COVENANTS WITHOUT SWORDS

GROUP IDENTITY, NORMS, AND COMMUNICATION IN SOCIAL DILEMMAS

Cristina Bicchieri

ABSTRACT

In one-shot social dilemma experiments, cooperation rates dramatically increase if subjects are allowed to communicate before making a choice. There are two possible explanations for this 'communication effect'. One is that communication enhances group identity, the other is that communication elicits social norms. I discuss both views and argue in favor of a norm-based explanation.

KEY WORDS • communication • cooperation • group identity • social dilemmas • social norms

'Covenants without the sword are nothing but words.'

Thomas Hobbes

Introduction

A social dilemma is, by definition, a situation in which each group member gets a higher outcome if she pursues her individual interest, but everyone in the group is better off if all group members further the common interest. Overpopulation, pollution, Medicare, public television, and the depletion of scarce and valuable resources such as energy and fish-rich waters are all examples of situations in which the temptation to defect must be tempered by a concern with the public good. There are several reasons why some individuals might not contribute to the provision of public goods or refrain from wasting common resources. Usually these resources are used by or depend upon very large groups of people for their

continued maintenance. It is easy, therefore, for an individual to consider her contribution to a public good or her personal consumption of a common resource as insignificant. Furthermore, in social dilemmas there is a huge difference between the costs and benefits accruing to an individual. Gains go to the individual, but the costs are shared by all. Given the structure of social dilemmas, rational, self-interested individuals are predicted to always defect. Yet almost 50 years of experiments on social dilemmas show cooperation rates ranging from 40% to 60%, and everyday experience shows people making voluntary contributions to public goods, giving to charities, volunteering and refraining from wasting resources. One variation in social dilemma experiments, which dramatically increases cooperation rates, is allowing subjects to discuss the dilemma. There are two possible explanations for this 'communication effect'. One is that communication enhances group identity, the other that communication elicits social norms. I argue that the reason for this departure from normative rational choice is the working of norms; if correct, this conclusion has important strategic implications for institutional design and public policies that encourage social cooperation.

Experiments

To examine how group members make their decisions in social dilemmas, two different research paradigms are used. In a public goods dilemma, such as contributing to the maintenance of a public space or funding a public television service, individuals must contribute resources to insure the provision of the public good. Since one can enjoy public broadcast service television without making a financial contribution, groups run the risk that members will not contribute, and that the public good will not be provided at all. In a resource dilemma instead, groups share a scarce resource from which individual members can harvest, and the group runs the risk of excessive harvesting, leading to depletion of the resource. Examples of such resources are common grazing land and clean air.

A typical social dilemma experiment uses the mixed-motive structure of the prisoner's dilemma to study choice behavior. Like prisoner's dilemma games, both public goods and resource dilemmas have the property that the individual's rational choice is

always defection, but if all refuse to cooperate, all are worse off.[1] The usual experimental procedure involves subjects previously unknown to one another, who may receive a monetary payoff or points, and form one or two groupings depending on the experimental design. Subjects are given instructions and are presented with a payoff matrix describing the monetary consequences of their actions. It is individually best for each to keep his money or to appropriate a large amount of a common resource (to defect), but all are better off if everyone makes a cooperative decision to contribute to the public good or take little of the common resource. When two separate groups are formed, subjects are given the choice between allocating money to the ingroup or to the outgroup; if there is only one group, individuals must choose between giving money to their group or keeping it themselves.[2] Choices are made privately, interactions may be one-shot or repeated, and discussion before playing may or may not be allowed. I consider here mainly one-shot interactions, because repeated interactions allow opportunities for reciprocation or reputation formation. In a repeated game, it might work to the advantage of a rational, self-interested player to develop a reputation for being a 'nice guy'. In this case cooperation is not surprising, and it is easily explained by the traditional rational choice model. Only when there is no apparent incentive to cooperate does pro-social behavior become really interesting.

As an example of what experimental subjects may face, consider the following 'Give some' game (Dawes 1980), which is an example of a public goods dilemma. There are five players, and each receives US$8 from the experimenter. The choice is between keeping the money or giving it away, in which case every other player gets US$3. What a player gets depends on his choice and the choice of the other players:

Table 1
The 'Give Some' Game (Dawes 1980)

Number of givers	*Payoff to keep*	*Payoff to give*
5	–	$12
4	$20	$9
3	$17	$6
2	$14	$3
1	$11	$0
0	$8	–

The table shows that it is always better to keep the money, at least in terms of monetary payoffs, but the outcome of everyone giving is much better than the outcome of everyone keeping (US$8 versus US$12).

An example of a resource dilemma is the following 'Take some' game (Dawes 1980). There are three players and each has to decide whether to pick a red chip, in which case he gets US$3 and all three players are fined US$1, or pick a blue chip, in which case he gets US$1 and there is no fine. Again, the individual outcome depends upon one's choice, as well as the other players' choices:

Table 2
The 'Take Some' Game (Dawes 1980)

Number picking blue chip	*Payoff to red chip*	*Payoff to blue chip*
3	–	$1
2	$2	$0
1	$1	–$1
0	$0	–

In this situation, too, it is better to defect (hold the red chip), but the collective outcome of defection is worse than the cooperative outcome.

What we know from years of social dilemma experiments is that a significant baseline of cooperation is found in all experimental conditions, contrary to the prediction of rational choice theory. Even more interesting, we also know that in one-shot games allowing subjects a short period of communication about the dilemma increases cooperation well above the baseline. Indeed, a meta-analysis of social dilemma experiments conducted from 1958 to 1992 (Sally 1995) shows that the mean cooperation rate across conditions was 47.4%, that communication increased cooperation by 40% and commitment and promising increased cooperation by 30%. Similar conclusions are drawn by Gerry Mackie (1997), who summarized the results of several social dilemma experiments devoting particular attention to the role of communication and commitments. His conclusions can be thus summarized:

- discussion about the dilemma (but not 'irrelevant' discussion) increases cooperation rates;

- the primary content of discussions about the dilemma is promises and commitments to cooperate;
- to be effective, promising must be unanimous;
- overhearing spoken commitments from another group does not increase cooperation;
- when subjects are instructed that pledges are 'non-binding', they treat them as such and pledges have no effect on cooperation;
- commitments tend to be kept even if the beneficiary is a computer;
- commitments made on the initial belief of benefit to the ingroup tend to be kept when the locus of benefit unexpectedly switched to the outgroup (carryover effect);
- discussion improves contribution to a step-level public good even when it is confined to subgroups smaller than the critical number necessary to attain the cooperative payoff;
- cooperation declines over repetitions.

A number of suggestions have been advanced to explain the effectiveness of communication in increasing cooperation rates in one-shot games. For example, communication may help subjects to understand the game, facilitate coordinated action, alter expectations of others' behavior, promote group solidarity, elicit generic norms of cooperation or result in commitments to cooperate (Kerr et al. 1997). The question is still open, but among these suggestions, only group identity and social norms have not been eliminated as explanations by experimentation.

At the heart of the controversy between the group identity and social norms explanations of the effects of communication on cooperation rates lie two different views of the relation between an individual and the groups to which she belongs. In a reductionist perspective, the basic explanatory unit is the individual, and the group is just the aggregate of its members. Group behavior is thus explained in terms of properties of the individuals that make up the group. Individuals may be motivated by rational considerations, social norms, or be 'driven' to behave in given ways by automatic, unconscious processes. Communication in this view increases cooperation rates by making individuals focus upon particular social norms, such as the norm of promise-keeping. A holistic perspective instead views the group as a primitive, distinct explanatory unit. Group membership has important cognitive consequences as to how we perceive ourselves and others, how we process and filter

information and how we represent other collectives. Thinking of oneself as a group member causes major shifts in motives and behavior. A basic tenet of social identity theory is that individuals incorporate groups into their self-concepts, and this internalization precipitates motivational changes, so that often behavior contrary to self-interest is activated. As far as I know, few have tried to merge the two perspectives.[3] It is entirely possible, however, to view group identity as a trigger for norm-abiding behavior. When we represent a collection of individuals as a group, we immediately retrieve from memory roles and scripts that 'fit' the particular situation, and have a tendency to follow the appropriate social norms.[4] I shall return to this important point later.

Group Identity

Dawes, Orbell, and van de Kragt are among the leading proponents of the social identity explanation of cooperation in social dilemmas. They reasoned that if individuals incorporate groups into their self-concept, a motivational shift would occur, and group welfare would matter more than individual welfare. Orbell et al. (1988) detail two experiments designed to investigate the role of discussion in increasing cooperation rates via a group identity effect. During each session, multiple groups of 14 subjects were randomly divided into subgroups of seven persons each; afterwards, they went to separate rooms. Each subject was given a promissory note worth US$6, which he could keep or give away. If they chose to give the money away, six other subjects would each receive US$2. If everyone cooperated, each would get US$12. In half of the subgroups, contributions benefited six outgroup members, whereas in the remaining half, contributions benefited the other six ingroup members. Half of the subgroups could discuss the dilemma for ten minutes before playing. At the end of the discussion period, half of the subgroups who were allowed to discuss were informed that the beneficiaries of their contribution had changed. If subjects originally believed that their contributions would benefit the ingroup, they were now told that the outgroup would receive the money, and vice versa. All experimental discussions were taped, and I shall later examine them to argue that it is not group identity, but norms of promise-keeping, that explain the high rate of cooperation after a period of discussion.

Subjects contributed much more when both the dilemma was discussed and they initially believed that their contribution would go to the ingroup, as the following table shows:

Table 3
Subject Contributions

	Initial belief that money goes to ingroup		*Initial belief that money goes to outgroup*	
	Belief before decision		*Belief before decision*	
	Own	Other	Own	Other
No discussion	37.5%	30.4%	44.6%	19.6%
Discussion	78.6%	58.9%	32.1%	30.4%

Since increases in cooperation rates were not uniform across conditions, but appeared only when discussion of the dilemma was allowed, the authors reject the hypothesis that general norms of cooperation motivate contribution. If a general norm of cooperation were at work – they argued – subjects would not have discriminated between groups. Their conclusion is questionable. If norms are interpreted as generic imperatives, always readily available and invariably followed by those who hold them, then of course Orbell et al. are right. Norms, though, are often context-specific, and subjects have to be focused on them. The effect of discussion on cooperation rates might precisely be due to the fact that discussing the dilemma often involves an exchange of pledges and promises, and the very act of promising focuses subjects on a norm of promise-keeping.

Social norms can be thought of as default rules that are activated in the right circumstances.[5] More often than not the activation process is unconscious, it does not involve much thinking or even a choice on the part of subjects. We may thus expect that, once a norm has been activated, it will show some inertia, in the sense that unless a major change in circumstances occurs, people will keep following the norm that has been primed. This absence of fine-tuning might explain an interesting finding from this experiment: When a group initially believed themselves to be the beneficiaries of their contributions, but were subsequently told prior to their decision that the outgroup would benefit instead, 58.9% still cooperated.

This *carryover effect* of discussion suggests that cooperation results from a norm of promise-keeping. Such a norm would only become salient in the context of ingroup giving but, once activated, would show some inertia and still be followed even if the beneficiaries have changed. If instead the commitments and pledges exchanged during the discussion period were just contracts with particular people (the ingroup), then knowing that the money will go to the outgroup should decrease cooperation rates. Identification with one's own group may encourage cooperative behavior, but once it becomes apparent that the money would go to the outgroup, the motivation to give should disappear.

There is some other indirect evidence supporting a norm-based explanation. The carryover effect is also present in a very different experiment by Isaac and Walker (1988). In it, subjects played a two-period game with ten trials per period. The experiment had three conditions: 1) no discussion in either period; 2) no discussion in period one, but discussion in period two; 3) discussion in period one, but not in period two. The third condition was the only one in which cooperation was almost 100% in period one (after the discussion), and there was a carryover effect in the second period, since cooperation started at 100% but eventually decreased to 85%. Without discussion, cooperation usually started at 50%, but quickly declined to 10%. The data seem to indicate that groups quickly agree on a behavioral norm, which is followed through the trials. Note that in the second condition (no discussion in period one) initial low rates of cooperation carried over into the second period, and only increased towards the end of the next ten trials.[6] A plausible explanation is that subjects followed what they perceived as the group norm until it gradually became apparent that other group members were behaving differently and another rule was in place, which they themselves then adopted.

The purpose of the second experiment by Orbell et al. (1988) was to clarify the relationship between promise-making and cooperation. This time all groups of 14 subjects participated in an initial discussion of the dilemma. Afterwards they were divided into subgroups of seven as in the first experiment. Half of the subgroups were allowed to discuss the dilemma for another ten minutes. Subjects could make one of three possible choices: They could keep their US$5; they could give it to their ingroup, in which case the other six members would each receive US$2; they could give it to the outgroup, in which case all seven outgroup members would

receive US$3 each. Since the initial discussion took place before each group of 14 subjects was split into two subgroups, and the best choice for the whole group of 14 was to give to the outgroup, promises to cooperate were exchanged among all the participants, with the understanding that – once they were split into two subgroups – the money would go to the outgroup. To investigate the relationship between promise-making and cooperation, the experimenters stratified groups into three categories: 1) groups in which everyone promised to cooperate with the outgroup; 2) groups in which some promised to cooperate with the outgroup and others didn't; 3) groups in which subjects decided to make their own independent choices. In more than half of the groups there was unanimous promising, and in that case 84% cooperated with the outgroup. Without universal promising, cooperation was a meager 58%.

Though this second experiment led Orbell et al. (1988) to reject the hypothesis that higher rates of cooperation occurring after discussion are due to generic norms of cooperation, one cannot exclude the possibility that more specific norms are at work. The data indicate that individuals are more likely to cooperate when everyone in the group promises to cooperate, that is, when a consensus on how to behave is reached and an informal social contract is established. But, one might argue, if a specific norm of promise-keeping is responsible for cooperative behavior, we should observe a linear relationship between the number of subjects who promise and the number of cooperators in each group, and no such relationship is shown by the data. This objection presupposes that the norm of promise-keeping is a personal, unconditional norm, since in the absence of external sanctions of any kind (choices are one-shot and anonymous) only a personal system of values would have sufficient motivational power to induce subjects to cooperate. Then if discussion is allowed and promises to cooperate are exchanged, those who promised should fulfill their obligations irrespective of how many others in the group promised. If the data show otherwise, cooperation cannot be imputed to the working of personal norms.[7]

The above mentioned objection presupposes an unduly restrictive view of how norms work. People may not have a personal norm prescribing a given behavior, yet they will display the behavior if a social norm encouraging it is made salient (Cialdini et al. 1990).[8] Not unlike Cialdini's littering experiments, unanimous promising points to a consensually held norm. It is both descriptive ('every-

body will cooperate') and injunctive ('it's the right thing to do'), and as such elicits conformity. Less than unanimous promising does not induce complete defection, though. The data from Orbell et al. suggest that the rate of cooperation is not completely discontinuous, with high cooperation under unanimity and almost no cooperation otherwise. Rather, the number of cooperators has some relation (though not perfectly linear) with the extent of promising. There seems to be a critical point at which a majority of subjects will cooperate, and it is reached when a majority of group members promises to cooperate. Even without unanimity, there is still significant conformity to what might be perceived as a norm held by a sufficiently large number of group members. Orbell et al., however, maintain that discussion has an effect on cooperative behavior mainly because it *creates* group identity. Though the data do not refute their hypothesis, there are several difficulties with it. For one, it is never independently tested and, as we shall see momentarily, the very concept of group identity needs clarification. Furthermore, an analysis of the taped discussions that occurred in the Orbell, Dawes, and van de Kragt (1988) first experiment lends support to a norm-based explanation.

Cheap Talk

Though each group had a unique personality and discussion style, there are common themes and concerns that arose in almost all groups, which provide insights into the causes of cooperation.[9] Many groups had leaders who dominated the discussion. They advocated a particular strategy and asked the rest of the group to concur. In the absence of group leaders, subjects found it difficult to reach an agreement, and often opted to end their discussion period early. Recall that in the first experiment discussion took place *after* the two subgroups were formed, and subjects had to choose whether to keep their money or, depending upon the experimental condition, to give it to either the ingroup or the outgroup. The content of these discussions is quite different, though, depending on whether the potential beneficiary of the money is the ingroup or the outgroup.

Groups sometimes wanted to talk with the outgroup to check if they planned to cooperate. The implication seemed to be that – if

they were to make a commitment – they would be considered more trustworthy. The question of whether to trust the outgroup frequently arose, and those groups who initially thought of cooperating with the outgroup were worried about being cheated by them. Many groups concluded that most outgroup members would defect.[10] This conclusion was reached by *projection*: If we were in their place – it was argued – we would certainly defect. Group members evidently considered themselves to be a statistically representative sample; knowing their own propensity to defect led them to predict with some confidence the outgroup behavior. The predictability of the outgroup's behavior was grounded upon an expectation that they would behave 'normally', given the circumstances. Why would most groups consider defection on the part of the outgroup a normal choice?

It seems that competitiveness, mistrust, discrimination, and even aggression towards outgroups are deeply rooted attitudes, ready to emerge even in relatively neutral situations such as those encountered in experiments. In 1948, Sherif's Robber's Cave experiment, in which young boys selected for good psychological adjustment and sociability were separated into two rival groups, showed how quickly hostility and aggression can develop among groups that have no cultural or status differences between them. Tajfel's *Minimal group paradigm* (1973) is even more disturbing, as it shows how the mere grouping of individuals on the basis of arbitrary category differences is sufficient to produce group behavior. Ingroup favoritism, group loyalty, and a preference for group members are common effects of arbitrary categorization, as is the tendency to exaggerate the similarities with the ingroup and the differences with the outgroup. Note that these effects occur in situations in which subjects know almost nothing about other group members, apart from the fact that they all share a common group membership. For example, one may just know that one's group is made of 'overestimators of dots' as opposed to another group of 'underestimators of dots' (after having quickly judged how many dots there are on a wall screen).[11]

Precisely when there is only limited personal information on other subjects, categorization alone can generate impersonal attraction (or preference) for the other group members, as well as a sense of cohesion. This is a particularly interesting observation, since it has been commonly assumed that group cohesiveness is linked to the

degree of personal attraction among group members, as well as to how well the group satisfies individual needs. According to Tajfel's theory, group behavior is ultimately induced by a cognitive effect. The moment we think of ourselves as members of a group, however randomly determined, our perceptions and motives change. When more personal information is available, however, for example due to a longer period of interaction, attraction becomes less impersonal and group behavior is less likely to occur. In one-shot social dilemma experiments, where exposure to one's or another group is minimal, we should observe uncontaminated, basic group behavior such as loyalty and cooperation with one's group and mistrust and hostility toward the outgroup. Indeed, in 'two groups social dilemmas' (Bornstein 1992) subjects tended to support their own group to the detriment of the other group and ultimately of themselves.

In the taped discussions, when subjects were discussing with members of their group, and in situations in which ingroups benefited from their own decisions, commitments to cooperate with the ingroup were frequently made. This choice was often seen as a gamble, and as such involving risk. Discussion probably decreased the perceived risk of a monetary loss not just because one was able to assess the trustworthiness of other members by looking at their facial expressions and body language. An important reason why cooperation was perceived as less risky was the exchange of pledges and commitments that took place during discussion. Such commitments are, in economic parlance, just 'cheap talk'. In a one-shot interaction, given the assurance of anonymity, the temptation to defect is strong. In the absence of a binding mechanism, it may be to one's advantage to make a public pledge to cooperate, but then defect in private. Commitments and promises to the ingroup, however, were generally trusted. Is this an effect of categorization alone, or is it mediated by some implicit normative implication produced by categorization? We must not think of an experiment as an isolated, unique situation. How many times, in the course of our lives, have we made promises to people we know, to members of one group or another to which we belong? We usually keep our promises, and expect others to keep theirs. The experimental circumstances are similar, in several respects, to many real life situations subjects have experienced.[12]

Precisely because they do not know the other group members well, and have only limited exposure to them, subjects are free to

categorize their interaction as typical. In a typical group interaction, one would trust and cooperate with members of one's own group. The default presumption is that they will not cheat on us, that they will be nice and helpful. This may be the reason why betrayal by an acquaintance is much more devastating than betrayal by a stranger. We do not expect the first to occur. Thaler (1992) noted that well-established groups are often less cooperative than newly formed ones. If group identity were the ultimate cause of co-operation, we would expect much higher rates of cooperation in established groups. What may happen instead is that, after an initial period in which a newly formed group adopts cooperative norms by default, 'deviant' behavior may lead members to reconsider the context of interaction and their understanding of the situation, and possibly reach the conclusion that the dominant descriptive norm is to defect. Similarly, in repeated social dilemma trials with no communication it has been observed that cooperation rates are high in the initial periods, and then steadily decline over trials. This pattern is probably due to the fact that subjects are initially uncertain as to what constitutes appropriate behavior. Hence they rely on default injunctive norms they deem appropriate to the situation. If, as trials continue, some group members defect, cooperators will revise their expectations and start defecting, too.

Another belief shared by many subjects was that cooperating with the ingroup was not that risky.[13] When all group members committed to cooperate, subjects held the belief that at least half of them would keep their word. In this case, a cooperator would not lose her money. Many were even more optimistic, and voiced the belief that more than half of those promising to cooperate would keep their word. Notice that subjects did not naively expect everyone to keep their promise; rather, they realistically expected most people to keep their commitments most of the time, and in so doing they must have relied on a shared norm of promise-keeping. Unanimity therefore should not be interpreted as fostering the expectation of universal compliance, nor as an indication that everybody 'buys into the cooperative solution', thereby creating an obligation on the part of the promisor.[14] Most likely, unanimous promising signals that there is a consensus on the appropriateness of cooperation, and that the group is highly cohesive in its judgment. This high cohesiveness might in itself be sufficient to create strong conformity pressures.

Creating Identities

When it is suggested that solutions to social dilemmas may be facilitated by exploiting the solidarity and bonding arising from a shared group identity (Brewer 1979), a big open question remains to be answered. How can we arouse group identification in such a way that group interest is promoted? For the proponents of the social identity explanation, inducing a salient group identity will cause a blurring of the boundaries between personal and group welfare, a change in preferences and perception that is ultimately responsible for the increased rate of cooperation we witness after discussion of the dilemma. It is therefore important to know what makes group identity salient not just in an experimental context, but especially in the large, anonymous groups that are a common setting for social dilemmas.

There are some minimal conditions for a collection of individuals to constitute a psychological group – a state of affairs where they feel to be a group and act as one. A prominent traditional theory holds that a psychological group is a collection of individuals characterized by mutual attraction reflecting the members' interdependence and mutual need-satisfaction. This definition is severely limited, though, since it applies only to small groups, whereas some of our most important group memberships refer to large-scale social affiliations such as nationality, gender, race, religion, and so on. Members of a nation are not usually united around a single common goal, they interact only with small subsets of people and not always amicably, and obey different norms depending upon the organizations and subcultures to which they belong. National membership is not usually chosen, we are born into it, and the moments in which we are most likely to feel psychological membership are not ones in which our individual needs are satisfied. Indeed, our loyalty to our nation may be fiercest in circumstances, such as a war, that require sacrifice and deprivation. Similarly, the fact that some groups of people are treated in a homogeneous way by others due to the color of their skin, religious background, or otherwise, may give them a sense that they belong to a group, even if the grouping is not the result of their choice and membership into the group may involve discrimination and abuse by the rest of society. It is often reported that during the Nazi period, many German Jews felt for the first time an identification with their fellow Jews. They had been completely integrated and considered themselves to be Germans

first and foremost, but finding themselves associated with other European Jews in a common fate gave them, for the first time, a sense of their separate identity.

It is the realization that there can be psychological group membership without interdependence, need satisfaction, personal attraction, social structure or common norms and values that led Tajfel, and later Turner and Brewer, to design experiments in the context of the so-called *minimal group paradigm*. In these experiments, people were divided into distinct groups on the basis of random and meaningless criteria (such as estimation of the number of dots on a screen), group membership was anonymous and there were no group goals or any link between group membership and self-interest. I discussed some of these experiments elsewhere, observing how individuals systematically discriminate in favor of ingroup and against outgroup members.[15] The data collected by Tajfel and his colleagues imply that group behavior and group membership can exist in the absence of any social contact, social structure or interdependence between members. It was concluded that the minimal (sufficient) condition for psychological group formation is the recognition and acceptance of some self-defining social categorization. Social interaction, common fate, proximity, similarity, common goals or shared threats are not necessary for group formation, even if they usually increase the cohesiveness of an existing group. It is an open question whether they can be sufficient conditions for group formation, in the absence of an explicit categorization of people into groups. Presumably the answer will lie in assessing how efficiently and under which conditions such variables function as cues to the formation of social categorizations.

Group behavior, as opposed to individual behavior, is characterized by distinctive features such as perceived similarity between group members, cohesiveness, the tendency to cooperate to achieve common goals, shared attitudes and beliefs and conformity to group norms. If social categorization is sufficient for group formation, by which mechanisms does it produce group behavior? According to Turner's *self-categorization theory* (1987), group behavior depends upon the effects of social categorization on the definition and perception of the self. Self-perception, or self-definition, is defined as a system of cognitive self-schemata that filter and process information, and output a representation of the social situation that guides the choice of appropriate behavior. This system has at least two major components, social and personal identity. Social identity

refers to self-descriptions related to group memberships. Personal identity refers to more personal self-descriptions, such as individual character traits, abilities, and tastes.

Though personal and social identity are mutually exclusive levels of self-definition, this distinction must be taken as an approximation. There are many interconnections between social and personal identity, and even personal identity has a social component. It is, however, important to recognize that sometimes we perceive ourselves primarily in terms of our relevant group memberships rather than as differentiated, unique individuals. Depending on the situation, personal or group identity will become salient.[16] For example, when one makes interpersonal comparisons between self and other group members, personal identity will become salient, whereas group identity will be salient in situations in which one's group is compared to another group. Within a group, all those factors that lead members to categorize themselves as different and endowed with special characteristics and traits are enhancing personal identity. If a group is solving a common task, but each member will be rewarded according to his contribution, personal abilities are highlighted and individuals will perceive themselves as unique and different from the rest of the group. Conversely, if the reward for a jointly performed task is equally shared by all group members, group identification is going to be enhanced. When the difference between self and fellow group members is accentuated, we are likely to observe selfish motives and self-favoritism against other group members. When instead group identification is enhanced, ingroup favoritism against outgroup members will be activated, as well as behavior contrary to self-interest.

According to Turner, social identity is basically a *cognitive mechanism* whose adaptive function is to make group behavior possible. Whenever social identification becomes salient, a cognitive mechanism of categorization is activated that produces perceptual and behavioral changes. For example, the category 'Asian student' is associated with a cluster of behaviors, personality traits, and values. We often think of Asian students as respectful, diligent, disciplined, and especially good with technical subjects. When thinking of an Asian student solely in terms of her group membership, we attribute to her the stereotypical characteristics associated with her group, so she becomes interchangeable with other group members. When we perceive people in terms of stereotypes, we depersonalize them and see them as 'typical' members of their group. The same

process is at work when we perceive ourselves as group members. Self-stereotyping is a cognitive shift from perceiving oneself as unique and differentiated to perceiving oneself in terms of the attributes that characterize the group. It is this cognitive shift that mediates group behavior.

The feature of group behavior most relevant to social dilemma experiments is the tendency to cooperate with the ingroup even when such behavior is contrary to self-interest. Through common group membership, individuals share the same self-stereotypes, and perceive themselves as 'depersonalized' and similar to other group members in the stereotypical dimensions linked to the relevant social categorization. Insofar as group members perceive their interests and goals as identical – because such interests and goals are stereotypical attributes of the group – self-stereotyping will induce a group member to embrace such interests and goals as his own, and act to further them. The dark side of this process is the shared perception of group members that their interests are in conflict with those of other groups or of unaffiliated individuals. A prediction of social identity theory is thus that the more salient group membership becomes, the greater will be the tendency to display cooperative behavior toward the ingroup and discrimination against outgroups.

How can group identification be aroused in social dilemmas in such a way that cooperation is promoted? In a multi-trial commons dilemma, Kramer and Brewer (1984) showed that subgroup categorization of a six-person group decreased cooperation when compared with a condition in which the group was not subdivided.[17] Kramer and Brewer interpreted the result as an instance of ingroup favoritism and ingroup/outgroup competition: the defectors in the subgroup categorization condition wanted to gain as much as possible for their own subgroup in comparison with the other subgroup. However, if we examine the payoff structure it appears that the benefits of defection accrued only to the individual, not the subgroup, whereas the costs of defecting were spread out over the whole group. The choice was thus either to serve one's private interest (to defect) or to serve the interest of the whole six-person group (to cooperate). There was no possibility to differentially benefit one's own subgroup. Also, from the additional results of a questionnaire that was filled in after the experiment, it appears that categorization manipulation did not affect subjects' perceptions of the fellow subgroup members and of the members of the other

subgroup, contrary to the prediction of social identity theory. However, since subjects received feedback about the other group members' choices after each trial, they may have used this information in their post-trial perception ratings of the other group members, thus mitigating the effects of the induced categorization.

In a subsequent series of experiments, Brewer and Kramer (1986) showed that when the subgroup identity was made salient, and subjects received a feedback suggesting the existence of a descriptive group norm (the group could be made of 'high users', who took large amounts of common resources, or 'low users' who took small amounts), they tended to follow the group norm. When instead a collective identity was made salient, and it was clear that resources were rapidly dwindling, individuals belonging to groups of 'high users' restrained themselves most. It is not clear, however, that this behavior results from group identification. Subsequent analysis of subjects' expectations of other group members' behavior revealed no effect of categorization, nor was an ingroup bias apparent from the data. The abandonment of the 'high-use' subgroup norm in the superordinate identity condition may be due to a perceived conflict between a descriptive subgroup norm and an injunctive norm prescribing restraint. The superordinate identity could have made the injunctive norm salient, and we know from the work of Cialdini et al. (1990) that when there is a conflict between these two kinds of norms, and the injunctive one is made salient, people tend to follow the latter. The identity manipulation in this case would have mediated the effect of injunctive cooperative norms through the cognitive salience of group membership.

In a typical social dilemma experiment, there is no imposed or suggested categorization on the part of the experimenter. Subjects do not know each other and, in one-shot experiments, do not expect to play or meet again. The *minimal group paradigm* was successful in producing group behavior because it created an explicit ingroup/outgroup categorization that, even in the absence of conflicting interests, induced ingroup favoritism. In a typical social dilemma, however, the choice is between favoring oneself and favoring the group. We know that the mere realization that universal cooperation is in the group's interest does not induce cooperative behavior, but the social identity hypothesis predicts that making group membership salient will induce a cooperative orientation. Common fate, perceived similarities and verbal interactions, among other things, should contribute to the process of perceptual

group formation, inducing people to categorize themselves as part of a more inclusive unit. We would expect a period of discussion, especially on a theme close to the subjects' lives, to engender co-operative behavior, as would the experience of sharing a common fate. There is no apparent reason to expect discussion of the dilemma to be more efficacious than relevant discussion *per se*, or the experience of a common fate.

Keeping Promises

There is now a handful of experiments aimed at directly testing the group identity hypothesis in social dilemmas. None of them explicitly consider the possibility that social norms are responsible for the increase in cooperation rates observed after a period of discussion of the dilemma, though the data can be interpreted as supporting a norm-based explanation. Since the behavioral effects of group identity might be indistinguishable from the effects of other variables, such as perceived consensus or commitment, these studies introduced an independent measurement of group identity, defined as a sense of belongingness or a feeling of membership to a group.

Kerr and Kaufman-Gilliland (1994) used self-efficacy as a variable to differentiate between group identity and commitment explanations of the effect of communication on cooperation rates. They proposed a distinction between *cooperation-contingent* remedies and *public-good* remedies. The former increase the value one puts on the cooperative choice; they include side-payments, sanctions, and feelings like pride and guilt. The latter increase the value one puts on group's welfare, and they include altruism and enhanced group identity. They reasoned that if cooperation was motivated by a public good remedy, then as the efficacy of one's contribution declines, it becomes less likely that one cooperates. Since the group identity explanation of the effects of discussion assumes that communication works by increasing the value one puts on group welfare, discussion is a public good remedy. Hence, whenever it is evident that one's action is less efficacious, discussion should not be expected to matter much to one's choice. An explanation based on commitments instead assumes that discussion increases the value of the committed choice itself. In this case the

efficacy of one's action should not matter: committed subjects would cooperate no matter what.

The experiment consisted of groups of five subjects playing an 'investment game'. Each player was given US$10 and a point allocation. In each play, 100 points would be randomly assigned among the five players. Each player only knew her share, but the larger one's share, the more effectual one's choice would be in providing for the public good. If choosing to give, a player would donate US$10 plus her allocated points. If 51 or more points were contributed to a step-level public good, then each group member would obtain US$15. The game was to be played 16 times, and half of the subjects were allowed a period of discussion before making their (anonymous) choices.[18] The discussion effect was replicated, with 74.2% cooperation in groups that discussed, and only 56.8% cooperation in groups in which no discussion was allowed. Cooperation, however, was stable across levels of efficacy, suggesting that the perception of personal significance in providing for the public good was not an important factor in the choice to contribute. As in other experiments, group discussion contained frequent promises to cooperate, and groups varied in the agreements they reached. Some groups achieved unanimous promising, and in those groups cooperation rates were highest and the minimal efficacy level at which subjects were willing to cooperate was lower than in other groups. Some groups agreed to conditionally cooperate depending on each subject's level of efficacy, and other groups decided instead that each individual would make his or her own independent choice.

Different groups thus seemed to develop their own norms, such as 'contribute only if you have a reasonable share' or 'contribute no matter what'. Yet there was no apparent difference in the respective levels of perceived group identity, as measured by Hinkle et al. (1989) GIS (group identity scale). The conclusion drawn by the experimenters is that group identity is not a good explanation of discussion-induced cooperation. Commitments, and the norm of promise-keeping that supports them, are the most likely candidate. I must hasten to add that, though I sympathize with the conclusions, I find them too swift. The assumption that group identity entails the desire to enhance group welfare overlooks the possibility that many actions we take have also a 'symbolic' value. In some of Tajfel's experiments with allocations, if given the choice subjects tended to maximize the difference between ingroup and outgroup, and in so

doing were ready to sacrifice their own group's welfare. For example, between an equal allocation of US$10 to a member of each group and an allocation of US$6 to a member of one's own group and US$2 to an outgroup member, many subjects would choose the second. It is a choice that penalizes both groups, but hurts the outgroup more. If actions have a symbolic value for the actor, she might perform them irrespective of their efficacy.

A better way to test the group identity hypothesis is to check whether several presumably equivalent ways to create or enhance group identity produce the same results in terms of cooperation. The group identity explanation predicts that any manipulation arousing group identity will be sufficient to induce cooperation. Bouas and Komorita (1996) ran a series of experiments to test whether discussion or common fate would have an effect on cooperation rates. If discussion of the dilemma has an effect on cooperation, but discussion of an irrelevant topic has no effect (Dawes et al. 1977), we cannot rule out the group identity explanation, since an insignificant discussion topic may not be sufficient to elicit group identity. Discussing a relevant issue, such as an increase in students' tuition when the experimental subjects are college students, should instead create a bond among them, as this topic touches their lives and they can sympathize with each other's concerns. Another way to induce group identity is common fate. Common fate may not involve a common objective or shared needs. It may simply mean that certain categories of people are treated in a homogeneous manner by others on the basis of their sex, color of skin, language, and many other attributes. And it may be as tenuous as participating to a lottery that will determine the monetary worth of the points owned by each subject. Though participating to a common lottery does not strike me as a strong inducement to social identity formation, there is some evidence about its effects on cooperation rates (Kramer and Brewer 1984, 1986).

The alternative explanation of discussion-induced cooperation that Bouas and Komorita favor is not one based on norms. In their view, discussion has an effect because it creates consensus, and consequently reduces risk and fosters the expectation that other group members will cooperate. Since the only discussion that can create a meaningful consensus is discussion about the dilemma, the perceived consensus explanation predicts that only

discussion of the dilemma will increase cooperation rates. To compare the different predictions generated by the social identity and the perceived consensus explanations, it is helpful to draw the following tables:

Table 4
Group Identity Prediction

	Common fate	*No common fate*
Control condition	–	Defect
No discussion	**Cooperate**	Defect
Discussion of relevant topic	**Cooperate**	**Cooperate**
Discussion of dilemma	**Cooperate**	**Cooperate**

Table 5
Perceived Consensus Prediction

	Common fate	*No common fate*
Control condition	–	Defect
No discussion	Defect	Defect
Discussion of relevant topic	Defect	Defect
Discussion of dilemma	**Cooperate**	**Cooperate**

The experiment consisted of groups of four subjects facing a typical social dilemma and consisted of four conditions. A control condition in which there was no discussion nor common fate manipulation. A second condition in which subjects were allowed to discuss a relevant issue and were then exposed to a common fate manipulation. A third condition in which the dilemma was discussed and common fate was present. Finally, a common fate condition in which no discussion was allowed. The common fate manipulation meant that subjects' payoffs were determined by a lottery. In this experiment, too, group identity was independently measured through Hinkle et al.'s GIS, after the decisions were taken. Consensus perception and expectations of group members' cooperation were also independently measured. The results are reported in the following table:

Table 6
Experiment Results

	Control	*Common fate*	*Discussion*	*Discussion of dilemma*
Mean cooperation	0.13	0.13	0.17	**0.81**
Group identity[19]	5.10	5.10	6.10	6.30
Consensus perception	2.25	2.40	5.45	6.65
Expected cooperation	1.05	1.20	1.25	2.47

Whereas 81% of the subjects involved in discussion of the dilemma and common fate cooperated, only 17% did so after discussing a relevant issue (an increase in tuition), and common fate manipulation alone did not even raise cooperation rates above the baseline. Group identity, however, was higher in both discussion conditions, while common fate had no effect on group identity. What seemed to matter was perceived consensus, which was highest under discussion of the dilemma, but was also quite high in the relevant discussion condition. These results led Bouas and Komorita to reject the group identity explanation.

A few comments on common fate and the effect of perceived consensus are in order. Common fate is introduced here as a chance event (a lottery). As such, it has no effect on cooperation rates. However, Kramer and Brewer (1984) claimed that common fate induces group identity when such identity is superimposed on a pre-existing subgroup identity. In this case, they show that common fate is in fact salient. If there is no prior group identity, perhaps the notion of common fate has to be strengthened to do its job. For example, it would be interesting to see what happens if common fate were to entail interdependence among the parties, as when subjects are involved in a common task, however briefly, before the social dilemma experiment proper.

Perceived consensus increased after discussing the dilemma, but it was also high when another relevant topic was discussed. Note that consensus is weaker than universal promising, in that it does not require unanimous agreement. Bouas and Komorita argue that perceived consensus is what causes greater expectations of cooperative behavior, hence it presumably lowers the risk of losing money. However, since discussion of the dilemma often entails promises to cooperate, we cannot rule out as an explanation of risk reduction

the expectation that others will keep their commitments because of a shared norm of promise-keeping. If norms were responsible for the increase in cooperation rates we observe after a period of discussion of the dilemma, the prediction of a norm-based explanation would be like the one in table (b). Since this prediction is fulfilled, we cannot exclude that it is norms, and not just perceived consensus, that cause higher cooperation rates. Moreover, perceived consensus in not in conflict with a norm-based explanation. Discussion of the dilemma, with the intervening exchange of promises and pledges, can trigger both a descriptive norm (we are all going to contribute) and an injunctive norm (one should always keep one's word). What is perceived is that the group reached a consensus on what the appropriate course of action should be. Reaching a consensus on, say, how unfair an increase in university tuition is does not increase cooperation rates. If consensus alone is not sufficient to motivate giving to the group, it must be that the norms activated during discussion are responsible for the perception of reduced risk that accompanies the expectation of cooperative behavior on the part of other group members.

Talking to Machines

Sometimes support for an hypothesis is found in unexpected places. The norm-based explanation I favor says that norms are like default rules that are triggered in the right circumstances but not otherwise. Since this process is largely unconscious, we do not expect individuals to be very discriminating or strategically oriented in their norm-following behavior (Bicchieri, 2000). Whenever a norm is made salient by the situation one is in, the first reaction is to follow the norm, unless something unexpected occurs that forces reconsideration and possibly reinterpretation of the situation. The field of human-computer interaction is particularly interesting in this respect since it studies, among other things, the reactions people have to various kinds of computer interfaces and the rules, if any, people adopt in interacting with computers.

Kiesler et al. (1996) examined human-computer interaction in a social dilemma experiment. Subjects were presented with an 'investment game', which was in fact a common prisoner's dilemma in which the choice to cooperate was dubbed 'project green', and the choice to defect was dubbed 'project blue', in order to devoid choices

of any evaluative undertone. Subjects played six rounds against one of the following types of partners: a human confederate, a computer who communicated through written text, a computer who communicated with speech, and a computer who communicated with a synthesized face and speech. Subjects knew whether the partner was a human or a computer.

The partner used the same strategy across conditions:

Round 1: the partner asks the subject to make a proposal and then cooperates
Round 2: the partner proposes cooperation and then cooperates
Round 3: there is no discussion and the partner cooperates
Round 4: the partner asks the subject to make a proposal and then defects
Round 5: the partner proposes cooperation and then cooperates
Round 6: there is no discussion and the partner defects

After each round, the choices of the players were revealed.

The results are surprising, since they show that discussion and commitment have a strong effect on cooperation, regardless of the nature of the discussion partner. In the first round, 80% of the subjects proposed cooperation to the human confederate, and 94% of them kept their commitment. In the same round, 59% of the subjects proposed cooperation to the computer, and 62% of them kept their commitment. Cooperation was consistently high in rounds 1, 2, 4 and 5, that is, when there was discussion with the partner. Even in round 5, after observing a defection in the preceding round, subjects were willing to cooperate with a partner who proposed cooperation. Evidently they trusted more their partner's willingness to cooperate than their previous experience, and this occurred even if the partner was a computer. In this case, discussion had the effect of discounting previous defection. In rounds 3 and 6, there was a sharp drop in cooperation under all conditions; these were the rounds in which there was no communication, hence no commitment to cooperate.

If group identity were elicited through discussion, we would have not observed a drop in cooperation in round 3, since previous discussion and commitments to cooperate should have carried over to this round. At the very least, we should have not observed a drop in cooperation with the human confederate, since identification with a computer may be harder than with another human being. The

results offer strong support for the hypothesis that discussion enhances cooperation rates because of its content: promises to cooperate are made, and subsequently kept. Individuals seem to adopt the same social rules to interact with computers as they do with other human beings, which lends credibility to the view that we are witnessing the operation of *default social rules*, which are situation-dependent and are made salient by environmental cues pointing to a particular interpretation of the circumstances and hence to appropriate behavior.[20] If commitments are pledges to behave in accordance with the object of the commitment, regardless to whom the commitment is made, we can easily explain the former results. When a powerful norm of promise-keeping is made salient, most individuals will obey it, whether the promisee is another person or a machine.

Cognitive Misers

To explain what happens in an experimental situation, and to assess the accuracy of the general conclusions we draw about behavior in social dilemmas, it helps to detail the cognitive processes and heuristics that result in a cooperative choice on the part of so many subjects. The processes we use to make inferences are far from ideal. Social inference is heavily schema-driven, we disregard regression effects and base-rate information, and are prone to perceive illusory correlations. We store information in long-term memory and retrieve schemata to interpret and understand our environment, as well as to make inferences, explain and predict others' behavior. Such schemata are cognitive structures that represent knowledge about people, events and the self (Schank and Abelson 1977). Most of the time, they work reasonably well, though they bias all aspects of information-processing and inference towards conservative, schema-confirming inferential practices. To apply schematic knowledge, one first needs to be able to categorize the person or situation one encounters as fitting a particular schema.

Categories can be better described as *fuzzy sets*, collections of instances that have a family resemblance and are organized around a prototype, a standard against which family resemblance is assessed and category membership decided (Rosch 1978). Whereas a prototype is an abstraction from many instances, or even an ideal category member, an exemplar is a specific instance one has encountered.

Categories may also be represented in terms of exemplars, and there is still uncertainty among social psychologists as to under which conditions we use prototypes versus exemplars (Fiske and Neuberg 1990). Be it as it may, once a person or situation is categorized, a schema is invoked. A new person or situation we encounter, however, possesses so many features that it is not obvious which of them will be used as a basis for categorization, and consequently which schema will apply. What determines which cues will be used as a basis for categorization and subsequent schema use?

Given our bounded memory and information-processing capabilities, we rely on cognitive short-cuts (heuristics) to reduce complex problem-solving to much simpler judgmental operations. Such heuristics are extremely useful in deciding how likely it is that a person or event is an instance of one category or another, and in the subsequent process of making inferences and drawing conclusions. Kahneman and Tversky (1973, 1974) have researched three main heuristics: representativeness, availability, and anchoring and adjustment. For example, when we assess social stimuli, looking for cues to identify them, our attention is often directed by the accessibility of categories we frequently use and are consistent with our current goals and expectations. Gender is a readily available category, and in fact it is easily primed and used to interpret a variety of behaviors.

When faced with an experimental setup, an individual will first search for cues to categorize, and thus interpret, the present situation as an instance of a well-known schema. For example, a new collection of individuals will be mentally compared to past groupings, and this comparison process will provide her with behavioral cues appropriate to the new situation. Categorizing a social situation as fitting a particular schema will typically elicit behavioral roles and norms. In similar, previously experienced contexts we had a role and expectations that we import into the new situation. For example, interpreting a situation as 'us' versus 'them', as it frequently occurs even in the *minimal group paradigm* studied by Tajfel, may activate interactive schemata that contain norms such as 'take care of one's own', which could explain the preferential treatment accorded to ingroup members.[21]

To compare the new experimental situation to stored prototypes or exemplars, we must search our memory. Cognitive heuristics are the short-cuts we adopt to accomplish this search in a fast and efficient way. For example, how likely is it that the current

experimental situation is an instance of a cooperative or a competitive interaction? To which extent the present situation is similar, or *represents*, a typical competitive or cooperative interaction? When a subject must choose between keeping the money or giving it to the ingroup or the outgroup, the way she represents the situation will influence her subsequent choice. Indeed, we know that expecting the outgroup to benefit from one's contribution consistently dampens the impulse to give, whereas if it is the ingroup who benefits from one's act of giving there is much more willingness to part with one's money. A period of discussion may work so well because – whenever promises are made – the situation is perceived as similar to many situations we have experienced in which we made promises, and typically when we make a promise we keep it.

The *availability* heuristic helps in inferring the likelihood of an event on the basis of how quickly instances come to mind. In taped discussions, many subjects expected promises to be kept most of the time. The norm of promise-keeping is highly available, and most of the time we and the people we know honor commitments. The fact that the experimental situation is strange and unnatural, since subjects are anonymous and there is no repetition, is overlooked in favor of an optimistic interpretation biased toward what we know well and have frequently experienced. Similarly, we know that in experiments in which subjects are given the choice to opt out, besides cooperating or defecting, cooperators typically decide to play and defectors instead opt out more frequently.[22] This happens because cooperators expect cooperation from their partners, and defectors instead expect defection. Having a cooperative orientation presumably makes cooperative behavior highly available, and increases the confidence that one will encounter kindred spirits.

When we categorize a situation as similar to previously encountered ones, we need a starting point, an *anchor* or initial standard that orients and directs our inferences. So for example inferences about other people are often anchored by beliefs about ourselves or the behavior of people we know well and interact frequently with. The importance of framing effects (Kahneman and Tversky 1973) in decision-making is due to the fact that how a situation is framed affects our reference point. It is well known, for example, that people are more likely to cooperate in resource rather than in public good dilemmas (Brewer and Kramer 1986). The experimental framing is different, and it activates different reference points. In a

commons dilemma, subjects start with no money or points; thus it is easy for them to refrain from taking too much of the common good. In a public good dilemma, they are initially given a sum of money, so they perceive themselves as owning a certain amount of wealth. Contributing to a public good is experienced as a loss, whereas in a commons dilemma refraining from taking from the common pool is perceived as foregoing a gain. In Ultimatum games, we know that the allocation of 'property rights' has a major effect on the amount of money offered, as well as on how much the responder is willing to accept (Camerer and Thaler 1995).

Cognitive heuristics help us in categorizing any new situation or person we encounter, and in searching for the appropriate schema. Depending on how a situation is interpreted, different schemata and thus different norms will be activated. Since our interpretation and understanding of a situation will depend both on a frame of reference and on past experience, different people may interpret the same situation differently. For example, we consistently find a minimum base-rate of cooperation in social dilemma experiments across a variety of conditions. For those inclined to cooperate, the experimental context probably makes salient other-regarding behavior, because it is perceived as relevantly similar to familiar situations in which solidarity with the group is demanded. Analogously to Cialdini's experiments, where there is a certain percentage of people who do not litter, no matter what, there is a certain percentage of people who cooperate in social dilemmas, even without expecting repeated interactions and in the absence of discussion. In the littering experiments, there was a sizable number of people who were ready to follow whatever descriptive norm was made salient and, in the presence of a conflict between a descriptive and an injunctive norm, were perceiving the latter as more salient.

Discussion of the dilemma may perform several functions, all important in increasing cooperation rates. When people face a new situation, they often turn to each other for cues as to how to interpret it. In this context, the role of a leader is substantial, since she provides an interpretation of the situation, or suggests a schema, that other group members can recognize as both familiar and relevant. Unanimous agreement on appropriate behavior is usually reached only with the help of leaders, who are instrumental in lending salience to a particular descriptive norm. Yet discussion also involves promises, and the act of promising has the effect of focusing people on the injunctive norm of promise-keeping by

representing the situation as an instance of situations we have experienced in the past, when we made commitments we usually honored and expected others to do likewise. Discussion of the dilemma – when successful – points to several norms at once: a descriptive cooperative norm that might come to be perceived as injunctive, or 'the right thing to do', and an injunctive norm of promise-keeping. Disentangling their respective effects on behavior is very difficult, and we will have to wait for further experiments to provide answers.

We already have, however, some scattered evidence hinting at the consequences of making salient descriptive norms in social dilemmas. Schroeder et al. (1983), for example, investigated the effects of observing the behavior of others in simulated social dilemmas. They found that subjects quickly conformed to the behavior of the interacting others, regardless of whether it was cooperation or defection. Pillutla and Chen (1999) found similar results in a study on the effects of context (economic or non-economic) and feedback on cooperative behavior. Information about the other members' behavior was the sole variable influencing cooperation rates. Similarly, Allison and Kerr (1994) found that individuals behaved consistently with the perceived group norm, which was inferred from information about past group performance. These data are interesting, since they contradict some other results that indicate how viewing or listening to other groups' taped pledges to cooperate had no effect on cooperation rates. Was the positive effect reported in the former studies due to the fact that subjects observed the behavior of their own group members? If so, individuals were conforming to what they perceived as a descriptive norm, or the 'normal' behavior of their group. Unanimous promising may play the same role as observing past behavior, indicating the group's convergence on a behavioral rule. Since conformity to norms is correlated with the perceived cohesiveness of the group, it should come as no surprise that only under unanimous promising do we observe almost universal cooperation.

Is an explanation in terms of norms incompatible with the social identity hypothesis? According to social identity theory, self-categorization entails perceiving oneself in terms of the group prototype and behaving in accordance with that. Though norms may regulate group members' behavior without being considered specific to the group, often groups develop their own special behavioral

norms. In that case, group members believe that certain patterns of behavior are unique to them, and use their distinctive norms to define group membership. Many close-knit groups, such as the Amish or the Hasidic Jews, enforce norms of separation proscribing marriage and intimate relationships with outsiders, as well as specific dress codes and a host of other prescriptive and proscriptive norms that make the group unique and differentiate it from outgroups.

Hogg and Turner (1987) called the process through which individuals come to conform to group norms *referent informational influence*. Group-specific norms have, among other things, the twofold function of minimizing perceived differences among group members and maximizing differences between the group and outsiders. Once formed, such norms are internalized as cognitive representations of appropriate behavior as a group member. Social identity is built around group characteristics and behavioral standards, hence any perceived lack of conformity to group norms is perceived as a threat to the legitimacy of the group. Self-categorization accentuates the similarities between one's behavior and that prescribed by the group norm, thus causing conformity as well as the disposition to control and punish ingroup members that transgress group norms. In this view, group norms are obeyed *because* one identifies with the group, and conformity is mediated by self-categorization as an ingroup member.

Experimental groups, however, have had no time to develop their unique norms, and even if successful discussion points to a descriptive norm (and perhaps a prescriptive norm, too) one can hardly claim that such norms are special to the group, make it unique, or differentiate it from other groups. At most, group identification will elicit generic norms favoring the group. To explain the biased allocation results he obtained after having grouped his subjects into different (but meaningless) categories, Tajfel concluded that in minimal groups there is a generic norm prescribing ingroup favoritism (Tajfel 1970). Similarly, there may be benevolence norms that prescribe cooperation and trust within a group. If so, we should observe higher cooperation rates in all circumstances in which group identity is salient, not just when discussion and unanimous promising occur. A norm-based explanation is independent of group identification, though it recognizes the importance of group identity in making certain group norms salient. Even without identifying with an ingroup, however, an individual may get sufficient

cues from the environment signaling that a descriptive and/or injunctive norm is in place. Most of the time, one does not need a conscious motivation to follow the norm. The mental process of categorization, searching for a script and finally following what appears to be the appropriate behavior is entirely automatic. According to this view, discussion of the dilemma, when it ensues in an agreement as to the appropriate behavior, signals what the descriptive norm is going to be, and offers enough cues to represent the situation as a familiar one in which promises are exchanged and kept, hence activating an injunctive norm of promise-keeping.

NOTES

I wish to thank Robyn Dawes, Peyton Young, Matteo Motterlini, and participants in workshops at the Center for Cognitive Science, University of Arizona; The Brookings Institution, Washington; the University of California, Berkeley Law School; and the Conference on Strategic Rationality in Economics, Associazione Sigismondo Malatesta, Italy.

1. Another class of social dilemmas is the so-called 'step-level' public goods problem in which, after a threshold number of participants is reached, the public good is provided. These dilemmas involve a coordination element, since less than the total number of participants is needed to provide the public good. Moreover, if one believes she is the critical person who will 'make or break' the public good, one has an incentive to cooperate. However, experiments with step-level public goods provision show that subjects behave as if they were involved in a pure social dilemma (Dawes et al. 1986).
2. I am referring to the experiments discussed in Orbell et al. (1988).
3. An example of such merging is found in Jetten et al. (1996).
4. For a discussion of scripts, see C. Bicchieri, 'Words and Deeds: a Focus Theory of Norms', in J. Nida-Rumelin and W. Spohn (eds.) *Rationality, Rules and Structure* (Kluwer, 2000). Hertel and Kerr (2001) have provided some evidence for the quick retrieval (via priming) of social norms.
5. See, for example, Bicchieri (1997, 2000).
6. In the first condition, cooperation in period one started at 50%, but then declined to 10%. In period two, it started at 40% and then declined to 0%. In the second condition, cooperation in period one went from an initial 50% to 10%. In period two, it started at 60% and then went to 90%. In the third condition, cooperation remained close to 100% in period one; in period two it went from an initial 100% to 85%. The second period of discussion had only a marginal effect.
7. I take personal norms to be unconditional, as opposed to social norms. The main difference between a social and a personal norm is that expectations of others' conformity play a crucial role in the former, much less so in the latter. There is a difference between conforming to a norm because one expects others to

conform, and conforming because one is convinced of its inherent value. In the first case, my preference for conformity is conditional upon expecting others to conform; in the second case, my preference for conforming is unconditional. Cf. Bicchieri (1990).

8. Bicchieri (2000) discusses the importance of making norms salient in particular situations.
9. I wish to thank Robyn Dawes for making the tapes available to me, as well as my student Colleen Baker for a careful analysis of their content. Colleen recorded, for each group, who spoke first and what he/she said. How the others responded and how many responded, whether there was unanimous agreement on the strategy proposed, and how the conclusion about the outgroup's expected behavior was reached.
10. For a discussion of how intergroup schemas that are based on learned expectations about the competitive nature of intergroup relations influence a group's assessment of the outgroup behavior and intentions, see Insko and Schopler (1987) and Schopler and Insko (1992).
11. There are several possible explanations for ingroup bias. Tajfel et al. (1971) originally proposed a generic social norm of group behavior, according to which people should treat ingroup members more favourably than outgroup members. Later, however, he favored a different explanation based on social identy (Tajfel 1982). He assumed that, since people are motivated to maintain a positive social identity, they tend to make their social group positively distinct from other groups. Recent experiments conducted by Yamagishi et al. (1999) lend support to a different explanation: Ingroup favoritism is based upon the expectation that favors made to ingroup members are more likely to be reciprocated than favors made to outgroup members. Expectations of generalized reciprocity seem to be based upon a 'generic norm' of group behavior. Such norm is, in turn, sustained by ingroup favoritism.
12. A similar argument in made by Hertel and Kerr (2001) in their study of how social norms that favor the ingroup are primed in the right circumstances.
13. I am referring here to the systematic analysis of taped discussions done by my student Collen Baker (cf. note 9).
14. Orbell et al. (1991), p. 121.
15. Cf. Bicchieri (2000).
16. Brewer (1991) has developed a theory of 'optimal distinctiveness' to explain under which conditions we make personal (or social) identity relevant.
17. The experimenters manipulated the salience of the collective or subgroup identity. In some conditions subjects were told the experimenters were interested in the choices of psychology students versus economics students, who were the remotely located members of the collective group. Such instructions aimed to elicit a subgroup identity. In other conditions, experimenters told students that they were interested in the decisions of students at their particular university versus students at other universities in order to elicit a collective group identity.
18. The experiment also tested anonymity conditions, showing that anonymity has no effect on the behavior of subjects. A later study by Kerr et al. (1997) extended the anonymity condition to the experimenters and also found it not to be a significant factor.

19. Social identity and perceived consensus were measured on a nine-point scale. Expectation of cooperation refers to the number of others (zero to three) expected to cooperate.
20. See C. Bicchieri (2000) for an extended discussion of the situation-dependence of social norms.
21. A similar argument is made in Hertel and Kerr (2001).
22. Cf. Orbell and Dawes (1993).

REFERENCES

Allison, S.T. and N.L. Kerr. 1994. 'Group Correspondence Biases and the Provision of Public Goods.' *Journal of Personality and Social Psychology* 21: 563–79.

Bartlett, F.C. 1932. *Remembering: a Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.

Bicchieri, C. 1990. 'Norms of Cooperation.' *Ethics* 100: 838–61.

Bicchieri, C. 1997. 'Learning to Cooperate.' In *The Dynamics of Norms*, eds. C. Bicchieri, R. Jeffrey and B. Skyrms. Cambridge: Cambridge University Press.

Bicchieri, C. 2000. 'Words and Deeds: a Focus Theory of Norms.' In *Rationality, Rules and Structure*, eds. J. Nida-Rumelin and W. Spohn. Dordrecht: Kluwer Academic Publishers.

Bornstein, G. 1992. 'The Free Rider Problem in Intergroup Conflicts Over Step-level and Continuous Public Goods.' *Journal of Personality and Social Psychology* 62: 597–602.

Bouas, K.S. and S.S. Komorita. 1996. 'Group Discussion and Cooperation in Social Dilemmas.' *Personality and Social Psychology Bulletin* 22: 1144–50.

Brewer, M. 1979. 'Ingroup Bias in the Minimal Intergroup Situation: A Cognitive Motivational Analysis.' *Psychological Bulletin* 86: 307–24

Brewer, M. 1991. 'The Social Self: On Being the Same and Different at the Same Time.' *Personality and Social Psychology Bulletin* 17: 475–82.

Brewer, M.B. and R.M. Kramer. 1986. 'Choice Behavior in Social Dilemmas: Effects of Social Identity, Group Size, and Decision Framing.' *Journal of Personality and Social Psychology* 50: 543–9.

Camerer, C. and R. Thaler. 1995. 'Anomalies: Ultimatums, Dictators and Manners.' *Journal of Economic Perspectives* 9: 209–19.

Cialdini, R., C. Kallgren and R. Reno. 1990. 'A Focus Theory of Normative Conduct: a Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior.' *Advances in Experimental Social Psychology* 24: 201–34.

Dawes, R. 1980. 'Social Dilemmas.' *Annual Review of Psychology* 31: 169–93.

Dawes, R., J. McTavish and H. Shaklee. 1977. 'Behavior, Communication, and Assumptions About Other People's Behavior in a Commons Dilemma Situation.' *Journal of Personality and Social Psychology* 35: 1–11.

Dawes, R., J. Orbell, R. Simmons and A. van de Kragt. 1986. 'Organizing Groups for Collective Action.' *American Political Science Review* 80: 1171–85.

Dawes, R., J. Orbell and A. van de Kragt. 1988. 'Not Me or Thee But We: The Importance of Group Identity in Eliciting Cooperation in Dilemma Situations.' *Acta Psychologica* 68: 83–97.

Dawes, R., A. van de Kragt and J. Orbell. 1990. 'Cooperation for the Benefit of Us, Not Me, or My Conscience.' In *Beyond Self-interest*, ed. J. Mansbridge. Chicago, IL: University of Chicago Press.

Estes W.K. 1986. 'Memory Storage and Retrieval Processes in Category Learning.' *Journal of Experimental Psychology* 115: 155–74.

Fiske, S.T. and S.L. Neuberg. 1990. 'A Continuum of Impression Formation, from Category-based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation.' In *Advances in Experimental Social Psychology*, vol. 23, ed. L. Berkowitz. New York: Academic Press.

Fiske, S.T. and S.E. Taylor. 1991. *Social Cognition*. New York: McGraw-Hill.

Hertel, G. and N. Kerr. 2001. 'Priming Ingroup Favoritism: The Impact of Normative Scripts in the Minimal Group Paradigm.' *Journal of Experimental Social Psychology* 37: 316–24.

Hinkle, S., L. Taylor, D. Fox-Cardamone and K. Crook. 1989. 'Intragroup Identification and Intergroup Differentiation: a Multi-Component Approach.' *British Journal of Social Psychology* 28: 305–17.

Hogg, M.A. and J.C. Turner. 1987. 'Social Identity and Conformity: a Theory of Referent Informational Influence.' In *Current Issues in European Social Psychology*, vol. 2, eds. W. Doise and S. Moscovici. Cambridge: Cambridge University Press.

Insko, C.A. and J. Schopler. 1987. 'Categorization, Competition, and Collectivity.' In *Review of Personality and Social Psychology: Group Processes*, vol. 8, ed. C. Hendrick. Beverly Hills, CA: Sage.

Issac, R. and J. Walker. 1988. 'Communication and Free-Riding Behavior: The Voluntary Contribution Mechanism.' *Economic Inquiry* 26: 585–608.

Jetten, J., R. Spears and A.S.R. Manstead. 1996. 'Intergroup Norms and Intergroup Discrimination: Distinctive Self-Categorization and Social Identity Effects.' *Journal of Personality and Social Psychology* 71: 1222–33.

Jones, E. and R. Nisbett. 1972. 'The Actor and the Observer: Divergent Perceptions of the Causes of Behavior'. In *Attribution: Perceiving the Causes of Behavior*, ed. E. Jones et al. Morristown, NJ: General Learning Press.

Kahneman, D. and A. Tversky. 1973. 'Availability: A Heuristic for Judging Frequency and Probability.' *Cognitive Psychology* 5: 207–32.

Kerr, N. and C. Kaufman-Gilliland. 1994. 'Communication, Commitment, and Cooperation in Social Dilemmas.' *Journal of Personality and Social Psychology* 66: 513–29.

Kerr, N., J. Garst, D.A. Lewandowski and S.E. Harris. 1997. 'The Still, Small Voice: Commitment to Cooperate as an Internalized versus a Social Norm.' *Personality and Social Psychology Bulletin* 23: 1300–11.

Kiesler, S., L. Sproull and K. Waters. 1996. 'A Prisoner's Dilemma Experiment on Cooperation with People and Human-like Computers.' *Journal of Personality and Social Psychology* 70: 47–65.

Kramer, R.M. and M.B. Brewer. 1984. 'Effects of Group Identity on Resource Use in a Simulated Commons Dilemma.' *Journal of Personality and Social Psychology* 46: 1044–57.

Kramer, R.M. and M.B. Brewer. 1986. 'Social Group Identity and the Emergence of Cooperation in Resource Conservation Dilemmas.' In *Psychology of Decisions and Conflict: Experimental Social Dilemmas*, eds. H.A. Wilke et al. Frankfurt: Verlag Peter Lang.

Mackie, G. 1997. '"Noncredible" Social Contracts Are Credible: Communication and Commitment in Social Dilemma Experiments.' Unpublished paper, Mimeo.

Nisbett, R., and T. Wilson. 1977. 'Telling More Than We Can Know: Verbal Reports on Mental Processes.' *Psychological Review* 84: 231–59.

Orbell, J., A. van de Kragt and R. Dawes. 1988. 'Explaining Discussion-Induced Cooperation.' *Journal of Personality and Social Psychology* 54: 811–19.

Orbell, J., A. van de Kragt and R. Dawes. 1991. 'Covenants Without the Sword: The Role of Promises in Social Dilemma Circumstances.' In *Social Norms and Economic Institutions*, eds. I.K. Koford and D.J. Miller. Ann Arbor, MI: University of Michigan Press.

Orbell, J. and R. Dawes. 1993. 'Social Welfare, Cooperators' Advantage, and the Option of Not Playing the Game.' *American Sociological Review* 58: 787–800.

Pillutla, M. and X.P. Chen. 1999. 'Social Norms and Cooperation: The Effects of Context and Feedback.' In *Organizational Behaviour and Human Decision Process*, Mimeo.

Rosch, E. 1978. 'Principles of Categorization'. In *Cognition and Categorization*, eds. E. Rosch and B. Lloyd. Hillsdale, NJ: Erlbaum.

Sally, D. 1995. 'Conversation and Cooperation in Social Dilemmas.' *Rationality and Society* 7: 58–92.

Schank, R. and R. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Hillsdale, NJ: Erlbaum.

Schopler, J. and C.A. Insko. 1992. 'The Discontinuity Effect in Interpersonal and Intergroup Relations: Generality and Mediation'. In *European Review of Social Psychology*, vol. 3, eds. W. Stroebe and M. Hewstone, pp. 121–51. New York: John Wiley.

Schroeder, D.A., T.D. Jensen, A.J. Reed, D.K. Sullivan and M. Schwab. 1983. 'The Actions of Others as Determinants of Behavior in Social Trap Situations.' *Journal of Experimental Social Psychology* 19: 522–39.

Sherif, M. 1966. *In Common Predicament*. Boston, MA: Houghton Mifflin.

Tajfel, H. 1973. 'The Roots of Prejudice: Cognitive Aspects.' In *Psychology and Race*, ed. P. Watson. Harmondsworth: Penguin.

Tajfel, H. 1970. 'Experiments in Itergroup Discrimination.' *Scientific American* 223: 96–102.

Tajfel, H., M. Billig, R. Bundy and C. Flament. 1971. 'Social Categorization in Intergroup Behavior.' *European Journal of Social Psychology* 1: 149–78.

Tajfel, H. 1982. 'Social Psychology of Intergroup Relations.' *Annual Review of Psychology* 33: 1–30.

Tetlock, P.E. and R. Boettger. 1989. 'Accountability: a Social Magnifier of the Dilution Effect'. *Journal of Personality and Social Psychology* 57: 388–98.

Thaler, R. 1992. *The Winner's Curse: Paradoxes and Anomalies in Economic Life*. New York: Free Press.

Turner, J.C. et al. 1987. *Rediscovering the Social Group: A Self-Categorization Theory*, Oxford: Blackwell.

Yamagishi, T., N. Jin and T. Kiyonari. 1999. 'Bounded Generalized Reciprocity: Ingroup Boasting and Ingroup Favoritism.' Advances in Group Processes 16: 161–97.

CRISTINA BICCHIERI is Professor of Philosophy and Social and Decision Sciences at Carnegie Mellon University. She has published widely in philosophy, sociology, political science and economics journals. She is the author of *Rationality and Coordination* (Cambridge University Press, 1993, 1997), and co-author of *The Dynamics of Norms* (Cambridge University Press, 1997) and *The Logic of Strategy* (Oxford University Press, 1999). Her current research interests are the emergence and dynamics of norms, social learning, and the foundations of game theory.

ADDRESS: Carnegie Mellon University, Pittsburgh, PA 15213, USA [email: cb36@andrew.cmu.edu]