From Self-Interest Motives to Justice Motives

The Challenges of Some Experimental Results

by HELENA LOPES*

ABSTRACT. This article begins by presenting experimental evidence that remains unexplained by standard and utility-extended economic models: experimental subjects tend to honor their promises even on occasions when an assessment of consequences asks them to defect; subjects voluntarily contribute to collective goods, and this contribution is highly conditional on others contributing as well; subjects evaluate and value the intentions behind actions as well as the consequences of actions. Arguments are sought for in moral philosophy that would more plainly explain the collected experimental evidence and that would help economists revise their explanatory frames. The hypothesis advanced is that the observed behavior may be interpreted as resulting from the moral strength of indignation and justice norms.

I

Introduction

IN THE 20TH CENTURY, rational choice theory replaced what was once a substantive concept of utility with purely formal requirements: the

*Helena Lopes is Associate Professor at ISCTE (Higher Institute of Social Sciences and Business Studies), Department of Economics, and researcher at DINAMIA, Lisbon, Portugal. Recent publications include *Utilité, normes et sentiments moraux*. This article was written in the context of the NORMEC (The Normative Dimensions of Action and Order), a 5th EU-FP project (ref. HPSE-CT2001-00081). The author greatly benefited from in-depth discussions with all team members, especially José Castro Caldas, João Rodrigues, Ana Costa, Ana Santos, and Luís F. Carvalho. Without them, this article would not have been possible. Much of Section II is based on Lopes, Caldas, Costa, and Rodrigues (2005) (unpublished ms). This collective paper presents the results of a set of experiments conducted in the frame of NORMEC, conceived and implemented by all team members. I am particularly grateful to all co-authors for allowing me to use this material here.

American Journal of Economics and Sociology, Vol. 67, No. 2 (April, 2008).

axiomatic framework of rationality. Action was henceforth conceived of as the product of the agent's beliefs and desires directed to securing what he or she most wants, and the issue of the motives for action was ruled out of the core of the theory. Statements about agents' preferences and beliefs, which were required to account for actual behavior, were given the status of auxiliary assumptions.

There is now a wide set of experimental results showing that, contrary to the predictions of standard rational choice theory, most individuals do not solely try to maximize their self-interest but instead behave as if they were guided by social or moral motives. In the light of these results, economists have been led to face the classical problems of moral philosophy, thus going back to the precise point where economics began in the 18th century. Hence, it might be interesting to find arguments in moral philosophy that would help economists to revise their explanatory frames.

This article begins by presenting the evidence, collected in an original set of experiments and reported in already vast experimental literature, that remains unexplained by conventional theory. The relevance of some economic models and of the kinds of motivations they put forward to account for the experimental findings is then examined in the light of moral arguments. We will advance the hypothesis that the experimental evidence collected to date may be interpreted as resulting from the moral strength of the sentiment of indignation and of justice norms.

II

Experimental Evidence: What is to be Explained?

A. The Puzzling Results of Public Goods Experiments

Public good provision experiments have been used to analyze the dilemmas of collective action. In these experiments, the goal, to be jointly achieved by the members of a group, is a public good defined by the nonrivalry and nonexcludability properties. Ideally, rational individuals are not expected to act consistently with the group's goals, preferring instead to *free-ride* on the contributions of others, since this is the strategy that maximizes their individual payoff.

In recent years, however, an impressive amount of evidence from experimental and field studies has challenged this prediction. Surveys of experimental studies on public good provision (Ledyard 1995; Ostrom 2000; Camerer and Fehr 2002) all point to the following stylized findings (Caldas et al. 2004):

(a) Subjects contribute considerable amounts of their endowment to the public good in the one-shot game.
(b) In repeated games, the level of contribution is high in the first round, but contributions steadily unravel over time.
(c) The individual's contribution to collective goods is conditional on others contributing as well. Contribution levels are highly dependent on whether subjects behaved cooperatively or uncooperatively early on (Kahan 2005).
(d) Opportunities for punishment or retaliation strongly increase the contributive levels. When such opportunities exist, most subjects tend to use them even at a cost to themselves. These penalties effectively deter *free-riding* and make contributions remain at a high and steady level (Fehr and Gachter 2000).
(e) Communication among experimental subjects strongly increases the contributive rates (Ledyard 1995).

Public good experiments are interesting to economists because the subjects' behavior reflects the behavioral patterns observed under many "real-world" circumstances when individual and collective interests appear to conflict (see Gneezy and Rustichini 2000; Kahan 2005; and Frey and Meir 2004 for reports of field experiments). Thus, experimental findings should not be viewed as relevant only within the laboratory walls. Even if the operation of inferring motives from behavior is always a delicate one, experimental evidence provides a valuable informative framework for examining the nature and motives of economically relevant behavior.

B. Evidence of Justice-Driven Behavior: An Experimental Exploration

Experimental studies based on ultimatum, dictator, gift exchange, and trust games highlighted the motivational strength of justice concerns. In bilateral bargaining situations—like ultimatum games—

anonymously interacting agents frequently agree on rather egalitarian outcomes while conventional game theory predicts strongly unequal outcomes. In competitive experimental labor markets with incomplete contracts, fairness considerations give rise to efficiency (higher) wages that deviate from the competitive outcome (Falk, Fehr, and Fischbacher 2003). In trust games, Players 1 often deviate from the subgame perfect strategy in order to achieve a future higher benefit for both and, in response, Players 2 often reciprocate positively instead of playing their dominant strategy. Both, therefore, incur an opportunity cost in order to arrive at the joint benefit (McCabe et al. 2003).

Until now, fairness concerns have been experimentally assessed by economists only in dyadic environments. A set of experiments was therefore conceived (Lopes et al. 2004)[1] that introduces different distributive rules in a public good experiment in order to try to isolate justice motives in multi-agent contexts.

Roth (1995: 22) identified three types of motivations to run experiments: (1) "Speaking to Theorists," experiments designed to test predictions of well-articulated formal theories; (2) "Searching for Facts," experiments designed to study the effects of variables about which existing theories have little to say; and (3) "Whispering in the Ears of Princes," experiments related to the dialogue between experimenters and policymakers. The experiments reported here fit under "Searching for Facts." They are not aimed at testing any theory or hypothesis but rather at observing behavior in conditions difficult to isolate in everyday life.

The study included four experiments.[2] Upon arrival in the laboratory, participants were all placed around a single table; instructions were distributed and explained by the experimenter. Participants then had a 10-minute period for dialogue. Afterward, they moved to isolated individual tables where they could no longer see nor communicate with each other.

While acknowledging that communication adds hard-to-control variables to experiments, as stressed by Ledyard (1995), and is therefore unusual in standard economic experiments, the research team claims that communication should not be assumed away, as it is a constitutive element of many real-life contexts of collective action. Furthermore, the existence of the 10-minute dialogue period allows the experiments to

be conceived as instances of "communicative action." According to Habermas (1986), (inter)actions are communicative when participants *agree* with one another to make plans for common action; in such cases, it is the common understanding and interpreting of the situation that constitutes the prevailing coordination mechanism. In contrast, in "strategic action," participants are focused on the *success* of their endeavor, that is, on the *consequences* of their action, and the mechanisms through which ends are achieved are rewards or sanctions. Communicative action is characterized by the use of language. The actual coming together and agreeing through the inter-understanding processes allowed by language adds a communicative dimension irreducible to the teleological dimension inherent in the pursuit of a common goal. As elaborated in subsequent sections of this article, participants are rationally or internally motivated to act collectively because of the commitments triggered by their speech acts. Strategic action, in contrast, is motivated by individual payoffs, and coordination depends on the convergence or not of the participants' individual interests. Strategic action is the very frame of action adopted by standard economics and game theory but, as we will see, it may be a misleading path.

In all four experiments, subjects received an amount of convertible token money, which they might keep to themselves or deliver to a common fund (i.e., the collective good), knowing that the amount delivered to that fund would be duplicated and that the final amount would be distributed among the group members. This situation was repeated 10 times, and at the end the token points earned were converted into money. Information about their individual share (or payoff) was privately handed on to the subjects, thus ensuring anonymity. Participants could infer the total contribution of the group from the information handed to them.

The particular feature of this experimental study was that the collective good was distributed among the group members in accordance with a rule that was different procedurally and with respect to sharing proportions in each experiment:

- Experiment A: The collective good was distributed in equal shares among the participants; subjects were informed about this ("equal") distribution rule—it was specified in the instructions.

- Experiment B: One of the participants (to be selected by a random draw once all subjects were isolated after the dialogue phase) was to receive 30 percent of the collective good, the remaining subjects receiving 10 percent each; subjects were informed about this ("unequal") distribution rule—it was specified in the instructions.
- Experiment C: The distribution rule was to be decided by a randomly chosen subject (the "rule-chooser") who was faced with two alternatives: (a) 30 percent for himself or herself with the remaining individuals getting 10 percent each; or (b) 44 percent for himself or herself with 8 percent for the remaining subjects.[3] Subjects were told in the instructions that the distribution rule would be decided by one of them, but the instructions ignored the alternatives that the rule-chooser would face.
- Experiment D: The distribution rule was to be chosen by the subjects themselves, and they were asked to do so during the dialogue phase.[4]

Figure 1 exhibits the series of the average contribution rates (the average share of the initial endowment that subjects delivered to the collective good) in the 10 rounds (or repetitions) of each session. The 10 series/sessions by experiment are represented (due to overlapping, some are not visible). Table 1 sums up the results in terms of average contribution rates of each experiment.[5]

By clustering the series represented in Figure 1, four types of contributive patterns were identified. Cluster 1 (the strongly noncooperative pattern) gathers sessions that start with comparatively low levels of contribution and end up with very low contribution rates in the 10^{th} round. Cluster 2 (the noncooperative pattern) includes sessions that start with high contribution rates and end up with very low contributions. Cluster 3 (the cooperative pattern) gathers sessions that start with high contribution rates and end up with low levels in the last two rounds (the last-round effect). Finally, Cluster 4 (the strongly cooperative pattern) corresponds to the sessions with high levels of contribution over the 10 rounds.

Table 2 presents the frequency of each contributive pattern in each of the four experiments.

Figure 1

Average Contribution Rates to the Collective Good by Session

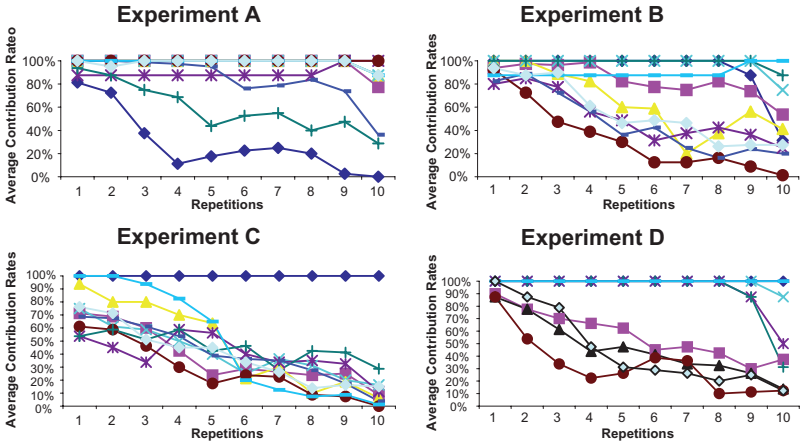

Table 1

Average Contribution Rates by Experiment

Experiment	A	B	C	D
Average Contribution Rates by Experiment	85%	71%	47%	77%
Average Contribution Rates by Experiment: 1st Round	96%	93%	75%	97%
Average Contribution Rates by Experiment: 10th Round	69%	46%	19%	55%

The most salient results are the high contributive levels in all experiments except for C when compared with standard experiments (with no communication); and the comparatively low contribution rate and comparatively high frequency of noncooperative contributive patterns in Experiment C.

By using the Kruskall-Wallis test for the differences in medians of the average contribution rates per session (the average of the

Table 2

Frequency of Contributive Patterns per Experiment

	Exp. A	Exp. B	Exp. C	Exp. D
Strongly Noncooperative (Cluster 1)	1	1	7	2
Noncooperative (Cluster 2)	1	4	2	2
Cooperative (Cluster 3)	1	2	0	2
Strongly Cooperative (Cluster 4)	7	3	1	4

contribution rates over the 10 rounds), a statistically significant difference was identified ($p = 0.02$) when the results of the four experiments are taken together. Further comparisons reveal that the sharpest difference is between Experiment C and the other three experiments. This is shown by the comparisons {C} vs. {A, B, D} (the difference is statistically significant with $p = 0.003$), {C} vs. {A} ($p = 0.014$), {C} vs. {B} ($p = 0.023$), and {C} vs. {D} ($p = 0.025$). Furthermore, no significant differences were identified when simultaneously comparing A, B, D, or {A} vs. {B}, {A} vs. {D}, and {B} vs. {D}.[6]

C. What is to be Explained?

During the dialogue phase preceding each session, the subjects jointly interpreted the context they were facing, pointing out the fact that the best solution for all would be for all to contribute, even if each of them would have an advantage from free-riding on the contributions of others. Extensive promise-making was observed, with a large part of the subjects expressing their personal intentions to contribute.

That people tend to commit to promises is strongly suggested by the very high average contribution rates observed in the first round (from 93 percent in B to 97 percent in D) and over all 10 rounds (from 71 percent in B to 85 percent in A) in all experiments except C (see Table 1). These rates are substantially higher than those obtained in the absence of communication, and they come as a confirmation of

previous experimental results that showed that communication is not mere "cheap talk" (Ledyard 1995), contrary to the standard prediction.

So a first conjectural statement might be formulated to account for subjects' behavior that will be the object of further elaboration in the next sections of the article:

> *Statement 1: Subjects tend to honor their promises even on occasions when an assessment of consequences, as measured by the expected utility of their payoffs, asks them to defect.*

However, communicating and expressing contributive intentions was not always a sufficient condition for the subjects to behave cooperatively. A large heterogeneity of contributive dynamics was observed (see Figure 1 and Table 2). Two main types of contributive dynamics can be identified: (strongly) cooperative patterns and (strongly) noncooperative patterns. The results suggest a strong dependence of contributive patterns on the initial level of contributions. It seems that subjects formed high expectations with regard to the cooperative dispositions in the group. When those expectations were confirmed by actual contributions in the first round, a sustained cooperative dynamics took place; that is, positive generalized reciprocity was observed. On the contrary, when those expectations were defrauded, even if marginally, and when noncontributions occurred, a spiral of noncontributions followed; negative generalized reciprocity prevailed.[7] This behavioral pattern is frequently explained in the literature by the agents' following a "reciprocity behavioral norm." But the very mechanism of reciprocity calls for further understanding.

The second and third correlated conjectural statements that need further scrutiny, taking together finding (c) referred to earlier and our own results, hence relates to positive and negative reciprocity:

> *Statement 2: In collective action settings, most subjects form the belief, and hence expect, that others will contribute if they themselves contribute.*
> *Statement 3: Most subjects cease to contribute as soon as one subject defects.*

The third main relevant result for our purpose is the significantly lower level of the average contribution rates in Experiment C (see Table 1) and the predominance of noncooperative patterns in this experiment (see Table 2). Recall that the contrast between Experiment C and the

other three experiments involves the fact that the (unequal) distribution rule is set by one of the participants, seemingly to his or her own benefit. Recall also that the incentive structure, that is, the distribution rule and the respective payoffs, were the same in Experiments B and C. Nevertheless, the contribution rates in Experiment B were significantly higher than in Experiment C.

In analyzing the observed difference, one must bear in mind that:

(a) the context in the dialogue phase was remarkably different in Experiment C. In ignorance of the distribution rule and of the alternatives that would be enacted, subjects used the preplay dialogue time to talk about the decision that was to be taken by the "rule-chooser" instead of expressing intentions and commitments regarding their future behavior; the available information was insufficient for participants to agree on a common plan.
(b) given that the subjects were unaware of the alternatives with which the rule chooser was to be faced, the interpretation of his or her intentions may have differed. While some subjects may have figured out that the rule-chooser chose the "fairer" available alternative (the one closer to the equal-share principle), others may have believed that he or she deliberately chose a rule that benefited the rule chooser to the detriment of others.

Both factors converge into a situation charged with a high degree of normative uncertainty, which stemmed from the difficulty of reading the intentions of their counterparts, including those of the rule chooser. In such conditions, many subjects refrained from full contribution in the first round, contrary to what was observed in the other experiments. This gave way to negative reciprocity spirals. As noted by McCabe et al. (2003), anything that makes the signals about others' intentions more difficult to discern reduces the likelihood of cooperation.

The fourth statement, inferred from the observation of subjects' behavior, on which our argument will develop is:

> *Statement 4: Subjects evaluate and value the intentions behind actions as well as the consequences of actions.*

III

Utility-Maximizing Behavior: An Unsatisfactory Explanation

A. Psychological Egoists Would not Comply with Promises nor Take Part in Generalized Positive Reciprocity Processes

As already mentioned, rational choice theory is mute about which desires and beliefs guide individual action. To generate empirical predictions, it must be supplemented with specific claims about motivations. Following Edgeworth's (1881: 16) famous statement,[8] self-interest has been pervasively prone as the basic assumption to explain a variety of behaviors. The coined term to designate self-interested agents is *psychological egoism*.

1. Trying to Explain Cooperation by "Egoistic" Self-Interest

Self-interested behavior may be interpreted in the narrow sense of strictly egoistically motivated behavior, without any regard for others' or for collective well-being. Conventional theory accordingly assumes that individuals in collective action settings behave like selfish maximizers and do not cooperate, even if acknowledging that contributing to the collective good leads to higher outcomes for each and all.

If psychologically egoist agents make promises, this is cheap talk unless utility calculations say otherwise. The making of a promise may be defined as the voluntary creation of an obligation by the use of words. But "nothing is more easily broken than a man's word," in the words of Hobbes ([1651] 1994: 81).

Economists have provided several types of solutions for collective action dilemmas, all of which consist of refinements to the games so as to induce egoistic agents to take actions that ultimately result in mutually beneficial outcomes (James 2002). Incentive-compatible devices, which may take the form of incentives (higher payoffs, contractual arrangements, benefits of reputation) or of sanctions (punishment, costs of bad reputation), are introduced into the games to create an individual incentive to cooperate. Most of the solutions proposed by economists and game theorists to explain promise-keeping and reciprocity amount to applying Hobbes's solution, that is, to creating a *Leviathan*:

> If a covenant be made . . . in the condition of mere nature (which is a condition of war of every man against every man) upon any reasonable suspicion it is void; but if there be a common power set over them both, with right and force sufficient to compel performance, it is not void. For he that performed first has no assurance the other will perform after, because the bonds of words are too weak to bridle men's ambition . . . and other passions, without the fear of some coercive power . . . He which performeth first does but betray himself to his enemy. (Hobbes [1651] 1994: 84–85)

So to escape the state of nature, that is, to make social living possible among egoistic agents, rules and institutions must be established.[9] The ultimate condition under which individuals become obligated to honor contracts is mutual trust, which is absent in the state of nature and comes into existence only if held by a strong central authority (Moss 2002). If only fear keeps people honoring promises and reciprocating, individuals will rationally defect whenever it pays and/or it is safe to do so, which will be often because sanctions and surveillance are costly and imperfect.

More recently, a "spontaneous" solution to social dilemmas was envisaged. It was shown that in repeated prisoners' dilemmas, cooperation may emerge as a plausible equilibrium. But for cooperation to emerge, interaction must be repeated infinitely among the same group of agents and perfect information on past contributions must be assured. This result obviously cannot be transposed to generalized reciprocity situations taking place between random individuals with indefinite horizons of interaction.

Hence, promise-keeping and positive reciprocity (Statements 1 and 2) are not to be expected from psychological egoists. In order to account for these behaviors, some economists propose to reintroduce a substantive moral psychology into rational choice theory[10]—which was exactly what Pareto and the subsequent formal theory of choice were trying to avoid.

2. Introducing Social and Moral Preferences into the Standard Rationality Model

In the "social preferences" models, agents are supposed to derive satisfaction from others' welfare. Interactions with others are subject to the same cost-benefit analysis as material payoffs in the form of

psychological or emotional pains and benefits. Agents are still doing what they most want to do, but now this also involves acting unselfishly or according to moral rules.

The tendency to choose "fair" strategies and to reject "unfair" offers, for instance, is interpreted as reflecting inequity aversion, and the justice motive is then modeled by augmenting the agents' utility functions with arguments that represent it (Bolton and Ockenfels 2005; Fehr and Schmidt 1999). The logic of the fairness equilibrium is quite simple (Rabin 1993): if *i* thinks that *j* is going to act fairly toward him, *i* is more likely to act fairly in return. When both act fairly, both derive positive (nonmaterial) utility from the exchange, in addition to any material utility. And acting fairly toward one who is unfair reduces one's materially based and psychologically based utility (Nelson 2002: 424).

But might the addition of other-regarding motives plainly explain promise-keeping and reciprocity?

If this specific kind of desire were to be considered as an argument of the utility function, the utility numbers would be altered. In some cases, this would automatically dissolve social dilemmas, but there would still be cases where payoffs, even when encompassing elements like distributive concerns or emotions, do not eliminate conflict between individual and collective interests. The individual's ordering of consequences expressed in utils may still result in the higher ranking of noncomplying to a promise or of not reciprocating. These models, instead of trying to understand what is at stake in such problematic cases, solve the problem by a posteriori interpreting that the agent had maximized his or her utility.

The social preferences and other utility-extended models construe the utility set broadly enough so as to include elements like desires as well as others like commitments or obligations. As argued by Searle (2001), this blurs the distinction between desire-dependent and desire-independent reasons for action, that is, between things we *want* to do and things we *have* to do:

> [T]he soldier who throws himself on a live hand-grenade in order to save the lives of his fellow soldiers is in exactly the same situation, rationally speaking, as the child who selects chocolate over vanilla when picking a

> flavor of ice cream. The soldier prefers death, the child prefers chocolate. In each case, rationality is just a matter of increasing the probability of getting to a higher run on the preference ladder. (Searle 2001: 168)

To distinguish between acting to satisfy desires and acting to honor obligations is, in our opinion, theoretically and practically essential. Honoring one's word and entering in reciprocity processes demands that agents sometimes rely on others to act contrary to the balance of their expected utility. As long as individuals are considered to have regard for the good of others only insofar as furthering it benefits them, cooperation and collective action are in peril. This difficulty is inherent in the concept of utility as the sole motivation for action and is not bypassed by substantive assumptions on preferences (Hollis 1998: 54).

So, subjects' behavior, as conjecturally described in Statements 1 and 2, remains unsatisfactorily explained. Cooperation in social dilemmas calls for *normative* expectations that can cement agreements despite calculations of utility. To put it plainly, what has to be explained is why people believe that they and others *ought to* comply with promises and cooperate even when it is contrary to their own interest.

B. Philosophical Egoists Might Comply with Promises and Enter in Reciprocity Schemes to Benefit from the Cooperative Surplus

In the experiments' dialogue phase, every subject agreed that the most preferred action was that everyone would contribute the whole amount to the collective good, for this would increase everyone's payoffs in the end. When individuals take into account their long-term interest, they acknowledge that it is in everyone's advantage to cooperate. But to benefit from the cooperative surplus, individuals need to be able to enter into mutual commitments. Prudential calculus thus requires binding behavioral rules and normative expectations.

A number of contractarian philosophers contend that compliance with moral principles could be shown to be compatible with rational action because it is conducive to the long-term interest of every member of society. Promises, agreements, obligations, and duties emerge as the

rational and functional artifices for the sake of better consequences for each and all. In these approaches, individuals remain solely moved by their own wants, and their actions are still guided by the ranking of the consequences. The utility maximization model still holds. These agents may be called enlightened egoists or *philosophical egoists*, to adopt Hollis's suggestive term (Hollis 1998: 21).

The most thorough attempt to equate morality with—or even to derive morality from—self-interest is made by David Gauthier (1986). Agents are conceived as being "constrained maximizers" in the sense that they do not try to maximize only their immediate marginal utility but also their "overarching life-plan," which takes into account the potential benefits from social interactions and cooperation. While not being naturally endowed with any moral sense, agents will progressively acquire a "moral disposition" that makes them honor promises and, more generally, play fair with those agents who also play fair with them. Morality, far from being inimical to one's interests, represents the condition for their effective pursuit:

> Each person assumes moral personality in accepting obligations that constrain her pursuit of her own good, but in this each demands that others accord her moral personality in being recognized as exercising rights, equally with her fellows, that allow the pursuit of that good. (Gauthier 1993a: 38)

But Gauthier's agents, if moral, do not cease to be strictly self-interested:

> Utilities (or the preferences they measure) and only utilities, provide reasons for acting . . . ; an agent has sufficient reason to choose one action rather than another if and only if the former affords her an expected utility greater than the latter. (Gauthier 1993b: 185–188)

As Hollis judiciously comments: "the disposition to play fair can be switched on and off. It is Kantian in the On position and self-interestedly consequentialist in the Off" (Hollis 1998: 97). Philosophical egoism will arbitrate between turning the moral disposition on or off. So, how are Gauthier's agents to be expected to behave?

In such a philosophical system, obligations and agreements are binding to the extent that they are instrumentally beneficial. Justice

concerns are but an enlightened form of self-interest. And "words have strength from their own nature" (to use Hobbes's words) because honoring one's word is a rule that rational people will agree to accept, for their mutual benefit, on the condition that others follow it as well. An instrumental morality does make individuals more trustworthy than psychological egoists. But although it may lead to the belief that agreements will be honored, this notoriously fails to guarantee that anyone will abide by them. Even though it may usually be to one's advantage to honor mutual commitments, one might sometimes gain by not honoring them. Thus it appears that not all moral obligations can be explained as derivable from self-interest.

In this framework, rational agents will indeed keep their promises and reciprocate *most of the time*, and so Statements 1 and 2 are partially accounted for. It should be noted that for Hobbes, sanctions and institutions are needed to enforce the social contract. In contrast, Gauthier's philosophy—which explicitly stands between Hobbes and Rawls (Gauthier 1993a)—does not acknowledge such requirements. But we may wonder if the agreed behavioral rules have a proper moral character, as their compliance will fade as soon as enlightened self-interest asks for noncompliance.

Cooperation and reciprocity actually demand two moral capacities that are profoundly alien to agents motivated as philosophical egoism supposes. First, they require that actions themselves, and not only their consequences, are evaluated and valued: having made a promise requires it to be honored just because it has been made, and this is so even on occasions when the consequences of breaking it increase one's utility. Second, successful cooperation supposes the adoption of an impartial standpoint, one from which each individual's interests are equally important. This is deeply incompatible with "egoism," whatever its nature. Philosophical egoists are only able to use a *I* language, and they lack the cognitive skills required by impartiality, which implies that agents are able to adopt a point of view that is exterior to their own (Habermas 1995).[12] The nature of morality has to be examined, as well as the kind of individualism in which it is grounded.

IV

Moral Norms and Moral Sentiments: Justice Motives as the Foundation of Reciprocity and Cooperation

A. The Crucial Role of Moral Sentiments in Moral Motivation

Even if we posit that individuals have internalized binding rules—be they the result of past social interactions or of enlightened self-interest—it remains that it is one thing to rationally accept and adhere to a norm; it is another to actually behave according to that norm, particularly in the cases when it conflicts with one's immediate interest. What motivates people to honor their moral duties?

The nature of the norms in question and their moral dimension need further examination. One may wonder if the behavioral rules are binding in *foro interno* (before an "interior tribunal," that is, conscience) or in *foro externo* (before an "exterior tribunal"). Do individuals genuinely care about justice and hence their behavior is driven by their *autonomous* adherence to justice principles, or do they behave fairly as a result of justice norms acquired because of the fear of sanctions or of other kind of consequences in the outside world?

If rule-following and internalized behavioral norms account for the formation of normative expectations necessary to keep promises and participate in reciprocity processes (Statements 1 and 2), Statement 3—a single defection is enough to trigger generalized defection—is more difficult to explain. Positive reciprocity is a rule that one should follow because it promotes the general welfare; it may be justified for "consequentialist" collective and personal reasons. On the contrary, negative reciprocity seems to involve other issues. If reciprocity is a commonly agreed-upon rule and if someone violates it, we may in turn feel released from our obligation. This is the very nature of reciprocity. But is this a satisfactory explanation?

What seems to underlie Statement 3 is the sense of justice that historically emerged with the experience of injustice—"That's not fair!," the moving exclamation everyone remembers as a child. The immediate reaction to someone's defection appears to be triggered by indignation, probably the most important of all moral sentiments. Indignation is not an affective reaction against a given person in a

particular situation. Its moral character derives from its having an *impersonal* form directed to all who betray norms and legitimate expectations. And indignation triggers the imperative to reestablish justice, and this with no consideration of the outcomes, be they personal or collective benefits or losses. In deciding to withdraw from contributing as soon as one defection is acknowledged, experimental subjects' behavior seems to stem both from a moral sentiment (indignation) and from a principled moral norm (justice).

Ricoeur (2001) conceives of morality as composed of obligatory norms (likely to satisfy Kant's universal principle) *and* moral sentiments because, according to Ricoeur, only moral sentiments, which belong to the domain of desires, can be a motive for action. Duties alone are not sufficient by themselves to drive human action; justice as a motivational drive has to be an object of desire. Acting from an abstract sense of duty or from a desire to "do the right thing" is not the same. And all actions have to be backed by reason(s); sentiments alone are not sufficient, either. Likewise, according to John Stuart Mill (who also explicitly refers to Kant's categorical imperative):

> The idea of justice supposes two things; a rule of conduct, and a sentiment which sanctions the rule. The first must be supposed common to all mankind, and intended for their good. The other (the sentiment) is a desire that punishment may be suffered by those who infringe the rule. (Mill [1863] 2001: 51)

Hence, people respond to a "bad" action by acting badly in return because of moral sentiments that impel one to act according to impartial rules and to punish the defectors. Therefore, the strict dichotomy established by the standard theory of choice between reason and sentiments should be questioned. In Mill's and Ricoeur's philosophies, moral sentiments are not to be assimilated to some "preference for dignity" or "aversion to injustice," and their inclusion in utility functions might be considered illegitimate:[12] "The order of moral sentiments constitutes a vast affective domain irreducible to pain and pleasure" (Ricoeur 2001: 59). In Mill's words, moral sentiments are incommensurable with all other considerations, and in this stands the distinction between the feeling of right and wrong and that of ordinary expediency and inexpediency (Mill [1863] 2001: 53). The last sentence of Mill's *Utilitarianism* is worth quoting:

> Justice remains the appropriate name for certain social utilities which are vastly more important, and therefore more absolute and imperative, than any others are as a class . . . ; and which, therefore, ought to be, as well as naturally are, guarded by a sentiment not only different in degree, but also in kind; distinguished from the milder feeling which attaches to the mere idea of promoting human pleasure or convenience, at once by the more definite nature of its commands, and by the sterner character of its sanctions.

Both authors thus acknowledge two distinctive traits of moral motives. On the one hand, they are irreducibly linked to moral sentiments; on the other, they refer to moral norms that seem to be of a deontological nature. Moral norms are committing and thus can hardly be assimilated to the teleological nature of preferences.

B. Acting on Deontological Moral Principles: Impartiality and Individuality

Let's now turn to the fourth statement, which is, in our opinion, the most crucial one to be explained as it probably stands behind the other three: *Subjects evaluate and value the intentions behind actions as well as the consequences of action.*

As recognized by the very authors of the social preferences approach, "there remain legitimate doubts whether [social preferences] models capture the phenomenon of reciprocal fairness in a fully satisfactory way" (Falk, Fehr, and Fishbacher 1999: 3). Other experimentalists recently proposed an approach that challenges the outcome perspective: "Whereas the outcome-based approaches imply that intentions are superfluous, intention-based models rely essentially on players reading each other's motives (and not merely their actions)" (McCabe, Rigdon, and Smith 2003).

The collected evidence shows that subjects evaluate not only the action's outcomes but also the intentions that drive or are signaled by the action. Behaviors such as honoring of commitments or reciprocating seem to be assessed against moral norms concerning the intentions themselves and not solely the consequences of the actions. These moral norms thus seem to be binding *in foro interno*. They are usually associated to deontological doctrines.

While Searle's theory (2001) explicitly stands outside of this philosophical class of doctrines, he shares the Kantian idea by which freedom and autonomy reflect the dominance of rationality over purely individual interests or desires.[13] For Searle, humans *intentionally* and rationally create commitments through the use of language, which commitments in turn create desire-independent reasons that motivate action. When someone makes a promise, the propositional content of the promise specifies some voluntary action that the speaker commits to taking in the future, regardless of what the other motivations might be for acting at that time. The speech act of promising is a commitment to a future action. Commitments and obligations are internally binding on the agent and are not just social constructs like any other because the speaker stands in a special relation to his or her own assertions in that the speaker creates them as his or her own commitments and thus binds himself or herself.

Searle (2001) establishes a distinction between desires and desire-independent reasons for action. Both *wanting* to eat chocolate and *wanting* to keep one's promises constitute reasons for action, but in the case of the promise, the desire is derived from the obligation; the reason is prior to the desire and is the ground of the desire, while in the case of chocolate, the desire is the reason (Searle 2001: 170). The obligation is not a prior desire functioning as a reason for the effectiveness of the obligation, as postulated in the utility-extended models. The commitment is a desire-independent reason that rationally grounds the secondary desire of wanting to fulfill the obligation. An effective reason to act in the future is thus rationally created. Hence, commitments create *internal* motivations for performing the action.

But for Searle, obligations are fulfilled for internally rational reasons in a naturalistic or self-sufficient way; there is no need for any external moral principle. By contrast, if Habermas (1986) follows Searle in closely relating reason and language, he departs from him in conceiving the obligation as the very fundamental *moral* phenomenon.

"Every man, as long as he does not violate the laws of justice, is left perfectly free to pursue his own interest his own way" (Smith [1776] 1976: 51). The special status or nature of the "laws of justice" has been the main focus of neo-Kantian philosophies.[14] According to Habermas

(1986), moral norms, such as the laws of justice, require rational justification and grounding, which is in turn associated with universality and impartiality requirements. He also argues that the universality of a maxim depends on it being validated through a discourse ethics, through communication; universal maxims cannot be the result of a monological operation of thought.

The experiments show that subjects act on norms suited to joint undertakings. Both Statements 2 and 3 suggest that they surpassed their instrumental rationality when they acted on a norm that stands above their own point of view. Is impartiality possible among totally separated individuals?

As asserted by Hobbes, "the notions of right and wrong, justice and injustice . . . are qualities that relate to men in society, not in solitude" (Hobbes [1651] 1994: 78). Morality is thus to be conceived of as being associated with socially embedded men, not with isolated individuals. Contrary to Hobbes, Ricoeur and Habermas are among those who contend that neither impartiality nor reason can be conceived as a property of isolated individuals. For Ricoeur, the identity of the self is constituted through a relational structure in which the dialogical dimension prevails over the monological. Ricoeur, as Habermas, strongly criticizes Kant for having conceived reason as possible among silent, isolated individuals; for both of them, reason cannot be monological, and impartiality would not grow among isolated individuals. In the constitution of the self and of reason, both the "near others" (the interpersonal relations of friendship) and the "distant other" (each one), without whom it would not be legitimate to speak of justice, count. The fundamental aim of humans, that of happiness, incorporated in the heart of economics, does not end in solitude, nor in friendship; it encompasses the city, as Aristotle would put it (Ricoeur 1995: 14–17). Our happiness depends on others'.

C. Exposing the Relation Between Justice Motives and Reciprocity and Honoring Promises

In the literature, reciprocity is often connected with fairness or justice. A commonly shared definition of reciprocity is to act fairly

toward someone. Or, in Nelson's terms (2002: 425): "Players' decisions to positively or negatively reciprocate depend on the perceived fairness of their partners' strategies." The perception of unfair intentions triggers indignation[15] and hence defection. Not to contribute can no doubt be a manifestation of greed. It may also be a consequence of fear because of a lack of confidence in others. Very often, however, it is first and foremost an expression of indignation in response to unjust behavior or the unjust intentions of others. In many experimental and real-life contexts, the only means available for agents to express their disapproval of opportunistic behavior is by defecting themselves. Conversely, the only means to express approval is by responding with further contribution. In repeated public games settings when opportunities for punishment or retaliation exist, most subjects tend to use them even at a cost to themselves (Fehr and Gachter 2000), acting in ways to maintain or restore justice. Reciprocal behavior is hence to be conceived as grounded in justice norms.

The association between justice and honoring of commitments, in which the following of reciprocity norms might be included, is less consensual. Among early modern philosophers, the point was noncontroversial. For Hobbes, "when a covenant is made, then to break it is *unjust*; and the definition of INJUSTICE is no other than the *non performance of covenant*" (Hobbes [1651] 1994: 89, emphasis in original).

Along the same line, J. S. Mill classifies the violation or the fulfillment of engagements as well as *desert*, that is, returning good for good and evil for evil, as pertaining to the class of *unjust* or *just* actions: "The precept of returning good for evil has never been regarded as a case of the fulfillment of justice, but as one in which the claims of justice are waived, in obedience to other considerations" (Mill [1863] 2001: 44).

Therefore, most of what was to be explained (as expressed in the four statements) is somehow related to justice motives. This is to be expected, since the concern for justice pinpoints the most serious difficulty for utilitarianism[16] and since the latter used to be the explicit or implicit philosophical background of economics.

V

Concluding Remarks

SELF-INTEREST IS INDEED a widespread and powerful pattern of motivation, and this makes it an obvious candidate for reconstructing moral concern. But the explanation of the "puzzling facts" that experimental findings have brought into light call for theoretical elements that are deeply alien to standard game and rational choice theory. If morally driven behavior is to be seriously taken into account in economics, the ontology of the individual is to be revised to encompass the social nature of the constitution of human identity and the communicative dimension of action. The other related challenges are to theoretically integrate the deontological moral elements and the specificity of moral sentiments that undeniably influence real agents' behavior and that are not reducible to utility maximization.

Putting forward utility maximization, pervasively interpreted as self-interest maximization, as normative behavior may have the regrettable effect of making it come true. As Hannah Arendt ([1961] 1983: 401) put it: "What is particularly awkward in modern theories of behavior is not that they may be false, it is that they might become true."

Notes

1. Although there is a remarkable diversity of public goods experiments, a standard one, according to Caldas et al. (2004), uses the following procedures: a group of n individuals (generally between four and ten, but sometimes more) is brought into a room (the lab); each of the participants is given a certain amount of money (an endowment z_i), which he or she has to divide into a part, x_i, that the participant keeps to himself or herself and another part, $t_i = z_i - x_i$, which is "invested" in the production of the public (or collective) good. The total amount invested, $T = \sum_{i=1}^{n} t_i$, is then used to produce the public good y, with $y = g\ (T)$ being the public good production function. The individual payoffs are then determined by the choices x_i, the amount of the public good y, and the way in which y, is distributed among the participants.

In standard experiments, y is equally distributed among participants, that is, each participant receives the same share of y (as it is the case in our Experiment A). For a full presentation of the experimental procedures and results, see Lopes et al. (2004), available at http://www.dinamia.iscte.pt.

2. Ten sessions of each of the four experiments were conducted, each involving eight subjects. In all, 320 students of different faculties participated in the study (4 experiments × 10 sessions × 8 subjects each).

3. As the objective was to compare the results of Experiment C with those of Experiment B, only cases where the selected subject choose the 30-10 rule were considered in data treatment (10 cases out of 13).

4. In all cases, the egalitarian distribution rule was unanimously chosen.

5. The average contribution rate by experiment stands for the average value of all participants' contribution rates for all rounds of the 10 sessions of each of the four experiments (average of 800 values). The average contribution rate in a given round by experiment gives the average value of all participants' contribution rates in this round (here, the first and the tenth) of the 10 sessions of each experiment (average of 80 values).

6. The same results hold for tests using the median of the average contributions in the first and last (10^{th}) round.

7. It is to be noted that most of the commitments voiced during the dialogue phase were indeed conditional: "If everyone contributes, so will I."

8. According to which: "The first principle of economics is that every agent is actuated only by self-interest."

9. Most economic models actually make such behavioral assumptions: agents are considered as potential opportunists, and institutions must be devised to ensure against moral hazard (see principal-agent models).

10. Economists are not the only social scientists to integrate "social preferences" in the rational choice framework to explain cooperation and give account of justice motives. Among social psychologists, "the general consensus is that people's desire for justice is neither distinct from self-interest nor of great motivational importance. Most contemporary social psychologists assume that people employ justice as a personal and social device to promote their acquisition of commonly desired resources" (Lerner 2003: 388).

11. A criticism, much in line with the argument developed here, of Rawls's adoption of the rational choice framework in what concerns Rawls's conception of moral agents can be found in Habermas (1995).

12. The inclusion of moral concerns as arguments in utility functions is proposed by the above referenced social preferences approaches, among others. Such an inclusion assumes moral "preferences" and selfish preferences to be commensurable: they are both evaluated through the same—unique—utility criteria, translated into satisfactions and, thus homogenized, treated instrumentally. It is thus not acknowledged that moral judgments, different from mere tastes, are rationally defendable. One does not have to argue about liking chocolate or not but one does have to rationally justify—and try to persuade others of—what is to be considered a moral action. In conflating preferences and values, these models deny the relevance of ethics and morality.

13. "Only a *categorical* value can constitute a real limit to one's individual interests. Such an imperative precisely requires . . . a faculty of detachment from oneself, a concept of freedom that is opposed to the rationality of interests" (Canto-Sperber 1992: 242–243).

14. It is not a coincidence that many contemporary thinkers integrate Kantian elements in their arguments. At a time when the conceptions of the good are irreducibly plural, any form of impartiality acquires a Kantian flavor. However, we will not subscribe here to an explanation exclusively in terms of categorical imperatives, as the facts to be explained obviously contain a conditional character.

15. Dawes et al. (1988) reported the same type of observations as those registered in our experiments: subjects strongly exhibit indignation at the end of noncooperative sessions.

16. As recognized by the very author of the term itself (see the last chapter of Mill's *Utilitarianism*). See also Smith's quotation above.

References

Arendt, Hannah. ([1961] 1983). *Condition de l'homme moderne*. Paris: Calmann-Lévy.

Bolton, Gary, and Axel Ockenfels. (2005). "A Stress Test of Fairness Measures in Models of Social Utility." *Economic Theory* 25(2): 957–982.

Caldas, José, João Rodrigues, and Luis Francisco Carvalho. (2004). "Economics and Social Psychology on Public Goods: Experiments and Explorations." *Notas Económicas* 18: 31–45.

Camerer, Colin, and Ernst Fehr. (2002). "Measuring Social Norms and Preferences Using Experimental Games." IERE, Working Paper 97. University of Zurich.

Canto-Sperber, Monique, ed. (1992). *Dictionnaire de Philosophie Morale et Politique*. Paris: PUF.

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

Edgeworth, Francis. (1881). *Mathematical Psychics*. London: Kegan Paul.

Falk, Armin, E. Fehr, and U. Fischbacher. (1999). *On the Nature of Fair Behaviour*. Institute for Empirical Research in Economics Working Paper 17.

——. (2003). "On the Nature of Fair Behavior." *Economic Inquiry* 41(1): 20–26.

Fehr, Ernst, and Simon Gachter. (2000). "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives* 14(3): 159–181.

Fehr, Ernst, and Klaus Schmidt. (1999). "Theories of Fairness, Competition and Cooperation." *Quarterly Journal of Economics* 114: 817–868.

Frey, Bruno, and Stephen Meir. (2004). "Pro-Social Behavior in a Natural Setting." *Journal of Economic Behavior and Organization* 54(1): 65–88.

Gauthier, David. (1986). *Morals by Agreement*. New York: Oxford University Press.

——. (1993a). "Between Hobbes and Rawls." In *Rationality, Justice and the Social Contract*. Eds. D. Gauthier and R. Sugden. Hertfordshire, UK: Harvester Wheatsheaf.

——. (1993b). "Uniting Separate Persons." In *Rationality, Justice and the Social Contract*. Eds. D. Gauthier and R. Sugden. Hertfordshire, UK: Harvester Wheatsheaf.

Gneezy, Uri, and Aldo Rustichini. (2000). "A Fine is a Price." *Journal of Legal Studies* 29(1): 1–17.

Habermas, Jurgen. (1986). *Morale et communication*. Paris: Flammarion.

——. (1995). "Reconciliation Through the Public Use of Reason: Remarks on John Rawls's Political Liberalism." *Journal of Philosophy* 92(3): 109–131.

Hobbes, Thomas. ([1651] 1994). *Leviathan*. Cambridge, UK: Hackett Publishing Company.

Hollis, Martin. (1998). *Trust Within Reason*. Cambridge: Cambridge University Press.

James, Harvey. (2002). "The Trust Paradox: A Survey of Economic Inquiries into the Nature of Trust and Trustworthiness." *Journal of Economic Behavior and Organization* 47(3): 291–307.

Kahan, Dan. (2005). "The Logic of Reciprocity: Trust, Collective Action and Law." In *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Eds. Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr. Cambridge: MIT Press.

Ledyard, John O. (1995). "Public Goods: A Survey of Experimental Research." In *The Handbook of Experimental Economics*. Eds. John Kagel and Alvin Roth. Princeton, NJ: Princeton University Press.

Lerner, Melvin. (2003). "The Justice Motive: Where Social Psychologists Found It, How They Lost It, and Why They May Not Find It Again." *Personality and Social Psychology Review* 7(4): 388–399.

Lopes, Helena, et al. (2004). "Public-Good Provision—Why People Do (Not) Contribute?" DINAMIA-ISCTE Working Paper 2004/36. Lisbon.

Lopes, Helena, José Caldas, Ana Costa, and João Rodrigues. (2005). "Collective Action and Justice: An Experimental Exploration." Manuscript, Dinâmia.

McCabe, Kevin, Mary Rigdon, and Vernon Smith. (2003). "Positive Reciprocity and Intentions in Trust Games." *Journal of Economic Behavior and Organization* 52(2): 267–275.

Mill, John Stuart. ([1863] 2001). *Utilitarianism*. Kitchener, Ontario: Batoche Books.

Moss, Laurence S. (2002). "Hobbes and the Early Uses of Economic Method." *Annals of the Society for the History of Economic Thought* 42: 1–17.

Nelson, William. (2002). "Equity or Intention: It Is the Thought that Counts." *Journal of Economic Behavior and Organization* 48(4): 423–430.

Ostrom, Elinor. (2000). "A Behavioral Approach to the Rational Choice Theory of Collective Action." *American Political Science Review* 92(1): 1–22.

Rabin, Matthew. (1993). "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83(5): 1281–1302.

Ricoeur, Paul. (1995). *Réflexion faite—Autobiographie intellectuelle*. Paris: Éditions Esprit.

——. (2001). *Le Juste 2*. Paris: Éditions Esprit.

Roth, Alvin. (1995). "Introduction to Experimental Economics." In *The Handbook of Experimental Economics*. Eds. John Kagel and Alvin Roth. Princeton, NJ: Princeton University Press.

Searle, John. (2001). *Rationality in Action*. Cambridge: MIT Press.

Smith, Adam. ([1776] 1976). *An Inquiry into the Nature and Causes of the Wealth of Nations*. Eds. R. H. Campbell, A. S. Skinner, and W. B. Todd. Oxford: Oxford University Press.