

BEYOND ALTRUISM, DUTY, OR COLLUSION INTRODUCING SOLIDARITY INTO ECONOMICS *

Christian Arnsperger
FNRS & Chaire Hoover
Université catholique de Louvain
arnsperger@etes.ucl.ac.be

Yanis Varoufakis
Department of Economics
University of Sydney
yanisv@bullwinkle.usyd.edu.au

1. Introduction

Throughout Europe, shoppers periodically pay at the supermarket counter for food which they then proceed to deposit in “solidarity bins” destined to be sent to Albania, Kurdistan, or East Timor. Workers the world over go on strike in support of dismissed colleagues, even though some of them are not inordinately worried about their *own* future job security. Every year, many Britons contribute to the Life Boat Service— which is run solely by voluntary contributions— even though they may dislike sailing or can never imagine themselves at sea. In laboratories, subjects give money to faceless people who are discriminated against by the experimental design (see Camerer and Thaler 1995 and Hoffman, McCabe and Smith 1996 for some recent evidence).

Economists tend to see these as examples of bounded rationality, of enlightened self-interest (i.e., seemingly other-regarding acts which are, nevertheless, intended to promote one’s own self-interest), or of some form of evolved social reciprocity (i.e., the following of norms of cooperative or seemingly altruistic behavior which evolved in the context of repeated games, but which agents may feel an obligation to follow even in one-shot interactions ; see Hoffman, McCabe and Smith 1996). Philosophers, on the other and, recognize in such acts other-regarding motives which might be independent of any prospect— real or imagined— of reciprocity. For instance, agents may harbor some (Humean) natural sympathy for others, or be utilitarian altruists, or, indeed, they might rationally work out what their obligations to others are (e.g., the Kantian imperative). An alternative, but still unfashionable, way of interpreting such potentially nonselfish acts is as a form of *solidarity with “others.”*

In this paper, we ask whether there is room in economic and social theory for a category of beliefs and actions which (i) would fall under the heading “solidarity” and which (ii) *cannot be satisfactorily accounted for* by more usual notions such as collusion between instrumentally rational

* This paper was written during Arnsperger’s visit to the University of Sydney’s Economics Department in May 1999. A generous travel grant was provided by the Belgian National Fund of Scientific Research. The Department’s hospitality, especially on the part of Phong and Heather, is very gratefully acknowledged. A preliminary version of this work was presented at the Economics Seminar of the Department; the comments and insights provided by the participants were very helpful in order to sharpen some of the ideas presented here.

egoists, utilitarian altruism, team reasoning, or reciprocity. We shall argue that the notion of solidarity can be usefully employed to describe such a specific class of other-regarding beliefs and acts. Our method will be one of “nested” characterizations: At the outset, our definition of solidarity will cover the conceptual ground of all explanations mentioned above, ranging from reciprocity to duty and pure altruism; however, as we progressively refine this definition, the existing categories of other-regarding behavior will be transcended. We shall claim that the new ground exposed as solidarity *proper*, and gradually distinguished from other forms of nonselfish behavior, promises important insights for economic and social theory.

In section 2, we survey the rare and eventually unsatisfactory appearances of the notion of solidarity in the existing economic literature. Thereafter, we explore the promise of solidarity as an increasingly *separate, specific analytical category*. Section 3 is devoted to a preliminary analysis of a concept of solidarity which, although it covers a wide range of nonselfish individual motivations—and hence already goes quite some way beyond the shortcomings of the existing occurrences of the concept—, is ultimately unsatisfactory in that it does not grasp the “depth” of the notion of solidarity. This deeper dimension of solidarity is explored in section 4, where we argue that to really distinguish solidarity proper from the motley crew of nonselfish motivations, we need to introduce two crucial conditions, which we label “autonomy” and “non-instrumentality.” Section 5 then specifies the content of solidarity even further, delineating “genuine” solidarity as that subset of autonomous and non-instrumental motivations which is directed specifically at victims of arbitrary social power. Section 6 concludes briefly.

2. “Solidarity” in economics: A few discrete appearances

2.1. “Solidarity” as the mystery which overcomes free riding

Early evocations of “solidaristic” elements in behavior are rooted in the idea that norms somehow evolve from collective attempts to overcome free riding. Schelling (1960) is an early source of the idea that individuals, when they are caught up in strategic situations which prevent them from communicating with one another, somehow *do* have a knack for thinking not as isolated selves, but as members of a team. Akerlof (1980) proposes a dynamic model in which a preestablished social custom may gather momentum and sway individuals away from selfish behavior, in direct proportion to the average expectation that its bandwagon will indeed roll. Varoufakis (1989, 1990, 1991) introduces a model of solidarity amongst trade union members which explains, along the lines of Akerlof’s bandwagon effect, the capacity of workers to overcome their free-rider preferences in favor of collective actions. How this can be accounted for and modeled, however, is never made plain.

2.2. "Solidarity" as an axiom

More recently theoretical economists, and especially game and social theorists, have embedded a notion of solidarity into either an axiom to be imposed on solution concepts (Heiding and Moulin 1991, Thomson 1995, Sprumont 1996) or into the intrinsic properties of solution concepts (Nowak and Radzik 1994).

Heiding and Moulin (1991) define solidarity within a parametric problem: "When the exogenous parameter changes, either the welfare of no agent decreases, or the welfare of no agent increases." In a behavioral context, this is somewhat akin to the mechanical movement of a well-structured herd, that is, a herd whose elements are mechanically fitted together: When a predator appears and changes the parameters of the situation, all animals move away in the same direction. A particular version of this mechanistic idea is Thomson's (1995) notion of population monotonicity: "When additional agents arrive, and the profile of welfare levels chosen by the solution for the initial group remains feasible only by 'ignoring the newcomers,' then none of the agents initially present gains. Conversely, the departure of some of the agents, if it permits a Pareto improvement for the remaining agents, is indeed accompanied by such an improvement."

In other words, this "solidarity" axiom ensures that, under normal conditions of scarcity or decreasing returns, a newcomer decreases everyone's utility while a participant's departure raises it. Paradoxically, when only scarce resources enter people's preferences, the axiom will be satisfied only when everyone resents the newcomer(s). This raises a question of implementation which the theory has never addressed. For unless a majority within a population have a reason *not* to turn against the newcomer(s), the solidarity axiom will never be accepted by those who would be responsible for implementing it. And yet, as long as individuals are modeled as narrowly self-interested, they have no reason to see their utility be reduced out of "solidarity" with strangers. How, indeed, do the various incumbents justify *to themselves* this utility loss?

Clearly, such "solidarity" axioms are no more than distributive conditions which impose unanimity across a homogeneous population of utilitarian egoists. Unlike the models of Akerlof and Varoufakis mentioned earlier, solidarity here is neither turned on nor switched off; there is no possibility of interior solutions in which some act solidaristically while others act selfishly. Consequently, such usage of the term *solidarity* would, quite unacceptably, deny that Oskar Schindler's celebrated exploits during the Nazi era were crucially solidaristic simply because of the lack of unanimity on this issue by most of his contemporary peers. Of course, the reason why unanimity is favored by some authors is that, in a world of egoists, it helps them derive well-behaved, herd-like behavioral patterns. Nevertheless, the latter offer only simulacra of solidarity as long as they are not given a more complex *motivational* content— which is what we shall attempt to do in this paper.

In another recent appearance, "solidarity" has emerged as an extension of Shapley's classic idea that each member of a coalition should receive her marginal contribution— the so-called Shapley value. Nowak and Radzik (1994) propose a "solidarity value" for transferable-utility games which differs in one small way from the Shapley value: Instead of positing, as Shapley did, that each

member of a coalition should receive her marginal contribution to the coalition, they suggest that each member collect her *average* marginal contribution. The point of this amendment is that, if for some reason one's marginal contribution falls below the average marginal contribution of the coalition of which one is part, one will benefit from having "been accepted to become a new member of the coalition," as the authors put it. Solidarity, in this case, is tantamount to a readiness "to support some 'weaker' members" of the coalition. However, as the authors themselves acknowledge, the "social or psychological" aspects of this idea are not made explicit since solidarity comes in at the level of the solution of the game, not in the structure of its characteristic function.

2.3. A solidarity game: The force of "fixed total sacrifice"

Finally, there is the most recent appearance in Selten and Ockenfels's (1998) so-called "solidarity game." Imagine three subjects *A*, *B*, and *C* participating in the following experiment, under conditions of complete anonymity. (This means that neither the other subjects nor even the experimenter are able map an individual subject into his/her choice. Moreover, all subjects know this from the outset.) Each subject is told that she/he will win 10 with probability $2/3$, and win nothing with probability $1/3$. Before the randomization, subjects are asked to state how much of their winnings— if they happen to win the 10— they are prepared to share with the other subjects in their team of three who, out of bad luck, did not win anything. Subject *A* is, for instance, asked to write on a completely anonymous piece of paper the sum she would want to donate to *B* (or to *C*) if *A* were to win 10 and *B* (or *C*) was the only loser in the trio. Let us call this sum *X*. Then *A* is asked to select her donation to both *B* and *C* if neither *B* nor *C* were to win any money from the lottery. Let this sum be *Y* (with "losers" *B* and *C* receiving $Y/2$ each from *A*). The authors report that, out of 120 subjects (split into forty groups of three), the average values of *X* and *Y* were 2.46 and 3.12, respectively.

Moreover, 52% of the subjects chose *X* roughly equal to *Y* (up to a rounding error)— a finding the authors label *fixed total sacrifice* and show to be inconsistent with standard utilitarian altruism. The significance of this result is that, for the first time in the literature, we may discern a substantive yet intuitive difference between altruism and solidarity: Altruism requires symmetry in the way we value others' utility from money or other resources *regardless of the others' social location*; by contrast, solidarity requires a basic asymmetry in favor of some target group— i.e., in this case, the victims of bad luck in some lottery.

2.4. Requisites for a more acceptable definition of solidarity

The lesson to be drawn from the discussion of these few occurrences of something resembling "solidarity" in economics is that (i) solidarity cannot be grasped in a satisfactory manner by an axiom imposed "outside" of the explicit individual motivations of agents and (ii) solidarity seems to have an intimate link with the targeting of a specific social group. Let us now make these attributes more precise. In the next section, we go a long way to satisfy requisite (i) by suggesting various explanations of other-regarding behavior which are rooted in explicit *individual motivations*. In

sections 4 and 5 we will then attempt to analytically *distinguish* solidarity from alternative other-regarding motives, beliefs, and actions.

3. Solidarity as general other-regardingness: The motley crew of beliefs and instrumental modes of reasoning

3.1. A broad definition of solidarity

We shall begin with a broad definition of solidarity as an inclination to incur some “sacrifice” in order to benefit others. Gradually, in subsequent sections, we will refine our definition so as to single out solidarity as a more specific form of non-selfish behavior. Let us start by introducing a few elements of notation. We will say that

- a_i is the action performed by individual i who belongs to group M
- $s_i(a_i) = u_i^* - u_i(a_i) \geq 0$ is the loss of direct utility by i who performs action a_i rather than some other action which would have yielded maximum utility u_i^*
- $W_N(a_i)$ is the welfare benefit to members of some group N accruing from a_i . (Under the assumption of cardinal utilities, a particular case would be a Benthamite aggregation such that N comprises the complete human population and W_N is the average utility. Another particular case would be for the set N to contain a single person: the one with the lowest utility— a type of welfarist-Rawlsian objective.)

Definition 1. Let $I_i(a_i) = \begin{cases} s_i(a_i) & \text{if } W_N(a_i) \geq 0 \\ 0 & \text{otherwise} \end{cases}$ and $h_i(a_i) = \begin{cases} s_i(a_i) & \text{if } W_N(a_i) < 0 \\ 0 & \text{otherwise} \end{cases}$

Thus a person i belonging to some group M ($i = 1, \dots, m$) performs an act of $?$ -solidarity towards members of some group N ($j = 1, \dots, n$) if she acts in a manner which benefits them at her own expense. (Note that the intersection of groups M and N may well be nonempty.) However, if she is prepared to make a sacrifice in order to *hurt* members of group N , she is to be known as $?$ -hostile towards them.

To delve deeper into i 's motivation, we define I_{ij} as i 's expectation of what some person j expects I_i to be. In other words, $I_{ij} = E^j[E^i(I_i)]$ for $j \neq i$. Then Λ_{Ni} and Λ_{Mi} are i 's expectation of the average prediction of I_i by persons belonging to groups N and M , respectively, and I_{M-1} (resp. h_{M-1}) is her expectation of the average $?$ (resp. $?$) value that others like her, i.e., belonging to group M , will choose, or would have chosen under similar circumstances. More explicitly,

$$\Lambda_{Ni} = E^i \left[\frac{1}{N} \sum_{j=1}^N E^j(I_i) \right] = \frac{1}{N} \sum_{j=1}^N I_{ij} \quad ; \quad \Lambda_{Mi} = \frac{1}{M-1} \sum_{j=1}^{M-1} I_{ij} \quad ; \quad I_{M-1} = E^i \left[\frac{1}{M-1} \sum_{j=1}^M I_j \right].$$

Similarly, for the case of hostility, we define

$$H_{Ni} = \frac{1}{N} \sum_{j=1}^N h_{ij} \quad ; \quad H_{Mi} = \frac{1}{M-1} \sum_{j=1}^{M-1} h_{ij} \quad ; \quad h_{M-1} = E^i \left[\frac{1}{M-1} \sum_{j=1}^M h_j \right].$$

Furthermore, in a bid to introduce normative beliefs into i 's deliberation, let us define \mathbf{x}_i as i 's belief about the value of I_i that she *ought* to choose, \mathbf{x}_{ji} as i 's belief about what j believes I_i *ought* to be, and \mathbf{m}_{jN} and \mathbf{m}_{jM} as i 's expectation of average opinion amongst group N and M respectively about the value of I_i that she *ought* to choose. That is,

$$\mathbf{m}_{jM} = E^i \left[\frac{1}{M-1} \sum_{j=1}^{M-1} \mathbf{x}_{ji} \right] \quad ; \quad \mathbf{m}_{jN} = E^i \left[\frac{1}{N} \sum_{j=1}^N \mathbf{x}_{ji} \right]$$

To illustrate, suppose $I_i = \mathbf{x}_i = \mathbf{m}_{jN} = 1$, while $\Lambda_{Ni} = \Lambda_{Mi} = I_{M-1} = 0$. In this case, individual i chooses to sacrifice one “util” in solidarity with members of group N when she believes (a) the average member of group N does not expect her to do this, (b) no other member of her own group M would do likewise, (c) she *ought* to sacrifice precisely one “util” for the sake of group N , (d) none of her fellow M -members think so, but (e) members of group N think, on average, that she *ought* to sacrifice one “util” for them (even if, by (a) above, they do not predict that she will).

Definition 2. Agent i 's ?-solidarity profile is given by

$$\langle I_i(a_i) > 0 \mid \Lambda_{Ni}, \Lambda_{Mi}, I_{M-1}, \mathbf{x}_i, \mathbf{m}_{jN} \rangle$$

Of course, such a profile encompasses a great deal of individual motivations, ranging from altruism to duty or norm observance. Before we refine our definition of solidarity in the following section, we shall delineate six different motivations consistent with ?-solidarity.

To facilitate our delineation, consider two simple one-shot games: Sets N and M coincide and each person $i \in N = M$ must choose an integer a_i from the interval $[1,9]$. The first game, which we call CG (for “coordination game”), is defined by a utility function for each player as follows: $u_i(a_i) = A \cdot \min(a_j) - a_i$, $\forall j \neq i \in N$, where A is a constant. Clearly, this is an n -person coordination problem featuring an infinity of Pareto-ranked Nash equilibria belonging to the continuum $[1,9]$. The optimal strategy for any i is to choose the smallest number in $[1,9]$ which she predicts will be selected by anyone within her group; in other words, she will set $a_i = \mathbf{a}_i$ where $\mathbf{a}_i = E^i \left\{ \min(a_j) \right\} \quad \forall j \neq i \in N$. As is well known, even the slightest degree of pessimism (i.e., $\mathbf{a}_i < 9$ for some i) suffices to lead players to an inefficient outcome (see e.g. the experimental evidence of van Huyck, Battalio and Beil 1990). The second game, which we call FRG (for “free-rider game”), obtains when the minimum of all a_j -choices is replaced with the mean in the payoff

function. That is, i 's utility function is given by $u_i(a_i) = A \cdot \text{mean}(a_j) - a_i$, $\forall j \neq i \in N$, where again A is a constant and $A < n$. This simple substitution turns the game into an n -person free-rider problem in which the optimal strategy is the dominant choice $a_i = 1$. According to our definition of α -solidarity, in FRG any choice of $a_i > 1$ constitutes an act of α -solidarity, characterized by a degree of solidarity $I_i = a_i$. By contrast, in game CG we have $I_i = a_i - a_i$ provided $a_i - a_i > 0$, or $I_i = 0$ otherwise.

Let us now turn to six possible foundations, or explanations, of α -solidarity in these two games. We shall see how α -solidarity can be accounted for by various assumptions about individual motivations, and we shall then argue that while these may be satisfactory at some level, none of them *ultimately* grasps the idea of solidarity in its specificity.

3.2. Instrumental explorations of solidaristic attitudes

(a) Solidarity as team-reasoning amongst instrumentally rational egoists

In CG the only impediment to successful coordination is fear that it will not occur. Instrumentally rational agents, whose utility functions above define their objectives fully, have no means of convincing themselves that coordination will occur. Indeed, even common knowledge of instrumental rationality (hereafter CKR) is incapable of rendering $a_i = 9$ the optimal strategy. Thus such agents will not have any reason to select $I_i > 0$ even though common knowledge that $I_i > 0$ for all i would lead to the Pareto-superior Nash equilibrium. One promising way out has been discussed by Sugden (1993) and, more recently, by Bacharach (1999): Provided members of group N recognize that they are *members of a team*, they may have no coordination problem to solve since they “do not need to form expectations about one another’s actions.” In this sense, agents select $a_i = 9$ regardless of a subjective expectation a_i which might be induced empirically (and proven to be less than 9).

Sugden defines an outcome as *focal* if it can be identified by the expression “Best for all” (hereafter BFA). Clearly, in the case of game CG above, $a_i = 9 \forall i$ satisfies this definition. The problem is that CKR cannot lead instrumentally rational agents to it with any degree of certainty. To accomplish this, we need a common knowledge of a different sort. Let us call it *common knowledge of α -solidarity* (denoted CKS α) and define it as follows:

Definition 3. In games with multiple Pareto-ranked Nash equilibria $a_i = x$ (where $x \in [a_{\min}, a_{\max}]$ and the strategy $a_i = a_{\max} \forall i \in N$ would yield maximum utility for each i , that is, $a_i = a_{\max}$ is consistent with BFA), there is common knowledge of α -solidarity (CKS α) if $E^i(I_j) = \epsilon$, $\forall i, j \in N (i \neq j)$, where ϵ is positive but arbitrarily small.

It is then possible to show that CKS α induces perfect coordination if all agents are instrumentally rational and know that all others are:

Proposition. Under CKR and CKS_?, the unique equilibrium in game CG is the Parto-superior Nash equilibrium.

Proof. CKS_? implies common knowledge of the following fact: if i anticipates that the common expectation of the minimum choice will be less than 9 (i.e., $a_i < 9$), then everyone else will be prepared to select a number just above that common-knowledge expectation a_i . But if this is true, then i will want to set $a_i > a_j$, something which is commonly known under CKR. As this train of thought applies to all values of $a_i < 9$, it follows that the only choice which is consistent with both CKR and CKS_? is the Pareto-optimal Nash equilibrium $a_i = 9 \quad \forall i \in N$. ?

In other words, in game CG, each player will be drawn to the collectively *and* individually optimal outcome as long as each is confident that everyone knows that (a) each member of the *team* is instrumentally rational and (b) *if* the anticipated minimum “contribution” might be less than the one which would have been best for all (i.e., $a_i < 9$), *then* each player will be prepared to make an infinitesimally small sacrifice (out of solidarity with the *team*) by setting $a_i = a_i + \epsilon$ (or, equivalently, set $I_i = \epsilon$). This mild form of solidarity, as long as it is commonly known, is sufficient to explain why in some cases coordination problems are solved instantly (e.g., soccer players who seem to have a “sixth sense” for one another’s on-pitch location) whereas in other cases the same problem leads to coordination failure (see van Huyck *et alii* 1990). The key seems to be the process which encourages people to feel they *belong* that is, the process that endows them with CKS_?.

Turning now to the second game, FRG, team reasoning amongst instrumentally rational, and egoistical, players cannot explain displays of ?-solidarity. The reason is that setting $a_i = 1$ is a dominant strategy and, therefore, CKS_? cannot dissuade agents from doing what maximizes their utility uniquely. Some authors have argued (see Gauthier 1986) that $a_i = 1$ is an irrational choice since it would be profitable to develop a disposition towards cooperation. However, we believe that in game FRG, values of I_i significantly greater than zero cannot be explained unless agents are motivated by something beyond the utility functions u_i postulated so far.

(b) Solidarity due to Humean natural sympathy or utilitarian altruism

Postulating the self as a series of cocentric circles, the Humean agent thinks of others’ interests as her own in inverse proportion to the psychological distance between her and “them.” Accepting without debate the reduction of Humean passions to utilities (which is something which, a.o., Sugden and Hollis 1993 warn us against, but which is convenient for the point we are trying to make at present), a simple explanation of $I_i > 0$ in either of our games is that i cares not only about her utility but also about that of the rest of the group. In effect, if i ’s (meta-)utility function is given by $U^i[u_i(I_i), W_N(I_i)]$ and if $\Lambda_{Ni} > 0$, then she chooses her degree of ?-solidarity to equal

$$I_i = \operatorname{argmax} \{ U^i [u_i, W_N] \mid \Lambda_{Ni} \}$$

Effectively such concern for the utility of other members of the group abolishes both the coordination and the free-rider problem in proportion to the valuation of others’ welfare (that is, to

$\forall U^i / \forall W_N$). Traditionally, the literature posits that a utilitarian altruist treats others symmetrically (e.g., W_N is a function whose arguments U^j are weighted equally).

(c) *Kantian solidarity*

A Kantian agent asks herself, “Which strategy would be best for me if *all* persons in my situation (including myself) were to follow it?” Clearly, the answer here is that $a_i = 9$ in both CG and FRG. The question then is, “Granted that this is so, why should I be interested in following this strategy in view of the fact that it may lead to individually suboptimal payoffs?” Such instrumental thoughts are very tempting. For indeed, if all you care about is your utility, then $a_i = 9$ may not be a utility-maximizing choice in CG (if $a_i < 9$) and is *never* utility-maximizing in FRG. Kant’s reply is that to be rational means to have a capacity to overcome such tempting thoughts and, instead, to understand that in social contexts, the rational person recognizes her duty to do *what is right* as opposed to what is expedient. And if she benefits in the end (since we may note that $a_j = 9 \quad \forall j \in N$ leads to the largest possible utility per player in CG, and to the Pareto-superior amongst all symmetrical equilibria in FRG), this is *not* the reason why it was rational for her to her duty— i.e., to choose $a_i = 9$ — but merely a satisfying byproduct.

An alternative way of interpreting this line of thinking is to say that Kantian solidarity is the urge one might feel to behind Rawls’s (1971) “veil of ignorance,” where each individual optimizes selfishly under uncertainty and subject to infinite risk aversion. The result of such optimization would procure a degree of ?-solidarity equal to

$$I_i \equiv x_i = \max \left[\min_{k=1, \dots, n} u_k \right], \text{ irrespective of } \Lambda_{Ni}, I_{N-1}, \text{ or } m_{iN}$$

In other words, a Kantian agent’s ?-solidarity makes itself felt in the form of sacrifices performed in the *line of duty*, that is, irrespective of (a) what she predicts to make out of it, (b) what she predicts others in positions similar to her to sacrifice, or (c) what others expect her to do. Contrary to the Humean case, our Kantian player treats as irrelevant both the degree of natural sympathy to others *and* her beliefs about others’ beliefs.

(d) *Conformity with the factual expectations of others*

In line with significant strands in sociology and psychology, Olson (1965) suggests that persons are motivated by an urge to “win prestige” amongst their peers. Becker (1974) adds the fear of being scorned. Such motivation would lead an agent to select $I_i = f_i(\Lambda_{Ni})$, $f_i > 0$, i.e., to choose a degree of ?-solidarity that depends positively on the average prediction she thinks her peers make as to her ?-solidarity. In other words, agents may act solidaristically in order not to disappoint (their own perception of) average factual expectations— even if, contrary to the Humean case, she does not care for their group welfare. Interestingly, a solidaristic agent may wish that she were *not* expected to be solidaristic by her peers although, *given that they do expect her to be so*, her utility is maximized by being solidaristic. (In connection with this, see what Geanakoplos, Pearce, and Stacchetti 1989 define as “psychological games” in which similar scenaria emerge in equilibrium.)

To illustrate this case of solidarity led by “public expectations,” suppose that *i*’s utility is given by

$$U^i = \begin{cases} u_i(I_i) - \Lambda_{Ni} & \text{if } I_i \neq \Lambda_{Ni} \\ u_i(I_i) & \text{otherwise} \end{cases}$$

Now suppose that in game FRG, $\Lambda_{Ni} = 8$ so that she believes to be expected to sacrifice up to 8 “utils,” and $a_i = 2$, so that she is expecting the average choice of solidarity in her group to equal 2. Then, unless she chooses to be fully solidaristic, i.e., to set $a_i = 9$, her utility will equal $2A-8$. On the other hand, if she acts selfishly and sets $a_i = 1$, she will collect utility $2A-9$. Thus she will be compelled by her perception of the expectations of others to make sacrifices. However, if she had a choice over what others expected of her, she would rather they predicted no solidaristic attitude on her part at all, in which case she would not display any (i.e., $I_i = 0$) and her utility would rise to $2A-1$. This confirms the suspicion that solidarity whose purpose is to carress others’ eyes and ears is rather hollow.

(e) *Conformity with the normative beliefs of others*

This is a mere variant of (d), with others’ normative beliefs replacing their factual ones in i ’s motivation. The latter then selects $I_i = g_i(\mathbf{m}_{jN})$, $g_i' > 0$, that is, she chooses a degree of ?-solidarity which depends positively on the average opinion she perceives her colleagues to hold regarding the degree of ?-solidarity she *ought* to choose. This is a clear case in which I sacrifice for others only because I think that they believe this is what I ought to do. (The utility analysis in (d) carries over here unchanged.)

(f) *“Biblical” solidarity*

Imagine person i plans to make sacrifices for her group *because* she thinks that other members of her group will make sacrifices for her (as well as for one another). Moreover, she thinks this *because* she believes they expect her (and others like her) to make sacrifices. Now suppose the same person would have been eager to sacrifice part of her utility in order to *hurt them* if she held a belief that they will be making sacrifices to *hurt her* (and each other). This is the nasty underbelly of ?-solidarity, which we defined earlier as ?-hostility: a tendency of a person to make sacrifices for those whom she thinks are similarly disposed towards her— but also a readiness to suffer personal costs in order to punish them if she predicts that they are trying to hurt her. (Rabin 1993 introduces this concept and derives a “fairness equilibrium” for a wide class of games.)

To see briefly how this motivation can be internalized, consider the following utility function: $U^i = u_i(I_i) + \mathbf{g}[I_i I_{N-1} + \mathbf{h}_i \mathbf{h}_{N-1}]$, where $\mathbf{g}_i > 0$ is some constant which reflects i ’s relative valuation of the means by which certain payoffs are produced. Such a maximand instructs i (as long as \mathbf{g}_i is large enough) to set $I_i > 0$ if she anticipates $I_{N-1} > 0$ (which of course goes along with the expectation that $\mathbf{h}_{N-1} = 0$). However, it also urges i to set $\mathbf{h}_i > 0$ whenever she expects $\mathbf{h}_{N-1} > 0$ (and $I_{N-1} = 0$).

3.3. Towards “deeper” solidarity

These explorations into possible foundations of ?-solidarity show that seemingly solidaristic behavior, defined simply as the readiness to sacrifice some personal utility for the sake of others (whether within one’s own group or within another group), can easily be explained by a combination of internalized (meta-) utility functions and of beliefs about duties, others’ beliefs, and so on. Some— or perhaps most— economists might stay content with such explorations: They are, indeed, able to explain quite a lot of instances of other-regarding behavior and attitudes, and they show that both sociological and philosophical basics can be integrated into a welfaristic perspective (albeit a broadened one).

However, our feeling is that solidarity is more specific than that. It must have something to do with a less conditional and less instrumental *predisposition* to help others in specific social situations. This is what we start exploring now, by first “deepening” our definition of solidarity in the next section, and by subsequently restricting this deeper attitude to specific social situations (linked to the notions of oppression and arbitrary social power) in section 5.

4. Deeper solidarity: Other-regardingness without belief-dependence and instrumentality

4.1. Life boats and union strikes: The issue of entitlement

Sugden (1993) discusses the case of the British Lifeboat Service, which is financed entirely through public donations. Why do people contribute money to the Service? Sugden correctly points out that utilitarian altruism cannot account for this economic fact. The reason is that if donors are motivated by an interest in ensuring that the Service has sufficient funds to perform its life-saving duties, they ought to think of each pound they contribute as a perfect substitute for each pound contributed by someone else. Yet the econometric evidence (see also Sugden 1982) contradicts this hypothesis. This is a finding which resonates with that in Selten and Ockenfels (1998), namely, that donations— by winners of some lottery to losers in the same lottery— largely reflect a tendency of donors to decide how much to give in aggregate *independently* of how much the recipients are to collect from others, or even of how their contributions are to be divided amongst a number of recipients (this is the hypothesis of “fixed total sacrifice” outlined at the end of section 2 above).

Sugden’s (1993) explanation of this failure of altruism to account for most donation patterns is that standard economic models assume that the agent only “looks for reasons that bear on *him*”; the players seem to be overlooking reasons that bear on *them*.” Thus the author calls for a different approach which acknowledges the fact that people can think in terms of collective goals and responsibilities. They can reason as members of an organic whole to which they rationally choose to belong, rather than as isolated selves in search of maximum utility *given their expectations about others’ actions* (for a model of such reasoning, see Bacharach 1999). However, the philosophical

difficulties of conceptualizing the individual as a team member notwithstanding, Sugden's team-thinking begs an important question regarding the Life Boat Service phenomenon.

If agent *i* is an occasional sailor who may, one day, find herself at sea and in peril, she has good cause to think of herself as a member of some sea-faring "team" and act accordingly when the Life Boat Service representatives call in and ask for a donation. However, why would *i* ever contribute if *she does not envisage ever leaving dry land*? And, to sharpen the point, why would she ever contribute a pound without treating *her* pound as a perfect substitute to the pound contributed by someone else? One facetious answer is that she may have a friend or neighbor who sails occasionally. This is fine, but what if she does not, and still contributes a fixed sum independently of her expectations of the Service's total income? (In fact, despite the lack of hard evidence, we strongly suspect that a large number of donors to the Service never expect to be in need of assistance at sea; yet they still contribute substantial funds.)

One simple explanation is that she might still contribute, not out of altruism, but *out of a kind of solidarity not subsumable in general other-regardingness*. She gives to the Service because she reflectively identifies an *image* which, somehow, makes it a worthy cause— an image which allows her to envision a set of persons who are *entitled* to her solidarity *because* of that image. She may even feel no natural sympathy, or altruism, towards these particular people as individuals whom she is about to benefit (e.g., rich round-the-world yachtsmen with more money than sense) but, nonetheless, she is drawn to make a contribution by something they have in common and which entitles them to *her* money: the objective fact that they will be *shipwrecked*. This is exactly what occurs for the donations observed by Selten and Ockenfels (1998); it may very well be that, in their solidarity game, winners of a lottery are prone to donate money to losers not because of some concern about how much money "fellow players" leave the laboratory with (including, altruistically, both losers *and* winners like themselves, since there is perfect substitutability), but through a feeling of solidarity with *the losers as losers*— a feeling which breeds an obligation within the agent to give them a share of *her* winnings.

As a last example, consider the case of a strike in sympathy with dismissed colleagues. Of course, an important aspect of such campaigns (powerfully publicized by union leaders) is the fear that their own jobs might be the next to go. However, workers who are about to retire anyway, or who feel safe in their jobs (or who predict that, even if other colleagues were to strike on their behalf, it would be in vain), have been known to strike against layoffs in a manner that reduces the welfare not only of themselves but, additionally, of *all* workers (both retained *and* sacked) on average. Since altruism cannot explain the latter, why do they strike? Our simple explanation here is that they strike in solidarity with those who lost their jobs because the latter are *entitled* to their solidarity purely on the objective basis that *they lost their jobs*.

Of course, the above raises the question of how one selects the trait or image (e.g., "shipwrecked," "loser in a lottery," "redundant worker," "refugee," "victim of torture," and so on) which entitles the "other" to one's solidarity. Briefly, the simple Humean answer is that such traits evolve in the same manner as conventions in indeterminate social interactions. For instance, the

solidarity traits which gather more adherents are the ones which will emerge by evolving. Their evolutionary fitness or stability then depends on their effect on the evolutionary fitness of the communities in which they have sprung up. We are not at all convinced that this is the whole story— in fact, genuine solidarity is more like a break away from evolutionary entrenchment. However, before we take up this crucial issue in the next section, we shall first codify the above *entitlement argument* and use the discussion of the British Life Boat Service as a pretext for refining our definition of solidarity.

4.2. A refined definition of solidarity

Suppose that person i will never go anywhere near the water; that is, $i \in M$ and $M \cap N = \emptyset$, where M is the set of non-seafaring persons and N is the set of those who might one day make use of the Life Boat Service. Yet she contributes some money to the Service ($I_i > 0$), which results in an infinitesimal increase in the reference welfare W_N of sailors. Why did she do this? Let us refer back to the six possible explanations developed in the previous section. Explanation (a) does not apply since i is not a member of the team whose goal she helps to further. Explanations (c) and (f) do not apply either since our donor neither expects any gains for herself if her donation were universalized¹ nor sees any room for reciprocity.

This leaves us with explanations (b), (d), and (e). Explanation (d) may hold as long as we identify solidarity with mere self-sacrifice, but it hardly qualifies as a case of deeper solidarity. Moreover, it is unlikely that those who might need the Service (that is, members of group N) would expect i to contribute. Explanation (e) seems more poignant: Person i contributes because she thinks that the Service's collectors believe she ought to contribute. However, with so many charities competing for her money, i would become destitute unless she was driven by something more focussed, beyond the mere moral expectations of those who request her support. Moreover, (e) is a delusion of solidarity since, as we have shown, i 's optimal scenario is that she does not contribute anything as long as she does not think she is expected to.

Therefore, unless explanation (b) comes up trumps, agent i 's contribution to the Life Boat Service will remain a motivational mystery. Now as we have seen already, it seems that a utilitarian account of her choice, à la explanation (b), may prove consistent with the evidence (as well as with this hypothetical discussion) provided i draws utility disproportionately from the benefit she has bestowed upon the potentially shipwrecked. But of course, there is no reason why we ought to confine ourselves to the philosophical egoism of instrumental accounts. Indeed, we will go along with Hollis (1987) and interpret our donor's behavior as a rational response to an *external* reason for action— as opposed to an action drawn from one's passions or one's utility ordering— which

¹ Note the "coldness" of Kantian solidarity: Any decision to set $I_i > 0$ is due *not* to some sympathy by i for the plight of others, but simply to the recognition that if all agents were to set $I_i > 0$, agent i would herself be better off. Indeed, if i could "will" that any of her behavior be universalized, she might well choose instead to give money to cancer research (since in that area she might benefit from universalized funding) rather than to the Life Boat Service.

compels her to fulfil an obligation to those about to drown. We may therefore suggest the following definition of solidarity, stripped of the possible explanations we submitted in section 3:

Definition 4. A person i acts out of α -solidarity if a parameter α can be defined as follows: $\alpha = I_i > 0$ if the following two conditions (I) and (II) apply; if these two conditions do not apply, then $\alpha = 0$ even if $I_i > 0$. The two conditions are:

- (I) *Autonomy:* I_i is independent of beliefs $(\Lambda_{Ni}, \Lambda_{Mi}, I_{M-1}, I_{N-1}, m_{jN})$.
- (II) *Non-instrumentality:* The choice of the set N of persons towards whom i 's solidarity is directed is irreducible to the maximization of i 's utility function.

Condition (I) distinguishes α -solidarity from those instances of α -solidarity in which sacrifices are made in accordance with either reciprocity or some custom or norm. For if I_i is not determined by i 's first-order or second-order expectations (either factual or normative), it can only be explained as a result of an autonomous deliberative process. Thus, norm or custom-following *à la* Akerlof (1980) or Varoufakis (1989) do not qualify as examples of α -solidarity. In this sense, nor does the concern for one's image within a group mentioned by Olson (1965) or Becker (1974). In short, α -solidarity is irreducible to social norms or public morals. Condition (II) prescribes that deep solidarity necessitates a capacity to act solidaristically towards some group N for reasons which cannot be fully reduced to the maximization of one's utility (e.g., " N is the group it is in my best interest to support") subject to constraints (e.g., "If I don't support members of N some of them will beat me up") and calculative beliefs (e.g., "I choose to support members of N because nowadays it is politically correct to do so and I don't want to look offbeat"). It can be interpreted in at least two ways.

One possible interpretation of condition (II) is explicitly non-instrumental. It relies on i choosing both the group N and the value α on the basis of some principle of choice external to her preferences (see Hollis 1987). A second interpretation envisages a two-stage decision process: In the first stage, i selects the set N toward which she must be solidaristic (e.g., the shipwrecked, the refugees or the sacked colleagues) on the basis of a principle external to both her preferences and (by condition I) any social expectations. Once N has been selected, i chooses the degree of solidarity

$$\alpha = \arg \max (U^i [u_i(I_i), W_N(I_i)] \mid N \text{ selected externally})$$

Conceptually, this two-stage process resembles Frankfurt's (1971) idea of a two-tier deliberation process for rational agents: one—the lower tier—where preferences determine outcomes and another—the higher tier—in which principles external to preferences decide which of the lower-tier deliberations should be "trumped" and which allowed to pass. But why claim that "deeper" solidarity cannot be accounted for on purely (meta-)utilitarian grounds? The reason is that if i maximizes her utility using set N and the value of α as choice variables, i.e., if

$$\{c_i, N\} = \arg \max_{I_i, N} (U^i [u_i(I_i), W_N(I_i)]),$$

we are back to an explanation of *i*'s attitude along the lines of utilitarian altruism, thus rendering redundant any notion of deep solidarity. The latter is, however, relevant— and its loss undesirable— when a utilitarian regard for others is not able to explain why, say, a non-sailor without special ties or natural sympathy with sailors might contribute to the Life Boat Service without being interested in knowing how much others have contributed to it. Furthermore, fully utilitarian non-sailors would find it hard to imagine why the optimal choice of *N* would be the group of shipwrecked sailors— as opposed to, say, cancer patients. Thus we must accept either that their contributions are the result of bounded rationality or that they are due to something akin to what we have termed δ -solidarity. Sugden (1993) concurs: “By restricting its attention to instrumental rationality, economics is neglecting a potentially significant form of human motivation.”

4.3. The import of the refined definition

Let us conclude this section with the two puzzling examples already discussed, and on which it now turns out δ -solidarity sheds some new light. The first concerns the results of Selten and Ockenfels. In their solidarity game, they show that their finding of a “fixed total sacrifice” is inconsistent with any standard utilitarian-altruistic optimization program. However, it is straightforward to demonstrate the compatibility of δ -solidarity with the FTC observation.² The second example concerns workers who may strike in support of dismissed colleagues even if they themselves do not believe that such action will have any positive impact on *their* future wages or job security.

To tease this point out, suppose there are *n* employed workers. At *t*=1, the firm announces that ***n*** of them will lose their jobs at *t*=2 when the list of the sacked workers is to be announced. The trade union organizes a strike to begin at *t*=2 in protest against the dismissals. Let *w* be the present value of keeping one's job— to be collected by each of the *n* – ***n*** retained workers— and let *g* be the part of *w* that each of the retained workers are prepared to sacrifice— through participation in the union's action— in support of their sacked colleagues. What proportion of *w* will a retained worker choose to sacrifice during such an industrial dispute? It is simple to demonstrate that a utility function in which a worker cares about the welfare of her *n*–1 colleagues will not lead her to strike for a period *which is independent of the number of sacked workers, n*. For doing so would imply that she values the expected loss of \$1 by a sacked colleague more than the expected loss of \$1 by a retained (but striking) colleague. On the other hand, δ -solidarity would indeed allow for such a preference provided worker *i* has an external reason to think of her sacked colleagues— but not, by the same token, her fellow striking workers— as the exclusive target of her support. Once she selects these

² According to δ solidaristic players who are asked to decide how much to give to losers of the lottery *if* they win, one first has to select the “losers” group *N* whose welfare (or payoff) one feels responsible for (a choice which has to be non-instrumental), and *only then* decide how much to sacrifice for the members of this group.

dismissed n workers as her reference group N , she will choose (instrumentally or otherwise) her level of g and hence her degree of solidarity $\alpha > 0$.

It would of course be absurd to imagine that a large proportion of any population are capable of α -solidarity towards sailors, sacked colleagues, refugees, etc. However, all other theories of other-regarding behavior, even the most cynical ones which argue that people show solidarity in order to keep appearance up, are “parasitic on moral theories that enjoin us to behave in ways that are not instrumentally rational” (Sugden 1993). Thus, the presence of even a small percentage of persons capable of α -solidarity may be the necessary initial condition for some bandwagon to start rolling. (This would alleviate what in section 2.1 we termed the “mystery” inherent in theories such as those of Akerlof 1980 or Varoufakis 1989.) If this is indeed so, then in our union example the firm may avoid marginal downward adjustments in its employment level— the reason being, of course, that workers will strike with the same ferocity in solidarity with 10 or 1,000 sacked colleagues. Either the firm will fire a large number, or none at all. ³

4.4. Towards “genuine” solidarity

It is no accident that our examples of α -solidarity feature victims of unfortunate circumstances. Indeed, solidarity seems to distinguish itself from alternative types of other-regarding motivations first and foremost in that it gravitates towards specifically identified sufferers— which we called the reference group N . But what if N is a group of mafiosi or, much worse still, a group of SS-officers? Instinctively, one feels very strong unease with the notion of solidarity *with* an oppressor. Yet solidarity *amongst* oppressors is not unknown. And there is nothing in our definition of α -solidarity which automatically precludes its being used to understand the process by which villains select a reference group N (which, in addition, may or may not coincide with their own group M of villains) before making sacrifices for its members. Our next refinements of the definition of solidarity are going to allow for such crucial distinctions to be drawn.

5. Genuine solidarity: Solidarity with the victims of arbitrary social power

We will start by utilizing evolutionary game theory in order to explain what we mean by systematic and arbitrary social asymmetries which allow some agents (the “oppressors”) to have social power over others (the “victims”), who are thus pushed into a state of subservience. This will be done in section 5.1. Then, in section 5.2, we will distinguish between collusion amongst oppressors (which we label α -solidarity), solidarity amongst victims (β -solidarity), and genuine solidarity (γ -solidarity).

³ From the point of view of union-employer bargaining theory, this implies that, at least within some range of employment levels, wage and employment bargains may indeed be efficient, since unions will have a credible threat in case of marginal adjustments away from the Leontief contract curve.

5.1. An evolutionary account of arbitrary social power

Consider the game pictured in figure 1, which is a variant of the Hawk-Dove game featuring an additional dominated cooperative strategy. There are three Nash equilibria in this game: two in pure strategies $[(h, d), (d, h)]$ and one in mixed strategies $[\text{Pr}(h)=1/3, \text{Pr}(d)=2/3, \text{Pr}(c)=0]$.

	<i>h</i>	<i>d</i>	<i>c</i>
<i>h</i>	-2, -2	2, 0	4, -1
<i>d</i>	0, 2	1, 1	0, 0
<i>c</i>	-1, 4	0, 0	3, 3

Figure 1

The Hawk–Dove–Cooperate Game

Now suppose that there are two identical but distinct populations M ($i=1, \dots, m$) and N ($j=1, \dots, n$) whose members play our game as follows. In each round two players are randomly matched, irrespectively of which group they belong to, and they play anonymously, with the single exception that each knows which group her opponent belongs to. In each round players are matched against fresh opponents. Anonymity and repeated play against new opponents ensures that no trigger strategies can emerge and that each round can be thought of as a one-shot game. However, repetition provides agents with information about general behavior, with the consequence that beliefs and strategies evolve across the population of players. Moreover, the fact that each player can condition her behavior on the group membership of her opponent leads to two evolutionary equilibria (*EE1* and *EE2*) featuring built-in patterns of discrimination. For K being either group M or group N , let us define the following probabilities:

p_K	$\text{Pr}(i \in M \text{ plays } h \text{ against an opponent belonging to group } K)$
q_K	$\text{Pr}(i \in M \text{ plays } d \text{ against an opponent belonging to group } K)$
r_K	$= 1 - p_K - q_K$
\mathbf{p}_K	$\text{Pr}(j \in N \text{ plays } h \text{ against an opponent belonging to group } K)$
\mathbf{q}_K	$\text{Pr}(j \in N \text{ plays } d \text{ against an opponent belonging to group } K)$
\mathbf{r}_K	$= 1 - \mathbf{p}_K - \mathbf{q}_K$

The two equilibria are then the following:

$$\begin{aligned}
 \text{EE1} \quad & p_N = \mathbf{q}_M = 1 ; \quad p_M = \mathbf{p}_N = 1/3 ; \quad q_M = \mathbf{q}_N = 2/3 \\
 & (\text{thus } r_K = \mathbf{r}_K = 0 \text{ for either } K)
 \end{aligned}$$

$$\begin{aligned}
 \text{EE2} \quad & \mathbf{p}_M = q_N = 1 ; \quad p_M = \mathbf{p}_N = 1/3 ; \quad q_M = \mathbf{q}_N = 2/3 \\
 & (\text{thus } r_K = \mathbf{r}_K = 0 \text{ for either } K)
 \end{aligned}$$

In summary, the group identity of one's opponent provides a toehold for asymmetry only in meetings of players belonging to different groups. In same-group meetings, the unique evolutionary equilibrium coincides with the mixed-strategy Nash equilibrium ($p_M = p_N = 1/3$; $q_M = q_N = 2/3$). However, in cross-group meetings, one of the two groups will eventually become "dominant" in the sense that, when its members meet players belonging to the other group (let us label the latter "subservient"), the "dominant" player will play h and the "subservient" one will play d . Which of the two groups— M or N — evolves as the dominant group is a matter of pure serendipity. (See Weibull 1995 for a formal analysis of this case of two-dimensional evolution.)

Suppose now that M emerged as the dominant group, that is, equilibrium $EE1$ became established. Thus, whenever $i \in M$ plays against $j \in N$, i plays aggressively and j retreats (they gain payoffs 2 and 0 respectively). When either an $i \in M$ plays a fellow member of M , or when a $j \in N$ is matched with another N -player, each agent plays aggressively with probability 1/3, opts for the d strategy with probability 2/3, and stays clear of the cooperative strategy. This simple game will now help us to define some crucial notions which will put us on the track towards a really specific and ethically acceptable notion of solidarity, stripped not only of instrumentality and belief-dependence but also of the risks of collusion between oppressors which remained inherent in ?-solidarity.

5.2. Refined types of solidarity

Let M and N be the dominant and subservient groups, respectively, under some two-dimensional (discriminatory) evolutionary equilibrium— e.g., $EE1$ above. We can then define the following notion of arbitrary social power:

Definition 5. Let there be two players i and j , who are identical in every respect other than their group membership ($i \in M$ and $j \in N$). Player i is said to exercise *socially evolved, arbitrary, relative power*, denoted Π_{MN} , over player j if the following holds: In meetings between i and j , an evolutionary equilibrium of a symmetric game is established such that

- (a) i 's expected payoff is greater than j 's, and
- (b) j 's expected payoff is lower than it would have been under the mixed-strategy Nash equilibrium.

Armed with this definition of abusive social power, we can define solidarity among the beneficiaries of this power, solidarity among its victims, and finally what we will call "genuine" solidarity.

Definition 6 (Solidarity among the beneficiaries [?solidarity] of Π_{MN}). When $i \in M$ is matched with $k \in M$ during round t , and both i and k exercise Π_{MN} over all $j \in N$, then i 's ?-solidarity toward k during round t equals $j_{it} = q$ if $q > 0$. Otherwise $j_{it} = 0$.

Definition 7. (Solidarity among the victims [?solidarity] of Π_{MN}). When $i \in N$ is matched with $k \in N$ during round t , and both i and k are victims of Π_{MN} in meetings with all $j \in N$, then i 's ?-solidarity toward k during round t equals $s_{it} = d$ if $d > 0$. Otherwise $s_{it} = 0$.

Definition 8. (Genuine [or ?] solidarity). When $i \in M$ is matched with $j \in N$ during round t , and therefore i exercises Π_{MN} over j , then i 's ?-solidarity toward j during round t equals $\xi_{it} = d$ if $d > 0$. Otherwise $\xi_{it} = 0$.

Let us see what new light these definitions shed on our understanding of solidarity in economic models. Social evolution favors person i by selecting her group, M , to be dominant. In other words, it endows her with arbitrary social power Π_{MN} over any member of the other group, N . Notice that Π_{MN} is arbitrary because it is independent of personal characteristics— since individuals across groups are assumed identical and the game is symmetrical— and is distributed solely by virtue of one's group membership. Any display of ?-solidarity among those who have been fortunate enough to be endowed with Π_{MN} is defined as ?-solidarity— or, alternatively, solidarity amongst the socially dominant. By contrast, ?-solidarity among the victims of social evolution is labelled ?-solidarity. Thus, we are able to discriminate between different instances of (apparently similar) ?-solidarity on the basis of the historical path of interaction which determined the social affiliation of the agents.

Genuine (or ?-) solidarity becomes possible when a dominant i comes across a subservient j towards whom she acts in solidarity even though she has the opportunity of exercising power Π_{MN} over him. More precisely, under an evolutionary equilibrium such as *EE1*, agent i anticipates agent j to acquiesce (i.e., to choose d) because— as i predicts on the basis of the established evolutionary equilibrium— j expects aggressive behavior on the part of i . If, such beliefs notwithstanding, agent i selects d , we have a case of ?-solidarity; in this precise case $\xi_i = 1$ so that i sacrifices one “util.” In short, genuine solidarity refers to instances in which a dominant player chooses to sacrifice some personal gain *by abstaining from using her social power*. Instead of taking advantage of a person who fully anticipates to be exploited, a ?-solidaristic agent is prepared to play a dominated strategy d in defiance of the evolved convention which casts her in the role of “exploiter.”

Interestingly ?-, ?-, and ?-solidarity are all inconsistent with common knowledge of their existence. (In fact, if they were to become common knowledge, condition (I) in the definition of ?-solidarity would cease to hold.) Indeed *if* they began to spread within the overall population, players would begin to anticipate a solidarity-drive towards the (d, d) disequilibrium outcome, with the result that their sacrifices would resemble Rabin's (1993) “fairness equilibrium,” an outcome consistent with our CKS_? (see definition 3) and our explanation (*f*) of biblical solidarity. However, when players are drawn through CKS_? towards outcome (d, d) , at some point they will automatically switch to the Pareto-superior ?-solidarity equilibrium (c, c) .

It is important to note that, in this evolutionary game, if i possesses social power Π_{MN} , her ?-solidarity has a capacity to contribute to such a ?-solidarity equilibrium (c, c) , an equilibrium under which she would suffer a substantial reduction in her future stream of payoffs from meetings with

members of the group of victims N (i.e., those without Π_{MN}): In each meeting with a player $j \in N$ she would collect payoff 3 rather than 4 (which she is guaranteed from such meetings on the basis of her access to Π_{MN}). By contrast, both α - and β -solidarities promise to increase average payoffs by eliciting some α -solidarity equilibrium (c, d) : In meetings between either dominant or subservient players, the expected payoff to each player from a non-solidaristic Nash equilibrium in mixed strategies (which *is* the unique evolutionary equilibrium in same-group meetings) is given by $2/3$; by contrast, if a α -solidaristic equilibrium (c, d) emerges, each player will collect a payoff of 1. This is why we use the epithet “genuine” to describe solidarity toward the victims of discrimination: Agents acting out of α -solidarity are effectively antagonistic towards the evolved convention which privileges them arbitrarily over others. If their solidaristic plans prove successful, they will see their own long-term payoffs decline substantially. Why would they ever desire this outcome, which no amount of instrumental or expectation-oriented thinking can ground? The answer is simple: out of *genuine solidarity* with victims of their own arbitrary, and therefore illegitimate, power.

Unlike α -solidarity which is compatible with Victorian philanthropy, β -solidarity is not. Genuine solidarity, unlike both altruism and philanthropy, relates to an urge to make personal sacrifices for those who, for reasons which are simultaneously social and impersonal, find themselves systematically discriminated against. What makes this *genuine* is that such solidarity is never exhausted by endeavors to diminish their suffering. More poignantly, it involves personally costly efforts to undo the sources of their systematic disadvantage, even if this means undoing by the same token the sources of one’s own privileges.

Some recent experimental evidence relating to the game above (see Varoufakis 1998) conveys the pessimistic view that genuine solidarity, though not absent, is not widespread. Perhaps genuine solidarity requires debate and social interaction before it takes root— both of which are rendered impossible in the laboratory. One surprising result, however, was that β -solidarity occurred far more frequently than α -solidarity. Indeed, once the evolutionary equilibrium had stabilized, almost 90% of the meetings between subservient players yielded mutual cooperation, as compared with a mere 3% in meetings between dominant players. Thus it seems that β is more likely among the victims of arbitrary social power than among its holders.⁴

6. Conclusion

Hurley (1989) castigates *homo oeconomicus* for lacking the mental capacities effectively to engage in the bewildering enterprise of acting in a manner *organically* consistent with the objectives of the team to which he belongs. In this paper we have taken Hurley’s theme further: We have argued that a rational person can also assess her actions in terms of their consequences for members of a group/team toward which she may (a) be indifferent at the individual level and to which she may

⁴ As usual, Aristotle had it all figured out a long, long time ago when, in his *Politics*, he asserted bluntly that “The weak are concerned about morality and justice. The strong do what they must.”

(b) not relate, i.e., neither belong to their group nor even wish she did, but which (c) share a trait (e.g., they are Kurdish or East-Timorian refugees) which *entitles* them to her concern and sacrifice.

If she is indeed indifferent to them *individually*, her solidarity cannot be endorsed unstrenuously by what Hollis (1998) has called “philosophical egoism.” Granted that many— even most— acts which *seem* solidaristic in reality spring from self-seeking motives (as we saw in section 3), their occurrence is nonetheless testimony to a wide cognition of the *possibility* of deeper solidarity (as defined in section 4). Thus, some humans, some of the time, *must* be capable of selfless sacrifices, moved neither by altruism nor duty, nor by mere collusion of the powerful, but by a fierce repugnance for the suffering caused by some accident of nature of of social evolution (as we saw in section 5).

Empiricism cannot help us to sort decisively the fake from the genuine acts of solidarity. Yet this does not lessen the importance of exploring philosophically the notion of genuine deep solidarity whose very *possibility* gives bogus, shallow or indeed “genuine” forms of solidarity the toehold they require in order to manifest themselves and produce many of the socio-economic phenomena which continually puzzle economists.⁵

⁵ Of course, while canvassing in favor of solidarity as a distinct and fairly general analytical concept, we have said nothing about the actual inner mechanisms by which some agents may motivate themselves to conjure up and *direct* their solidarity (e.g., select a particular group *N* in connection with condition II in definition 4). Some thoughts on this can be found in Arnsperger and Varoufakis (1999).

References

- Akerlof, G. (1980), "A Theory of Social Custom, of Which Unemployment May be One Consequence", *Quarterly Journal of Economics* **95**, 749-775.
- Arnsperger, Ch., and Y. Varoufakis (1999), "Solidarity and Contemplation: Probing the Rational Foundations of Solidaristic Motivation", mimeo, Catholic University of Louvain and University of Sydney.
(<http://www.usyd.edu.au/econ/research/index.html>)
- Bacharach, M. (1999), "Interactive Team Reasoning: A Contribution to the Theory of Cooperation", mimeo, Oxford University.
- Becker, G. (1974), "A Theory of Social Interactions", *Journal of Political Economy* **82**, 1063-1093.
- Camarer, C. and H. Thaler (1995), "Anomalies: Ultimatum, Dictators and Manners", *Journal of Economic Perspectives* **9**, 209-219.
- Frankfurt, H. (1971), "Freedom of the Will and the Concept of Reason", *Journal of Philosophy* **68**, 5-20.
- Gauthier, D. (1986), *Morals by Agreement*, Oxford University Press.
- Geanakoplos, J., D. Pearce and E. Stachetti (1989), "Psychological Games and Sequential Rationality", *Games and Economic Behavior* **1**, 60-79.
- Heiding, H., and H. Moulin (1991), "The Solidarity Axiom in Parametric Surplus-Sharing Problems", *Journal of Mathematical Economics* **20**, 249-270.
- Hoffman, E., K. McCabe and V. Smith (1996), "Social Distance and Other-Regarding Behavior in Dictator Games", *American Economic Review* **86**, 653-660.
- Hollis, M. (1987), *The Cunning of Reason*, Cambridge University Press.
- Hollis, M. (1998), *Trust Within Reason*, Cambridge University Press.
- Hollis, M. and R. Sugden (1993), "Rationality in Action", *Mind* **102**, 1-35.
- Hurley, S. (1989), *Natural Reasons*, Oxford University Press.
- Nowak, A. S., and T. Razik (1994), "A Solidarity Value for n -Person Transferable Utility Games", *International Journal of Game Theory* **23**, 43-48.
- Olson, M. (1965), *The Logic of Collective Action*, Cambridge, Mass., Harvard University Press.
- Rabin, M. (1993), "Incorporating Fairness Into Economics and Game Theory", *American Economic Review* **83**, 1281-1302.
- Rawls, J. (1971), *A Theory of Justice*, Cambridge, Mass., Harvard University Press.
- Schelling, Th. (1960), *The Strategy of Conflict*, Cambridge, Mass., Harvard University Press.
- Selten, R., and A. Ockenfels (1998), "An Experimental Solidarity Game", *Journal of Economic Behavior and Organization* **34**, 517-539.

- Sprumont, Y. (1996), "Axiomatizing Ordinal Welfare Egalitarianism When Preferences May Vary", *Journal of Economic Theory* **68**, 77-100.
- Sugden, R. (1982), "On the Economics of Philanthropy", *Economic Journal* **92**, 341-350.
- Sugden, R. (1993), "Thinking as a Team: Towards an Explanation of Non-Selfish Behavior", *Social Philosophy and Policy* **10**, 69-89.
- Thomson, W. (1995), "Population Monotonic Allocation Rules", in W. Barnett *et alii* (eds), *Social Choice, Welfare and Ethics*, Cambridge University Press.
- Van Huyck, J., R. Battalio and R. Beil (1990), "Tacit Coordination Games, Strategic Uncertainty and Coordination Failures", *American Economic Review* **80**, 238-248.
- Varoufakis, Y. (1989), "Worker Solidarity and Strikes", *Australian Economic Papers* **28**, 76-92.
- Varoufakis, Y. (1990), "Solidarity in Conflict", in Y. Varoufakis and D. Young (eds), *Conflict in Economics*, Hemel Hempstead and New York: Harvester Wheatsheaf and St Martin's Press.
- Varoufakis, Y. (1991), *Rational Conflict*, Oxford, Blackwell.
- Varoufakis, Y. (1998), "On the Evolution of Cooperation, Discrimination, and Perceptions of Fairness: An Experimental Study of the Hawk-Dove Game", mimeo, University of Sydney.
(<http://www.usyd.edu.au/econ/research/index.html>)
- Weibull, J. (1995), *Evolutionary Game Theory*