Falk and Fischbacher 2006).[1] These theories presuppose that players are hardwired with a notion of fair or kind behavior, as exogenously defined by the theorist: since they implicitly assume that all players have internalized a unique, exogenous, normative standpoint (as reflected in some notion of fairness or kindness), these theories do not explicitly model *normative* expectations. Hence, players' preferences are assumed to be conditional solely on their empirical beliefs; that is, preferences are conditional on whether others will behave fairly (according to an exogenous principle) or not.

We stress that social preferences should not be conflated with social norms. Social preferences generally capture stable dispositions toward an exogenously defined principle of conduct (Bicchieri 2006; Binmore 2010). By contrast, social norms are better studied as group-specific, context-dependent solutions to strategic problems (Bicchieri 1993; Young 1998; Sugden 2005). Such solutions are brought about by a particular class of preferences, conditional on the relevant set of empirical beliefs and normative expectations. More precisely, in what follows we define social norms as behavioral regularities emerging in a mixed-motive (that is,

social dilemma) game,[2] as a result of preferences for conformity conditional on an endogenous set of beliefs and expectations (Bicchieri 2006). In this regard, we stress that what constitutes fair or appropriate behavior often varies with cultural or situational factors (Henrich et al. 2001; Cappelen et al. 2007; Ellingsen et al. 2012): accounting for endogenous expectations is therefore key to a full understanding of social norms.

The remainder of the chapter focuses on a few game-theoretic frameworks that allow for the above characterization of social norms. The chapter proceeds as follows: section 2 reviews theories that explicitly feature normative expectations; section 3 reviews models that link norms to some formal specification of social groups or categories; section 4 surveys relevant experimental evidence; and section 5 concludes.

<a>2.NORMATIVE EXPECTATIONS

Bicchieri (2006, p. 11) proposes a set of conditions for the existence of a social norm. Specifically, a social norm exists if two conditions are satisfied. First, an individual is aware that he or she is in a situation in which a particular rule of behavior applies ('contingency clause'). Second, an individual prefers to conform to that rule if ('conditional preference clause'):

- he or she holds the relevant (first-order) *empirical* beliefs, that is, he or she believes that sufficiently many others will conform to it;
- he or she holds the relevant *normative* expectations, that is, he or she believes that sufficiently many others believe he or she ought to conform to it.[3]

A social norm exists and is followed by a group, if the above conditions are satisfied and beliefs are correct for members of that group. When that is the case, Bicchieri (2006, p. 52) proposes a utility function that captures norm-driven preferences. Considering an n-player normal form game, let S_i denote the strategy set of player i and let $S_{-i} = \prod_{j \neq i} S_j$ denote the set of strategy profiles of players other than i (with generic member s_{-i}). A norm N_i is defined as a

correspondence from one player's expectation about the opponents' strategies to the 'strategies one ought to take'; that is, N_i: $L_{-i} \rightarrow S_i$, with $L_{-i} \subseteq S_{-i}$.[4] From the viewpoint of player j, a strategy profile $s = (s_j, s_{-j})$ is said to instantiate a norm for player j, if N_j is defined at s_{-j}. When that is the case, s is said to violate a norm if player j does not follow a recommendation, that is, if $s_j \neq N_j(s_{-j})$. Player i's utility function is a linear combination of i's material payoff $\pi_i(s)$ and a component that depends on norm compliance:

$$U_i(s) = \pi_i(s) - k_i \max_{s_{-j} \in L_{-j}} \max_{m \neq j} \{\pi_m(N_j(s_{-j}), s_{-j}) - \pi_m(s), 0\}, \quad \text{<rr>(9.1)</rr>}$$

where $k_i \geq 0$ represents i's sensitivity to the norm and j refers to the norm-violator (with $j = i$ or $j \neq i$). The norm-based component represents the maximum loss suffered by any norm-following player m, as a result of j's violation. Specifically, the first maximum operator ranges over all the strategy profiles that instantiate a norm for j, while the second maximum operator ranges over all players other than the norm-violator j. We note that Bicchieri's (2006) framework makes it possible for the experimenter to test whether subjects exhibit preferences for conformity to a social norm, however specified, given that the conditions for compliance with that norm are satisfied (that is, contingency, and preferences conditional on the relevant empirical beliefs and normative expectations).

Bicchieri and Sontuoso (2015) propose an application extending Bicchieri's framework to dynamic psychological games (Geanakoplos et al. 1989; Battigalli and Dufwenberg 2009). This application accounts for (non-degenerate) conjectures about the recommendations of alternative rules of behavior. In doing so, it allows for the above-mentioned conditions to be more explicitly reflected in the player's motivation. In what follows we draw on Bicchieri and

Sontuoso (2015). We begin by introducing some standard notation on extensive form games. Let $H\backslash Z$ denote the set of non-terminal histories (that is, sequences of actions), with generic member h. A node of the game tree is identified with the history h leading up to it. For each player i, let S_i denote the set of pure strategies of player i, with generic member $s_i = \left(a_{i,h}\right)_{h \in H_i}$: note that $a_{i,h}$ indicates the action that would be selected by strategy s_i if history h occurred, with H_i denoting the set of nodes where i is active. Material payoffs are defined by functions $m_i: Z \to \mathbb{R}$ for each player i, where Z denotes the set of terminal histories (which correspond to end-nodes), with generic member z. Let $z(s)$ denote a terminal history induced by strategy profile $s \in S$. Player i holds a system of conditional first-order beliefs α_i (that is, beliefs about the co-players' strategies conditional on h);[5] player i also holds a system of second-order beliefs β_i about the first-order belief system of each of the co-players. It is assumed that players' beliefs at different information sets must satisfy Bayes' rule and common knowledge of Bayesian updating.

A *behavioral rule*, or more concisely a rule, is defined as a correspondence that assigns to every non-terminal history $h \in H\backslash Z$ one or more elements from the available set of strategy profiles $S(h)$. That is, a rule serves as a navigation system recommending some patterns of behavior at each node of the game tree; these recommendations may reflect a general principle (see López-Pérez 2008 for a related definition of rule that is not embedded in a psychological game theory framework). For example, consider a rule that prescribes behavior minimizing payoff inequality. When we evaluate this rule at the initial history, the rule will recommend strategy profiles that minimize any difference in payoffs among players, considering that every terminal node can be reached. Note that, if a deviation occurs, when we evaluate this rule at a history following such a deviation, the rule will make a recommendation that minimizes payoff inequality conditional on the terminal nodes that can still be reached.

Assume that players are aware of multiple rules of behavior, potentially making different recommendations (for example, think of rules reflecting principles of equity, efficiency, reciprocity, and so on). Consider all the recommendations made by every rule at the initial history: we refer to the set of actions contained in those recommendations as 'rule-complying actions'. Given that, we define a *norm-conjecture* as a collection of independent probability measures $\rho_i = \left(\rho_i(\cdot\,|h)\right)_{h\in H\setminus Z}$, such that the support of $\rho_i(\cdot\,|h)$ is a weak subset of the rule-complying actions.[6] That is, ρ_i represents a conjecture about the actions that might be considered appropriate for the current play of the game. More explicitly, when facing a social dilemma, players consider alternative patterns of behavior (as recommended by the rules they are aware of), and assign positive probability to the actions that may be appropriate for the current play of the game.

A norm-driven individual i is modeled as a player whose expected utility function is a linear combination of his or her material payoff and a component representing some anticipated negative emotion; that is, a function of the sum of losses that players (j) other than i would suffer because of a rule violation. To calculate such potential losses, we need to define player j's expectation of his or her material payoff, given strategy s_j and initial belief $\alpha_j = (\cdot\,|h^0)$ about the strategies of the co-players: drawing on the Battigalli and Dufwenberg (2007) concept of simple guilt, this expectation is given by $\mathrm{E}_{s_j,\alpha_j}\left[m_j|h^0\right] = \sum_{s_{-j}} \alpha_j\left(s_{-j}\middle|h^0\right) m_j\left(z\left(s_j, s_{-j}\right)\right)$. (Note that $\mathrm{E}_{s_j,\alpha_j}\left[m_j|h^0\right]$ is the *expected value* of m_j, calculated with respect to s_j and α_j at the initial history.) Bicchieri and Sontuoso (2015) assume that if player $j \neq i$ expects his or her co-players to follow some rule, then j will derive his or her first-order belief α_j from norm-conjecture ρ_j (player i will in turn estimate α_j on the basis of ρ_i).[7] That is, players will make use of the rules

they are aware of to form their first- and second-order beliefs, and accordingly calculate expected payoffs.

Formally, a norm-driven individual has conditionally conformist preferences characterized by a utility function of the form

$$u_i\left(z, s_{-i}, \alpha_j\right) = m_i(z) - k_i d_i^E \left(1 + \sum_{j \neq i} max\left\{0, \mathrm{E}_{\rho_i, s_j, \alpha_j}\left[m_j | h^0\right] - m_j(z)\right\}\right), \quad \text{<rr>(9.2)</rr>}$$

where $k_i \geq 0$ represents player i's sensitivity to the presumed norm, and d_i^E is a dummy variable equal to one if i believes that every $j \neq i$ will follow some rule (equal to zero otherwise). We note that if $k_i = 0$ or $d_i^E = 0$ the norm-based component vanishes, and the utility function reduces to the standard material payoff; otherwise, the norm-based component is a function of any positive difference between the initially expected payoff to j and the payoff j would get in the event of a rule violation.[8] In summary, we note that Bicchieri and Sontuoso's (2015) framework explicitly accounts for players' reasoning as to what constitutes appropriate behavior in the current play of the game; in doing so, it represents a social norm as a group-specific solution to a (dynamic) game.

We conclude this section by noting that an alternative approach to modeling norm-driven behavior has focused on the reputation-orientated signaling of prosocial acts, such as an individual's decision to contribute to a public good or share an asset (Benabou and Tirole 2006; Andreoni and Bernheim 2009). We note that these frameworks do not explicitly model normative expectations, but posit that an individual cares about the others' perception of his or her status (type) – as inferred from observed actions – under the assumption that acting altruistically is good.[9]

In particular, Benabou and Tirole (2006) model the individual's utility from contributing to a public good (which may be viewed as upholding an unwritten contribution rule) as a function of three components: intrinsic rewards (for example, joy of giving), material gain and reputational payoffs. The first two components vary across types, with types being private information. The third component (that is, reputational payoffs) depends on the observers' posterior expectations of the individual's type, in such a way that the individual would like to be perceived as public spirited rather than greedy (note that reputational payoffs can be affected by the level of observability of the actions). The model allows for multiple solutions to emerge as equilibria by virtue of the interplay of prosocial orientations, reputational concerns and the level of observability of the actions. For example, weakening the observers' ability to draw inferences about an individual's type may diminish the stigma associated with a deviation from the contribution rule. In summary, Benabou and Tirole's (2006) focus is on public goods games with reputational concerns: their model shows that different inferences (that is, different empirical beliefs about people's adherence to the contribution rule) may imply different equilibria.

<a>3.CATEGORY-SPECIFIC PRESCRIPTIONS

Theories of social identity partition the set of players into a number of social categories (for example, owner or employee, ingroup or outgroup members, and so on), and presume each category to be associated with a norm specifying the ideal behavior of a member of that category. It is often assumed that different contexts may trigger different identities (for example, work or family), hence different norms. These theories draw on the social-psychology notion of 'scripts' (Schank and Abelson 1977), that is, prescriptive sequences of actions that people automatically engage in, and are expected to engage in while in particular situations (see

Bicchieri 2016 for further discussion of the relationship between scripts and norms). In what follows we survey some of the models that account for category-specific prescriptions.

Akerlof and Kranton's (2000) seminal work introduces a generic utility function such that players' preferences depend on actions as well as on an identity-based component:

$$U_j = U_j\left(\boldsymbol{a}_j, \boldsymbol{a}_{-j}, I_j\right), \qquad \text{<rr>(9.3)</rr>}$$

where $\boldsymbol{a}_j$ and $\boldsymbol{a}_{-j}$ respectively denote the action vector of player j and that of j's co-players, while I_j denotes j's identity. The latter is defined as follows:

$$I_j = I_j\left(\boldsymbol{c}_j, \epsilon_j, \boldsymbol{P}, \boldsymbol{a}_j, \boldsymbol{a}_{-j}\right), \qquad \text{<rr>(9.4)</rr>}$$

where $\boldsymbol{c}_j$ denotes player j's categorization, ϵ_j denotes j's given characteristics, and $\boldsymbol{P}$ denotes prescriptions (that is, ideal behaviors and characteristics of different categories). Specifically, $\boldsymbol{c}_j$ is player j's assignment of individuals (including herself) to categories: this is interpreted as the player's conception of his or her own categories and those of others; for example, given two players and two genders, one such assignment may be $\boldsymbol{c}_j = \left((player\ 1, \text{woman}), (player\ 2, \text{man})\right)$. Now, j's identity I_j depends on $\boldsymbol{c}_j$ and on the degree to which j's own characteristics ϵ_j match the ideal (as indicated by prescription $\boldsymbol{P}$) of j's self-assigned category. Identity I_j further depends on the degree to which players' actions $\boldsymbol{a}_j, \boldsymbol{a}_{-j}$ match the ideal behavior (as per prescription $\boldsymbol{P}$). Akerlof and Kranton assume that each player j chooses actions $\boldsymbol{a}_j$ to maximize utility (9.3), taking as given $\boldsymbol{c}_j, \epsilon_j, \boldsymbol{P}, \boldsymbol{a}_{-j}$ (the authors informally

acknowledge that individuals may more or less consciously affect $\boldsymbol{c}_j, \epsilon_j, \boldsymbol{P}$ through their actions). Akerlof and Kranton (2000) go on to discuss a few examples in which there is an assignment of individuals to categories and a category-specific prescription affecting individual preferences (for example, think of gender discrimination in the workplace or household division of labor).

Benjamin et al. (2010) build on Akerlof and Kranton's work by proposing a specific functional form to capture identity-based preferences in decision problems. Suppose that an individual belongs to some social category C, and has to choose an action $x \in X$ (more generally, x may represent a vector of actions). Let x_0 and x_C respectively denote the individual's privately preferred action (regardless of identity considerations) and the action prescribed by her category. The individual's utility is given by:

$$U = -\big(1 - w(s)\big)(x - x_0)^2 - w(s)(x - x_C)^2, \qquad \text{<rr>(9.5)</rr>}$$

where $w(s)$ is the weight the individual puts on conforming with x_C, with s indicating the strength of the individual's affiliation with her category. That is, the individual suffers a weighted disutility when departing from the privately preferred choice x_0 or the prescription x_C. Benjamin et al. (2010) assume that the weight $w(s)$ of a zero-strength category is nil, and that the disutility of deviating from a prescription increases with the strength s of the affiliation. Then, the individual chooses x to maximize his or her utility for a given value of s. The first-order condition gives the optimal decision as:

$$x^*(s) = \big(1 - w(s)\big)x_0 + \ w(s)x_C, \qquad \text{<rr>(9.6)</rr>}$$

which is just a weighted average of the privately preferred action and the prescription. Expression (9.6) shows that as the strength s of the affiliation with one's category increases, the optimal choice x^* gets closer to the category prescription x_C. Benjamin et al. (2010) go on to show how exogenously evoking a category (for example, in an experimental setting) affects an individual's strength of affiliation with that category, and hence the individual's decision.

We conclude this section by noting two related lines of research. The first is concerned with the individual's self-reputation, and hence characterizes identity-dependent behavior as a self-signaling game where an agent has to decide how much to invest in social assets (in such a way to achieve or preserve a favorable self-conception, under the assumption that investments are valuable or appropriate; Benabou and Tirole 2011). The second line of research broadly defines categories in terms of 'social distance' (Akerlof 1997; Durlauf 2004): this class of models may be interpreted as a special case of the identity-based frameworks surveyed previously, in that it explains group-specific behaviors as the result of an incentive to conform with individuals whose inherited 'social locations' are close. That is, these models do not assume that utility increases as someone conforms to some category-specific prescriptions, but rather increases as someone conforms to his or her neighbors.

<a>4.EXPERIMENTAL EVIDENCE

<b>4.1Normative Expectations

We now turn to survey some experimental findings. Norms of cooperation and punishment are thought to persist as a consequence of the internalization of a principle of conduct or may be enforced out of fear of social sanctions (Elster 1989; Henrich and Boyd 2001; Cialdini and

Goldstein 2004). In the following we focus on laboratory experiments that identify social norms by explicitly measuring normative expectations.

Xiao and Bicchieri (2010) designed an experiment to investigate the impact on trust games of two potentially applicable, but conflicting, principles of conduct, namely, equality and reciprocity. Note that the former can be broadly defined as a rule that recommends minimizing payoff differences, whereas the latter recommends taking a similar action to others (regardless of payoff considerations). Xiao and Bicchieri's experimental design involved two trust game variants: in the first, players started with equal endowments; in the second, the investor was endowed with twice the money that the trustee was given. In both games, the investor could choose to transfer a preset amount of money to the trustee or keep it all. Upon receiving the money, the trustee could in turn keep it or else transfer back some of it to the investor: in the equal endowment condition (baseline treatment), both equality and reciprocity dictate that the trustee transfer some money back to the investor; by contrast, in the unequal endowment condition (asymmetry treatment), equality and reciprocity dictate different actions as the trustee could guarantee payoff equality only by making a zero back-transfer. Xiao and Bicchieri elicited subjects' first- and second-order empirical beliefs ('how much do you think other participants in your role will transfer to their counterpart?' and 'what does your counterpart think you will do?') and normative expectations ('how much do you think your counterpart believes you should transfer to her?'). Xiao and Bicchieri's (2010) results show that a majority of trustees returned a positive amount whenever reciprocity would reduce payoff inequality (in the baseline treatment); by contrast, a majority of trustees did not reciprocate the investors' transfer when doing so would increase payoff inequality (in the asymmetry treatment). Moreover, investors correctly believed that less money would be returned in the asymmetry treatment than in the baseline treatment, and

most trustees correctly estimated investors' beliefs in both treatments. However, in the asymmetry treatment empirical beliefs and normative expectations conflicted: this highlights that when there is ambiguity as to which principle of conduct is in place, each subject will support the rule of behavior that favors him or her most.

To verify whether a social norm reflecting a principle of equality was effectively in place, in a follow-up study Bicchieri and Mercier (2013) asked third parties to judge the appropriateness of trustees' behavior. Results show that third parties sided with trustees in stating that, in the asymmetry treatment, it is more appropriate to guarantee payoff equality than reciprocate the counterpart's action. This provides evidence that, while upholding a (self-serving) principle of equality, trustees were ultimately following a shared norm.

Reuben and Riedl (2013) examine the enforcement of norms of contribution to public goods in homogeneous and heterogeneous groups, such as groups whose members vary in their endowment, contribution capacity, or marginal benefits. In particular, Reuben and Riedl are interested in the normative appeal of two potentially applicable rules: the efficiency rule (prescribing maximal contributions by all) and the class of relative contribution rules (prescribing a contribution that is fair relative to the contributions of others; for example, equality and equity rules). Reuben and Riedl's (2013) results show that, in the absence of punishment, no positive contribution norm emerged and all groups converged toward free-riding. By contrast, with punishment, contributions were consistent with the prescriptions of the efficiency rule in a significant subset of groups (irrespective of the type of group heterogeneity); in other groups, contributions were consistent with relative contribution rules. These results suggest that even in heterogeneous groups individuals can successfully enforce a contribution norm. Most notably, survey data involving third parties confirmed well-defined yet conflicting

normative views about the aforementioned contribution rules; that is, both efficiency and relative contribution rules are normatively appealing, and are indeed potential candidates for emerging as contribution norms in different groups.

Bicchieri and Chavez (2010) designed an experiment to investigate norm compliance in ultimatum games. Specifically, their experiment involved a variant of the ultimatum game whereby the proposer could choose one of the following three options: ($5, $5), ($8, $2), or Coin (in which case one of the other two allocations would be selected at random). This design allows for two plausible notions of fairness: as an equal outcome ($5, $5) or as a fair procedure (Coin).[10] Bicchieri and Chavez (2010) elicited subjects' normative expectations about the actions they thought would be considered fair by most participants: proposers and responders showed a remarkable degree of agreement in their notions of fairness, as most subjects believed that a majority of participants deemed both ($5, $5) and Coin to be appropriate. Further, Bicchieri and Chavez (2010) had subjects play three instances of the above ultimatum game under different information conditions. In the full information condition, all participants knew that the Coin option was available, and that responders would know if their respective proposer had chosen Coin. In the private information condition, responders did not know that Coin was available to proposers, and proposers were aware of responders' ignorance. In the limited information condition, participants knew that the Coin option was available, but responders would not be able to distinguish whether their respective proposer had implemented one of the two allocations directly or had chosen Coin instead. Bicchieri and Chavez's (2010) results show that when normative expectations supporting the Coin option were either absent (in the private condition) or could be defied without consequence (in the limited condition), the frequency of choice of (5, 5) and (8, 2), respectively, were considerably higher than those of Coin. Moreover, the frequency

of Coin choices was highest in the public information condition, where this option was common knowledge and its outcome transparent: this shows that there proposers followed the rule of behavior that favored them most, and that this rule was effectively a social norm. However, substantial norm evasion characterized proposers' behavior in the limited information condition, where (8, 2) was the most frequent choice.

In a subsequent study, Chavez and Bicchieri (2013) measured empirical beliefs and normative expectations (as well as behavior) of third parties who were given the opportunity to add to or deduct from the payoffs of subjects who had participated in an ultimatum game. Third parties tended to reward subjects involved in equal allocations and to compensate victims of unfair allocations (instead of punishing unfair behavior); however, third parties were willing to punish when compensation was not an available option. Chavez and Bicchieri's (2013) results further show that third parties shared a notion of fairness (as indicated by their normative expectations), and that this notion was sensitive to contextual differences.

We move on to note Krupka and Weber's (2013) procedure for identifying social norms by means of coordination games. Using alternative (between-subjects) variants of the dictator game, Krupka and Weber had participants assess the extent to which different actions were collectively perceived as socially appropriate: subjects providing these ratings effectively faced a coordination game, as they were incentivized to match the modal response given by others in the same situation (this coordination game was intended to verify the presence of shared normative expectations). Krupka and Weber (2013) went on to use these elicited assessments to predict other subjects' compliance with the relevant social norm in each dictator game variant (for another application of the same elicitation procedure, see Gächter et al. 2013).

In this connection, we turn to present Schram and Charness's (2015) proposed procedure for inducing a shared understanding of the relevant rule of behavior, in the laboratory. Schram and Charness had participants in dictator games receive advice from a group of third parties. The information received simply revealed what a group of uninvolved subjects thought dictators ought to do: the information received generated an exogenous variation in the dictators' normative expectations. Schram and Charness's (2015) results show that choices are affected by this information.

Bicchieri and Xiao (2009) designed an experiment to investigate what happens when empirical beliefs and normative expectations conflict. Participants in a dictator game were exposed to different pieces of information. Specifically, two groups of dictators were given some descriptive information; that is, they were told what other subjects had done in another session (that is, one group was told that previous participants had made for the most part a generous offer, while the other group was told that most participants had made a selfish offer). Further, another two groups of dictators were given some normative information; that is, they were told what previous subjects said ought to be done (that is, one group was told that most previous participants thought that they should make a generous offer, while the other group was told that most participants thought that they should make a selfish offer). Other groups were given both descriptive and normative information. Bicchieri and Xiao's results show that, whenever this information did not conflict, both descriptive and normative messages had a significant influence on the dictators' own expectations and subsequent choices. When messages conflicted in that one indicated generosity and the other indicated selfishness, only the descriptive information affected dictators' behavior. This suggests that if people recognize that others are breaching the norm, then they will no longer feel compelled to follow the relevant rule of behavior themselves.

To conclude, the studies surveyed here provide evidence of the role played by normative expectations in affecting behavior in a variety of social dilemmas. In this regard, we note that in contrast to the vast literature on empirical beliefs, the number of studies that directly measure normative expectations is somewhat limited: more research is clearly needed to investigate the interplay of empirical and normative information about applicable rules of behavior.

<b>4.2Category-Specific Prescriptions

Research on the relationship between identity and norm-driven behavior typically evokes category-specific prescriptions by priming individuals for some, more or less arbitrary, identity. In this subsection we survey some of the recent, most relevant studies in the field (for some of the early experiments, see Tajfel and Turner 1986).

Cohn et al. (2015) designed a field study involving prison inmates. In a baseline treatment, subjects were asked to report the (privately observed) number of heads resulting from 10 tosses of a fair coin: inmates misreported 20 percent of the time. In a treatment in which subjects were primed for their criminal identity, inmates misreported even more (32 percent of the time). In another pair of treatments, subjects were recruited from the general population: in this case when primed for criminal identity (by reading a short text describing a criminal profile) the percentage of misreported heads fell slightly, relative to a control (from 14 percent to 10 percent). In brief, the experiment shows how priming criminal identity has a negative effect, on honest behavior, only in the case of inmates.

Benjamin et al. (2010) tested the predictions of their model of identity-dependent decision making by investigating subjects' preferences in the laboratory. To do so, before taking any decision, participants were provided with a set of questions that highlighted race or gender

identities: this questionnaire was intended to prime participants' social identities. Benjamin et al. (2010) went on to examine the marginal effect of category-specific prescriptions on discount rates and risk aversion (by measuring how subjects' choices change when an aspect of their social identity is made salient). When Asian-American subjects were primed for their ethnic identity they ended up making more patient choices, which the authors interpret in light of the presumed norm of patience characterizing the Asian identity. The same effect was observed in the case of African-Americans, who, when primed for race, additionally exhibited a higher degree of risk aversion (the latter effect is interpreted in light of the common credence in the hypothesis that racial risk norms depress native blacks' stock market participation). By contrast, primed and unprimed whites did not exhibit any difference; similarly, making gender identity salient had no effect on intertemporal or risk choices.

Charness et al. (2007) investigate the effect of group membership on behavior in the prisoner's dilemma and the battle of the sexes. Participants were divided into two groups, namely, row players and column players (thereby creating ingroup and outgroup categories for each participant). Subjects' payoffs consisted of two components: their own payoffs in periods (that is, games) in which they made an active decision, and a fraction of the payoffs of their ingroup members when the latter made active decisions. Charness et al.'s treatments revolved around two dimensions: audience and feedback. In the audience treatment ingroup, members directly watched the active player's decision (in a control, ingroup members did not watch), and in the feedback treatment ingroup, members immediately learned the outcome of each game (in a control, they learned it at the end of the whole session): both treatments are intended to make group membership more salient. Charness et al.'s (2007) results show that increasing the salience of an individual's group membership induces more aggressive behavior, that is, he or she is more

likely to defect in the prisoner's dilemma and play his or her favorite action in the battle of the sexes.

Chen and Li (2009) study the impact of social identity on other-regarding preferences in a series of two-player dynamic games. They designed an experiment in which subjects' group identity is induced using subjects' own preferences over different artworks; participants are then prompted to make some allocation decisions involving ingroup or outgroup co-players. Chen and Li's (2009) results show that when participants were matched with an ingroup member, there was a 47 percent increase in charity concerns and a 93 percent decrease in envy. Similarly, participants were 19 percent more likely to reward an ingroup member for good behavior, but 13 percent less likely to punish an ingroup member for misbehavior. Further, participants were significantly more likely to implement social-welfare maximizing allocations when matched with an ingroup member. Chen and Li's (2009) results are consistent with the intuitive hypothesis that participants are more altruistic toward an ingroup member (for related experiments on ingroup/outgroup effects, see Hargreaves Heap and Zizzo 2009 and Goette et al. 2012).

We conclude by noting that, while a review of the vast literature on stereotypes is beyond the scope of this chapter, there is ample evidence showing that priming different categories implies evoking different prescriptions. In particular, the evidence suggests that a subject is not bound by a rule of fairness when interacting with outgroup members (for field studies, see Fershtman and Gneezy 2001 and Bernhard et al. 2006); further, the evidence suggests that subjects believe that outgroup members are not bound by any rule either. In this connection, the literature on trust games has shown that if subject i belongs to a group whose members were untrustworthy, then participants from other groups will expect subject i to be untrustworthy as well (even when it is common information that group membership was assigned arbitrarily;

McEvily et al. 2006). These findings suggest that individuals may be hardwired, perhaps as a result of an evolutionary process, to take on social identities that are tied to specific (and appropriate) behavioral repertoires.

<a>5.CONCLUDING REMARKS

Social norms and social preferences have become an integral part of the economics discourse. After disentangling the two notions, this chapter has focused on a few formal accounts of social norms. Particular emphasis has been put on the role played by endogenous normative expectations in the context of (norm-driven) preferences for conformity. The relevant experimental literature has provided evidence in support of these accounts.

In summary, social preferences should not be conflated with social norms. Social preferences capture stable dispositions toward an exogenously defined principle of conduct. Specifically, social preference theories presuppose that players are hardwired with a notion of fair or kind behavior, as exogenously defined by the theorist; given that, players' preferences are assumed to be conditional on whether others will behave fairly (according to the exogenous principle) or not.

By contrast, social norms are better studied as group-specific, context-dependent solutions to strategic problems: more precisely, we have stipulated social norms as behavioral regularities that occur in mixed-motive games. Such solutions are the result of preferences for conformity conditional on an endogenous set of empirical beliefs and normative expectations.

We stress that different situational factors often come to be associated with different notions of appropriate behavior (think of contextual variables such as the framing and characteristics of the strategic problem, the role we are assigned, the social category with which

we identify, as well as historical and chance events). Accounting for endogenous normative expectations is therefore key to a full understanding of social interactions.

We note that our analysis has focused on the short run. The long-run study of norm formation and persistence necessitates an evolutionary account. While such an account would be beyond the scope of this survey, we note that a general implication of evolutionary models is that some norms are resilient to changing circumstances (Young 1998; Binmore 2005). Sometimes behavioral regularities come to be viewed as right and necessary simply because of their longevity, even if they originally emerged owing to contingencies or chance; thus, normative expectations often develop around long-established patterns of behavior, thereby reinforcing these behavioral regularities. More research, both theoretical and experimental, is needed to further illuminate the origin of normative expectations and their impact on social interactions.

<a>REFERENCES

Akerlof, G.A. (1980), 'A theory of social custom, of which unemployment may be one consequence', *Quarterly Journal of Economics*, **94** (4), 749–775.

Akerlof, G.A. (1997), 'Social distance and social decisions', *Econometrica*, **65** (5), 1005–27.

Akerlof, G.A. and R.E. Kranton (2000. 'Economics and identity', *Quarterly Journal of Economics*, **115** (3), 715–53.

Alesina, A. and P. Giuliano (2015), 'Culture and institutions', *Journal of Economic Literature*, **53** (4), 898–944.

Andreoni, J. and B.D. Bernheim. (2009), 'Social image and the 50–50 norm: a theoretical and experimental analysis of audience effects', *Econometrica*, **77** (5), 1607–36.

Arrow, K.J. (1972), 'Models of job discrimination', in A.H. Pascal (ed.), *Racial Discrimination in Economic Life*, Lexington, MA: Heath, pp. 83–102.

Battigalli, P. and M. Dufwenberg. (2007), 'Guilt in games', *American Economic Review: Papers and Proceedings*, **97** (2), 170–76.

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

Benabou, R. and J. Tirole (2006), 'Incentives and prosocial behavior', *American Economic Review*, **96** (5), 1652–78.

Benabou, R. and J. Tirole (2011), 'Identity, morals, and taboos: beliefs as assets', *Quarterly Journal of Economics*, **126** (2), 805–55.

Benjamin, D.J., J.J. Choi and A.J. Strickland (2010), 'Social identity and preferences', *American Economic Review*, **100** (4), 1913–28.

Bernhard, H., E. Fehr and U. Fischbacher (2006), 'Group affiliation and altruistic norm enforcement', *American Economic Review*, **96** (2), 217–21.

Bicchieri, C. (1993), *Rationality and Coordination*, Cambridge: Cambridge University Press.

Bicchieri, C. (2002), 'Covenants without swords: group identity, norms, and communication in social dilemmas', *Rationality and Society*, **14** (2), 192–228.

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

Bicchieri, C. (2016), *Norms in the Wild*, New York: Oxford University Press.

Bicchieri, C. and A.K. Chavez. (2010), 'Behaving as expected: public information and fairness norms', *Journal of Behavioral Decision Making*, **23** (2), 161–78.

Bicchieri, C. and H. Mercier (2013), 'Self-serving biases and public justifications in trust games' *Synthese*, **190** (5), 909–22.

Bicchieri, C. and A. Sontuoso (2015), 'I cannot cheat on you after we talk', in M. Peterson (ed.), *The Prisoner's Dilemma*, Cambridge: Cambridge University Press.

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

Binmore, K. (2005), *Natural Justice*, New York: Oxford University Press.

Binmore, K. (2010), 'Social norms or social preferences?', *Mind & Society*, **9** (2), 139–57.

Bolton, G.E. and A. Ockenfels (2000), 'ERC: a theory of equity, reciprocity, and competition', *American Economic Review*, **90** (1), 166–93.

Bolton, G.E., J. Brandts and A. Ockenfels (2005), 'Fair procedures: evidence from games involving lotteries', *Economic Journal*, **115** (506), 1054–76.

Bowles, S. (1998), 'Endogenous preferences: the cultural consequences of markets and other economic institutions', *Journal of Economic Literature*, **36** (1), 75–111.

Camerer, C.F. (2003), *Behavioral Game Theory. Experiments in Strategic Interaction*, Princeton, NJ: Princeton University Press.

Cappelen, A.W., A.D. Hole, E.Ø. Sørensen and B. Tungodden (2007), 'The pluralism of fairness ideals: an experimental approach', *American Economic Review*, **97** (3), 818–27.

Charness, G. and M. Rabin (2002), 'Understanding social preferences with simple tests', *Quarterly Journal of Economics*, **117** (3), 817–69.

Charness, G., L. Rigotti and A. Rustichini (2007), 'Individual behavior and group membership', *American Economic Review*, **97** (4), 1340–52.

Chavez, A.K. and C. Bicchieri (2013), 'Third-party sanctioning and compensation behavior: findings from the ultimatum game', *Journal of Economic Psychology*, **39** (December), 268–77.

Chen, Y. and S.X. Li (2009), 'Group identity and social preferences', *American Economic Review*, **99** (1), 431–57.

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

Cohn, A., M.A. Maréchal and T. Noll. (2015), 'Bad boys: how criminal identity salience affects rule violation', *Review of Economic Studies*, **82** (4), 1289–308.

Dhami, S. (2016), *The Foundations of Behavioral Economic Analysis*, New York: Oxford University Press.

Dufwenberg, M. and G. Kirchsteiger (2004), 'A theory of sequential reciprocity', *Games and Economic Behavior*, **47** (2), 268–98.

Durlauf, S.N. (2004), 'Neighborhood effects', J.V. Henderson and J.F. Thisse (eds), *Handbook of Regional and Urban Economics*, vol. 4, Amsterdam: Elsevier, pp. 2173–242.

Ellingsen, T., M. Johannesson, J. Mollerstrom and S. Munkhammar (2012), 'Social framing effects: preferences or beliefs?', *Games and Economic Behavior*, **76** (1), 117–30.

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

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

Fehr, E. and K.M. Schmidt (1999), 'A theory of fairness, competition, and cooperation', *Quarterly Journal of Economics*, **114** (3), 817–68.

Fehr, E. and K.M. Schmidt (2006), 'The economics of fairness, reciprocity and altruism – experimental evidence and new theories', in S.-C. Kolm and J. Mercier Ythier (eds), *Handbook of The Economics of Giving, Altruism and Reciprocity*, vol. 1, Amsterdam: North-Holland/Elsevier, pp. 615–91.

Fershtman, C. and U. Gneezy (2001), 'Discrimination in a segmented society: an experimental approach', *Quarterly Journal of Economics*, **116** (1), 351–77.

Gächter, S., Da. Nosenzo and M. Sefton (2013), 'Peer effects in pro-social behavior: social norms or social preferences?', *Journal of the European Economic Association*, **11** (3), 548–73.

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

Goette, L., D. Huffman and S. Meier (2012), 'The impact of social ties on group interactions: evidence from minimal groups and randomly assigned real groups', *American Economic Journal: Microeconomics*, **4** (1), 101–15.

Hargreaves Heap, S.P. and D.J. Zizzo (2009), 'The value of groups', *American Economic Review*, **99** (1), 295–323.

Henrich, J. and R. Boyd (2001), 'Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas', *Journal of Theoretical Biology*, **208** (1), 79–89.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis and R. McElreath (2001), 'In search of homo economicus: behavioral experiments in 15 small-scale societies', *American Economic Review*, **91** (2), 73–8.

Kagel, J.H. and A.E. Roth (2016), *Handbook of Experimental Economics*, vol. 2, Princeton, NJ: Princeton University Press.

Kolm, S.-C. and J. Mercier Ythier (2006), *Handbook of The Economics of Giving, Altruism and Reciprocity*, Amsterdam: North-Holland/Elsevier.

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

Ledyard, J. (1995), 'Public goods experiments', in J.H. Kagel and A.E. Roth (eds), *The Handbook of Experimental Economics*, Princeton, NJ: Princeton University Press, pp. 111–94.

Lewis, D. (1969), *Convention: A Philosophical Study*, Cambridge, MA: Harvard University Press.

López-Pérez, R. (2008), 'Aversion to norm-breaking: a model', *Games and Economic Behavior*, **64** (1), 237–67.

McEvily, B., R.A. Weber, C. Bicchieri and V.T. Ho (2006), 'Can groups be trusted? An experimental study of trust in collective entities', in R. Bachmann and A. Zaheer (eds), *Handbook of Trust Research*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 52–67.

North, D.C. (1991), 'Institutions', *Journal of Economic Perspectives*, **5** (1), 97–112.

Ostrom, E. (2000), 'Collective action and the evolution of social norms', *Journal of Economic Perspectives*, **14** (3), 137–58.

Rabin, M. (1993), 'Incorporating fairness into game theory and economics', *American Economic Review*, **83** (5), 1281–302.

Reuben, E. and A. Riedl. (2013), 'Enforcement of contribution norms in public good games with heterogeneous populations', *Games and Economic Behavior*, **77** (1), 122–37.

Schank, R.C. and R.P. Abelson (1977), *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*, Oxford: Lawrence Erlbaum.

Schelling, T.C. (1960), *The Strategy of Conflict*, Cambridge, MA: Harvard University Press.

Schram, A. and G. Charness (2015), 'Inducing social norms in laboratory allocation choices', *Management Science*, **61** (7), 1531–46.

Sontuoso, A. (2013), 'A dynamic model of belief-dependent conformity to social norms', MPRA Paper No. 53234, University Library of Munich.

Stigler, G.J. and G.S. Becker (1977), 'De gustibus non est disputandum', *American Economic Review*, **67** (2), 76–90.

Sugden, R. (1984), 'The supply of public goods through voluntary contributions', *Economic Journal*, **94** (376), 772–87.

Sugden, R. (2000), 'The motivating power of expectations', in J. Nida-Rümelin and W. Spohn (eds), *Rationality, Rules and Structure*, Amsterdam: Kluwer, pp. 103–29.

Sugden, R. (2005), *The Economics of Rights, Co-operation, and Welfare*, 2nd edn, Oxford: Basil Blackwell.

Tajfel, H. and J.C. Turner (1986), 'The social identity theory of inter group behavior', in S. Worchel and W.G. Austin (eds), *Psychology of Intergroup Relations*, 2nd edn, Chicago, IL: Nelson, pp. 7–24.

Xiao, E. and C. Bicchieri (2010), 'When equality trumps reciprocity', *Journal of Economic Psychology*, **31** (3), 456–70.

Young, H.P. (1998), *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*, Princeton, NJ: Princeton University Press.

Young, H.P. (2008), 'Social norms', in S.N. Durlauf and L.E. Blume (eds), *The New Palgrave Dictionary of Economics*, 2nd edn, London: Macmillan.

---

[1] Such intention-based theories of social preferences distinguish themselves from inequity aversion models, since they allow for beliefs to directly enter a player's utility function. The first contribution to the field of 'psychological game theory' is from Geanakoplos et al. (1989), while the first application is that of Rabin (1993). Battigalli and Dufwenberg (2009) have extended the analysis of psychological games to account for higher-order empirical beliefs; for an application, see Battigalli and Dufwenberg (2007).

[2] Rules guiding behavior in a coordination game (as opposed to a social dilemma) are usually regarded as conventions (Schelling 1960; Lewis 1969; Bicchieri 2006). For a short survey covering both social norms and conventions, see Young (2008).

[3] Normative expectations are second-order beliefs about appropriate behavior (that is, our belief about the 'personal normative beliefs' of others): from the viewpoint of a player, I believe that others believe that a certain behavior ought to be followed.

[4] For example, in an n-player prisoner's dilemma, a shared norm may be to cooperate: in that case, L_{-i} includes the cooperative strategies of all players other than i. Note that in cases where (given the others' strategies) there is not a prescription as to how player i should behave, N_i is not defined.

[5] For instance, in a game with perfect information, at each h player i holds an updated belief $\alpha_i(\cdot\,|h)$ so that he or she believes that players have taken all the actions leading to h with probability 1.

[6] Given a set of actions $A_i(h)$, the support of a probability measure $\rho_i(\cdot\,|h)$ on $A_i(h)$ consists of the actions that are assigned positive probability by ρ_i.

[7] That is, player i will derive his or her second-order belief β_i from norm-conjecture ρ_i. The reader can anticipate that in equilibrium $\rho_j = \rho_i$ for all j, i. In this connection, we may imagine that the rules an individual is aware of have been acquired through experience. When people have shared the same experiences, it is more likely that norm-conjectures coincide, and hence that first- and second-order beliefs are correct. See Sontuoso (2013) for a discussion of the relevant equilibrium notion.

[8] Note that the motivation captured by expression (9.2) is different from that of (9.1) in several respects. Among them, note that if no co-player j is harmed by i's deviation (that is, if $\sum_{j\neq i} max\left\{0, \mathrm{E}_{\rho_i,s_j,\alpha_j}[m_j|h^0] - m_j(z)\right\} = 0$), player i will still suffer a psychological disutility in the amount of $k_i \cdot 1$. This is thought to capture an intrinsic distaste for violating rules, regardless of the losses inflicted on others.

[9] When addressing social norms, Benabou and Tirole (2006, pp. 1665–9) consider a binary public goods game, where each agent can choose whether to contribute or not: the default rule of behavior is to contribute. Andreoni and Bernheim (2009) consider dictator game variants: the default code of conduct is the equality rule.

[10] See Bolton et al. (2005).