CHRISTIAN ARNSPERGER and YANIS VAROUFAKIS

TOWARD A THEORY OF SOLIDARITY

ABSTRACT. Many types of 'other-regarding' acts and beliefs cannot be accounted for satisfactorily as instances of sophisticated selfishness, altruism, team-reasoning, Kantian duty, kin selection etc. This paper argues in favour of re-inventing the notion of solidarity as an analytical category capable of shedding important new light on hitherto under-explained aspects of human motivation. Unlike altruism and natural sympathy (which turn the interests of specific others into one's own), or team-reasoning (which applies exclusively to members of some team), or Kantian duty (which demands universalisable principles of action), the essence of solidarity lies in the hypothesis that people are capable of responding sympathetically to (or empathising with) a *condition* afflicting 'others', irrespectively of who those others are or whether one cares for them personally. And when that condition is a social artefact, we argue, solidarity turns radical.

1. INTRODUCTION

While the consensus regarding the state's responsibility for sustaining the unfortunate and empowering the weak remained intact, the notion of solidarity invoked images of Polish dissidents and striking British miners. However, for some time now the tide has been going out on many arguments in support of state-welfare systems. As it recedes, the few weedy posts it leaves behind seem to have inspired a variety of European politicians and institutions[1] to re-evoke solidarity, often as a means of counter-balancing the heightened emphasis on entrepreneurship and self-reliance. However, it is not at all clear what calls for 'greater solidarity' could possibly mean. Is it a euphemism for organised philanthropy? For social constructivism funded by means other than taxation?

More likely than not, politicians and activists make use of the term because of its emotive value, with minimal clarity regarding what solidarity actually means. This paper began with a query: Is solidarity a potentially useful analytical category? The result is an essay-in-retrieval on solidarity's *potential* meaning. It reflects the view that solidarity can be meaningfully distinguished from similar, and far better researched, other-regarding, dispositions; e.g., reciprocity, duty and altruism. Is solidarity just a sloppier term for what is already well-defined? Or does it open up a window to useful, fresh insights?

Erkenntnis **59:** 157–188, 2003.

We suggest one basic prerequisite for solidarity; namely, a generous disposition; a propensity to sacrifice something one values (even if it only amounts to lost peace of mind) on behalf of some targeted group of people (e.g., refugees) whose welfare one deems important. Such generosity is defined formally in Section 2 but nothing is said specifically on solidarity until Section 4. Section 3 demonstrates that even *minimal* generosity, as long as it is commonly anticipated, can change the complexion of several classic social interactions (e.g., Rousseau's stag-hunt game). Six popular explanations of generosity are then discussed (ranging from natural sympathy and altruism to fairness equilibria) before solidarity is defined (see Section 4) as an analytically distinct other-regarding disposition. Finally, Section 5 examines the special case of radical solidarity and links it to the evolution of arbitrary social power.

Section 4 presents solidarity in juxtaposition to competing other-regarding notions. To give a flavour of the argument, we believe that solidarity differs from altruism in that, whereas the latter is about treating the interests of other persons as one's own (or acting *as if* this were the case), solidarity is about identifying a *condition* which makes those who 'suffer' it worthy of one's concern *independently* of (a) who those unfortunates are, (b) whether or not one cares for them *personally*. Put differently, altruism is a response to others' needs, interests and character. Solidarity, in contrast, is defined here as a reaction to a *condition* which afflicts certain 'others' independently of their personal or social character. And when this unfortunate *condition* is a product of social evolution, a social artefact in other words, then generosity turns radical and solidarity becomes subversive (see Section 5).

To the extent that our hypothesis is sensible and solidarity is, indeed, a form of targeted empathy toward strangers whose personal character is not the issue, it is considerably more puzzling than other forms of 'other-regarding' propensities. Unlike the conundrum of altruism, which has been addressed exhaustively in various ways,[2] solidarity-with-selected-strangers is almost as bewildering as Nietzsche's (1956) paradox of trust.[3] Of course, the analysis in this paper does no more than to scratch the problem's surface. At best, it opens up the debate and sets the scene for analytical treatments of a concept which is slowly re-gaining prominence in European political culture.

2. GENEROSITY

All other-regarding deeds appear, at some level, as expressions of kindness or generosity. Thus it seems natural to start our search for solidarity

by postulating some minimal generosity that must characterise an act, or intention, before the latter is even considered as a possible expression of solidarity. Later we shall propose additional (sufficient) conditions which such acts must meet before specific cases of kindness can qualify as 'solidarity'; in juxtaposition to altruism, natural sympathy etc. So, we begin with a simple definition of *perceived* generosity: We *believe* we are being generous to others if we act in a manner costly to ourselves but beneficial to them.

Suppose person i (who belongs to group M) is facing some choice problem and define S_i, $a_i \in S_i$, and $u_i(.)$ as, respectively, i's set of feasible actions or strategies, i's chosen action, and i's intertemporal utility function. Suppose further that, in i's mind, there is a group of people, say N, who are affected by her choice. Then, the prerequisites of perceived generosity (see the previous paragraph) are in place if:

(a) i's choice $a_i \in S_i$ entails a sacrifice s_i for i, and
(b) i thinks that group N members somehow benefited by her sacrifice s_i.

For action a_i to involve some sacrifice it cannot, by definition, be optimal from i's own perspective. Thus i's optimal choice $a_i^* = \text{argmax}\{u_i(.)\}$ must be different from her actual choice a_i and, thus, her sacrifice can be expressed in utility terms as $s_i = u_i(a_i^*) - u_i(a_i) > 0$. As for prerequisite (b) above, suppose that $W_N^i(a_i)$ is an index of group N's welfare as perceived by i following i's choice of $a_i \in S_i$. Then $w(a_i) = W_N^i(a_i) - W_N^i(a_i^*) \geq 0$ is an index of how much i thinks that her sub-optimal choice a_i benefited group N.[4] Note that the usual aggregation problem does not apply here since the units of welfare utilised (w and $W_N^i(.)$) represent no more than i's *perceived* effect on the welfare of group N, as opposed to any real welfare effect.[5] In summary, the prerequisites for generosity, as stated above, take the form of simple inequalities: $s_i(a_i) > 0$ and $w(a_i) \geq 0$.

DEFINITION 1. Person i's λ-*generosity* to members of target group N is given by

$$\lambda_i(a_i) = \begin{cases} s_i(a_i) \times w(a_i) & \text{if } s_i(a_i) > 0 \text{ and } w(a_i) \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus person i (belonging to some group M; $i = 1, \ldots, M$) performs an act of λ-*generosity* toward members of group N ($j = 1, \ldots, N$) if she acts in a manner which benefits them at her own expense. This definition is important for three reasons. First, it distinguishes sharply between generosity and reciprocity in the sense that, while acts of reciprocal kindness are underpinned by an expectation of something in return, the genuinely 'generous' are generous for nothing.[6] Secondly, Definition 1 marks generosity

out from its 'darker side'; namely, from spiteful acts intended at hurting others, at one's own cost (see note 6). Thirdly, because λ-*generosity* is the foundation upon which our solidarity concept is erected in Section 4. Before we can delve deeper into these issues, we need to say more about the beliefs in the background of λ-*generous* acts.

While generosity can be random and lack reasons, to qualify as something more 'substantive' (e.g., as justice in action, or solidarity, or team-reasoning) actions must be grounded on specific reasons.[7] To begin with, for $i \in M$ and $j \in N$ we let $\lambda'_{ji} = E^j[\lambda_i]$ denote the expectation of j regarding i's λ-*generosity* to her group and $\lambda''_{ij} = E^i[\lambda'_{ji}]$ i's estimate of j's expectation of λ_i. For example. when $\lambda_i > 0$ and $\lambda''_{ij} > 0$, i intends to be generous to N-members *and* thinks that this is precisely what they expect her to do. The rationale here is that other-regarding acts are often driven by the power of others' expectations.[8] And when one's 'sacrifice' is directed at a whole group (N), then average expectations among that group, as well as one's assessment of what people in a similar position as hers might do (fellow M-members), play a crucial role in capturing our agent's situation. So, to complete a profile of other-regarding actions and beliefs, we define $\lambda_{M\sim i}$ as i's expectation of the average λ_i that others like her (i.e., also belonging to group M) will choose (or would have chosen under similar circumstances) and λ_{Nj} as the average λ_i i anticipates members of the target group N expect (on average) of M-members like her.

$$\Lambda_{M\sim i} = E^i\left[\frac{1}{M-1}\sum_{i\neq k=1}^{M-1}\lambda_k\right]; \quad \Lambda_{Nj} = \frac{1}{N}\sum_{j=1}^{N}\lambda''_{ij}.$$

So far we have looked at i's calculative second order predictions[9] viz. members of both her group and of those she wishes to benefit. Typically though there is another type of belief that plays a significant role in motivating agents: normative beliefs. To introduce such beliefs in i's deliberations, we define ξ_i as i's belief about the value of λ_i that she *ought* to choose; ξ''_{ij} as i's (predictive) belief about what $j (\in N)$ believes λ_i *ought* to be,[10] and $\Xi_{M\sim i}$ and Ξ_{Nj} as i's expectation of average opinion regarding the value of λ_i that she *ought* to choose amongst her fellow M-members as well as N-members respectively. That is,

$$\Xi_{M\sim i} = E^i\left[\frac{1}{M-1}\sum_{i\neq k=1}^{M-1}\xi''_{ik}\right]; \quad \Xi_{Nj} = E^i\left[\frac{1}{N}\sum_{j=1}^{N}\xi''_{ij}\right].$$

DEFINITION 2. Agent i's λ-*profile* is given by $\langle\lambda_i(a_i)|\{\Lambda_{M\sim i}, \lambda_{Nj}\}, \{\xi_i, \Xi_{M\sim i}, \Xi_{Nj}\}\rangle$.

In brief, a person's λ-*profile* is defined by (a) her λ-*generosity* toward members of group N, as conditioned on, (b) her calculative (the Λs), and (c) her normative beliefs (the Ξs). To illustrate, suppose $\lambda_i = \xi_i = \Xi_{Nj} > 0$ while $\Lambda_{Ni} = \Lambda_{M\sim i} = \Xi_{M\sim i} = 0$. In this case, i makes a sacrifice which she expects to have positive effects on the welfare of some target group N; she believes that she *ought* to be making such a sacrifice (and that N-members think so too); she also thinks that no one predicts that she would in fact prove so generous. Indeed, she is of the view that fellow M-members dismiss any notion that she is morally obliged to make sacrifices on the behalf of group N.

3. THE IMPACT AND SOURCES OF GENEROSITY

So far we have said nothing that pertains exclusively to solidarity. Indeed, the definition of an agent's λ-*profile* above may be helpful in depicting, and dissecting, all sorts of other-regarding behaviour, including altruism, or even love. While our particular hypothesis on what distinguishes solidarity from related concepts will have to wait until Section 4, it might be useful to emphasise one impression that the word 'solidarity' conjures up: solidarity, by nature, involves large numbers of people. In contrast, love and altruism seem to be better suited (though not exclusively so) to small groups.

The object of one's romantic love is, usually, a sole person. Altruism may be confined to a mother's feelings towards her offspring. Collusion usually involves no more than a handful of agents. By comparison, it seems harder to envision solidarity in a similar context; for it usually entails a generosity of spirit that extends to larger numbers, in which love and altruism have a tendency to dissolve. Coal miners caught up in some underground emergency are more likely to expect of their colleagues a degree of solidarity, or team-reasoning, rather than love, pure altruism or the type of reciprocal logic that motivates collective action against the employer.

Of course these are just preliminary thoughts which we shall return to after our attempt at a definition of solidarity in Section 4. Meanwhile, it is interesting to examine one common thread running through different types of other-regarding motivations which, like solidarity, are more relevant when more than two people are involved (e.g., reciprocity, norm-driven behaviour, team-reasoning etc.): The common thread in question is the thought that such other-regarding behaviour toward, as well as within, some target group N is inextricably linked to the group's shared identity. Moreover, a shared identity allows agents to coalesce to the *common* expectations of group N, or to *knowledge* regarding their generous dispos-

ition toward them. And when this happens, as we shall see below, some interesting results follow.

DEFINITION 3. Commonly known λ-*generosity* (CKG^*_λ) toward group N requires that, (a) each agent $i \in M$ chooses a sacrifice level s_i at least equal to $s^* > 0$ for the purpose of boosting the welfare of group N by $w^* > 0$; (b) knows that all other agents $j(\neq i) \in N$ know (a); (c) all agents $j(\neq i) \in N$ know (b); *ad infinitum*. By definition, $\Lambda_{Ni} = \Lambda_{M\sim i} = \lambda^* = s^* \times w^*$, while the normative expectations (ξ_i, Ξ_{Nj}, $\Xi_{M\sim i}$) could differ from λ^*. When, however, $\xi_i = \Xi_{Nj} = \Xi_{M\sim i} = \lambda^*$, we have a stronger case of CKG^*_λ in the sense that the agents' calculative beliefs are reinforced by (identical) normative ones.

A vivid illustration of the analytical value of *commonly known* λ-*generosity* (or CKG^*_λ) can be given in the context of a simple interaction in which an infinitesimal λ can solve a perennial problem in game theory, as long as it is commonly known. By *minimal generosity* we shall henceforth refer to a case of CKG^*_λ with $s^* = \epsilon$, where ϵ is vanishingly small but never zero. Consider the following one-shot game in which, for simplicity, sets N and M coincide: Suppose each person $i \in N(= M)$ must choose a real number a_i from the interval [1, 10]. The payoff function for each player is: $u_i(a_i) = A \times \min(a_i, a_j) - a_i \forall i, j(i \neq j) \in 1, \ldots, N$ where $A \neq 1$. Clearly this game is of the N-person co-ordination-problem type (also known as the Stag-Hunt game, see note 11) featuring an infinity of Pareto-ranked Nash equilibria within the continuum [1, 10]. Ultimately everyone is best off when each chooses $a_i = 10$ [in which case $u_i = 10(A - 1)\,\forall i$] and no player has an incentive to select a number below that chosen by others. Nevertheless the Nash best reply strategy is to choose the smallest number in [1, 10] that one predicts will be selected by anyone within the group [i.e., set $a_i = m$ where $m = E^i\{\min(a_j)\}, \forall i, j \in N$]. Thus even the slightest degree of pessimism (i.e., $m < 10$) suffices to lead players to an inefficient outcome. Indeed experimental work has shown that, often, the greater the experience of subjects with this game the lower their payoffs.[11]

Instrumental rationality, even when commonly known, cannot guarantee successful co-ordination in this game despite the absence of inbuilt incentives to 'cheat' or 'defect'. [12] However if, additionally, players act under common knowledge of *minimal generosity*, successful co-ordination on the Pareto-dominant Nash equilibrium is guaranteed. To see why, suppose that, in equilibrium, each player expects a Pareto-dominated Nash equilibrium with everyone in the group choosing $a_i = \alpha(<10)\ \forall i \in N$. *Minimal generosity* means that each player will be prepared to make a

tiny sacrifice $s_i = \epsilon$, an act of λ-*generosity*, in order to benefit the rest and will thus choose $a_i = \alpha + \epsilon$ (Nb. It is easy to show that if $s_i = s^* = \epsilon$ then $a_i = m + \epsilon$). At this stage, we have $\langle \lambda_i(a_i) = \lambda' > 0 \,|\, \{\Lambda_{M\sim i} = \Lambda_{Nj} = 0\}\rangle$ where $\lambda' = \epsilon^2(N-1)(A-1)$.[13] But then, courtesy of *minimal generosity*, everyone will anticipate i's new profile $\langle \lambda_i(a_i) = \lambda' > 0 \,|\, \{\Lambda_{M\sim i} = \Lambda_{Nj} = \lambda'\}\rangle$ and thus their estimates of α will be revised upwards. All of a sudden the ϵ-increase in a_i is no longer an act of sacrifice, or λ-*generosity*, since choice $\alpha+\epsilon$ is a Nash best reply strategy to the new expectations. In other words, the agent's profile is transformed again to $\langle \lambda_i(a_i + \epsilon) = 0 \,|\, \{\Lambda_{M\sim i} = \Lambda_{Nj} = 0\}\rangle$. Since by this stage of their deliberation no generosity is required, to be minimally generous is to choose $\alpha + 2\epsilon$; that is, each player's λ-*profile* is revised upwards to $\langle \lambda_i(a_i = a+2\epsilon) > 0 \,|\, \{\Lambda_{M\sim i} = \Lambda_{Nj} = 0\}\rangle$. And so on, until each player's λ-profile becomes $\langle \lambda_i(a_i = 10) = 0 \,|\, \{\Lambda_{M\sim i} = \Lambda_{Nj} = 0\}\rangle$. At that point all choose $a_i = 10$, the Pareto-dominant Nash equilibrium is achieved, and no *actual* generosity is necessary.[14] To sum up, once minimal generosity it taken for granted by all, co-ordination is achieved without any need for mutual sacrifices. This interesting result can in fact be generalised for a class of continuous, finite N-player co-ordination games.

PROPOSITION 1. In N-person interactions with continuous strategy/payoff spaces, multiple Pareto-ranked Nash equilibria, and risk-dominance of the Pareto inferior equilibria (over the Pareto-superior ones), the Pareto-dominant mutual-maximum (Nash) outcome will occur if players act under common knowledge of (a) *minimal generosity* (minimal CKG^*_λ) and (b) instrumental rationality (CKIR). Moreover, no generosity will be shown in equilibrium.

Proof. We consider games in which each player's strategy a_i is chosen from a continuous, closed and bounded set $S_i \in \mathbb{R}$ with a common upper bound $(\bar{a})$. Further, the players' pay-off functions u_i are also continuous mappings such that: (i) the game features multiple Pareto-ranked Nash equilibria; i.e., $n_i(a_j = a) > u_i(a_j = a - \epsilon)\ \forall i,\ j \in N$ and $\epsilon > 0$; and (ii) no player has a capacity to increase her pay-offs by choosing an a_i below the smallest choice in the group; that is, i's best reply to the expectation that the smallest choice will equal m [i.e., $m = E^i\inf(a_j),\ \forall i,\ j \in N$] is to select strategy $a_i = m$. By definition, when everyone selects $\bar{a}$ as their strategy, each collects the highest available pay-off; a mutual-maximum equilibrium: $u_i(a_j = \bar{a}) \geq u_i(a_j \leq \bar{a})\ \forall i,\ j \in N$. In this equilibrium, private and social optimisation is achieved and, by our earlier definition of λ-generosity, no agent gets a chance to put their generosity on display since each chooses the behaviour that serves their narrow self-interest. Under

the assumption of a continuous strategy/utility space, it is evident that no other symmetrical outcome (i.e., a strategy choice of $a_j < \bar{a}, \forall j \in N$) is consistent with both Nash equilibrium and *minimal generosity*. To see this, suppose that each player is contemplating strategy $a_j < \bar{a}$ and everyone knows this. Under *minimal generosity*, each expects everyone else to be ready to make a slight sacrifice on behalf of the rest; that is choose $a_j + \epsilon$ instead of a_j. Due to the continuity assumption, a new Nash equilibrium exists in pure strategies: players choose $a_j + \epsilon, \forall j \in N_i$, and thus anticipate a uniform rise in their pay-offs. Once this stage in the iterative process is reached, agents again optimise (and their sacrifice level returns to zero). A new iteration therefore starts as *minimal generosity* and, once more, motivates players to revise their strategy upwards to $a_j + 2\epsilon$. And so on until the iterative process reaches its upper barrier at the mutual-maximum equilibrium at which, as shown above, actual generosity is neither necessary nor possible.

The interesting feature of the above result is that *generosity*, even in tiny doses, succeeds (as long as it is commonly known) where hyper-rationality has hitherto failed: in procuring an all-round beneficial (that is, Pareto superior) equilibrium. As long as the players' payoffs are continuous functions defined in a continuous strategy space, even infinitesimal values of ϵ will gradually dispel pessimistic expectations and push players' strategies in the direction of the mutually most beneficial equilibrium. The above proposition is of course relevant for a fairly narrow class of social interactions: Continuous co-ordination games in which i's higher than average contribution (or sacrifice s_i) benefits the other player(s) (however infinitesimally). *Minimal generosity* suffices in such games because Jill has an opportunity to be minimally λ-*generous* to Jack in every Nash equilibrium (and vice versa). This is the hook that the algorithm requires to generate full co-ordination out of minimal solidarity.

By contrast, no such hook is available either in pure co-ordination problems or in antagonistic games (e.g., such as hawk-dove, prisoner's dilemma). In the former case (i.e., pure co-ordination), once they have homed in on some equilibrium (however Pareto inferior it might be), agents have no way of making the requisite minuscule sacrifice on behalf of fellow players. Similarly in the case of antagonistic games; once a conflict of interest emerges (e.g., when different equilibria are favoured by different people or players have clear incentives to 'defect', as in the prisoner's dilemma), *minimal generosity* fails to make a difference. In those richer contexts we shall need to examine the connection between a person's degree of λ-*generosity* and the underlying beliefs within her λ-profile. Nevertheless, it was still rather important to have shown (see

above) that there *does* exist a class of social interactions in which even minimal, commonly known, generosity can forge hearty bonds between atomistic individuals. The question now is: What motives underpin commonly anticipated generosity? In the remainder of this section we review a number of well-researched sources of such motives. In the next we argue that solidarity is quite distinct from these and deserves to be treated as a separate notion.

(a) *Team-reasoning*: According to Sugden (1993) and Bacharach (1999) individually rational persons sometimes manage to see themselves as members of a team whose common purpose bears significantly upon their private passions. When this happens, a general commitment to the team's objectives is taken for granted and various coordination difficulties disappear. Precisely the same point was made in the previous section; namely that several coordination failures are avoided once agents are embroiled in *minimal generosity* (or minimal CKG^*_λ). Thus *team-reasoning* and *minimal generosity* are analytically equivalent. They help resolve the same class of coordination problems while, at the same time, they fail in equal measure to foster cooperation at the slightest hint of conflicting interests between agents. For example, in interactions of the prisoner's dilemma type, *team-reasoning* dissolves in the wake of the centrifugal forces created by private agendas and *minimal generosity* is too brittle to overcome the destructive logic of free-riding. Something more is needed. Indeed in the context of a free-rider problem (or N-person prisoner's dilemma) that 'something' is *maximal generosity* (or maximal CKG^*_λ).

DEFINITION 4. *Maximal generosity* toward group N requires commonly known λ-generosity (CKG^*_λ) among members of group M with $\lambda^* = \lambda_i(a_m)$ for each $i \in M$ and $a_m = \arg\max_{a_i}(\lambda_i(\alpha_i))$.

Example. Consider a free-rider variant of the earlier N-person interaction. Each player selects a real number in the [1, 10] interval and receives payoffs:

$$u_i(a_i) = A \times \frac{1}{N} \sum_{j=1}^{N=M} a_j - a_i \; \forall i, j \in N(= M),$$

with $N > 1$ and $A \neq 0$.[15] In the previous game *minimal generosity* guided instrumentally rational agents safely to the mutually maximum outcome. In this free-rider version, however, the dominant strategy is to choose 1 regardless of what the others will do and, therefore, nothing less than a (commonly known) readiness to be maximally λ-*generous* (i.e., choose 10 rather than 1) will do the trick.[16]

PROPOSITION 2. In free-rider/prisoner dilemma games, the mutual-maximum (non-Nash) outcome will be selected if players act under common knowledge of *maximal generosity* (maximal CKG^*_λ) *given their beliefs regarding their opponents' choices*.

Proof. Consider the simple two-person prisoner's dilemma in which each player chooses between strategies 'defect' (d) and 'co-operate' (c) and faces the following utility preference ordering $u_i(d,c) > u_i(c,c) > u_i(d,d) > u_i(c,d)$ for $i = 1,2$; where $u_i(a,b)$ is i's utility from playing strategy a while the other player chooses b. By virtue of strict dominance, their optimal action a_i^* is to select strategy d independently of their expectations. To do otherwise (i.e., to select $a_i \not\equiv a_i^*$) requires *maximal* λ-*generosity*: If 1 expects 2 to choose her dominant strategy d, in choosing c 1 is selecting the maximum sacrifice s_1 possible $\{s_1 = u_1(c,d) - u_1(d,d)\}$ *and* the largest welfare benefit to her opponent $\{w = u_2(d,c) - u_2(d,d)\}$. If on the other hand 1 expects 2 also to be λ-*generous*, that is to play strategy d, in choosing c player 1 is selecting the sacrifice level $s_1 = u_1(c,c) - u_1(d,c)$ *and* estimates the welfare benefit to her opponent as $w = u_2(c,c) - u_2(c,d)$. In the special case where $u_i(d,c) - u_i(c,c) = u_i(d,d) - u_i(c,d)$, the degree of λ-*generosity* necessary to bring about a co-operative action is maximal and independent of the actor's beliefs regarding her opponent's intentions. When this equality does not hold, then a necessary and sufficient condition for co-operative moves is that players adopt maximal λ-*generosity* given their beliefs about the opponent's move. A similar result holds in N-player versions of the game. For instance, in the free-rider game above, any strategy choice $a_i = a > 1$ corresponds to a sacrifice level equal to $s_i(a) = (1-(A/N))(a-1)$ while the welfare impact of such λ-*generosity* to the remaining $(N-1)$ players equals $w = (A/N)(a-1)(N-1)$. Thus the precise level of λ-*generosity* by player i (whenever she strays from her dominant strategy $a_i = 1$) is given as $\lambda_i = s_i \times w = A(N-1)(N-A)((a-1)/N)^2$. In this case, due to the linearity of the pay-offs, it is clear that a player's λ-*generosity* is independent of her beliefs regarding how others will behave. Moreover, to reach the decision to play in a fully co-operative manner (that is, set $a_i = 10$), we require maximal λ-*generosity*.[17] When pay-off functions are non-linear, again we require maximal λ-*generosity* only this time the latter will vary with the players' beliefs about their opponents' behaviour.

Some authors have argued, controversially, (see, for example, Gauthier, 1985) that, in the context of free-rider games, any level of λ-*generosity* below its maximal value is an instrumentally irrational choice. Their point is that it would be profitable to develop a *disposition* toward *conditional co-operation*, which in our terms translates into arguing that there are good

instrumental reasons for cultivating in our hearts and souls an λ-*profile* which comprises maximal λ values as long as $\Lambda_{M\sim i}$ and Λ_{Nj} exceed some threshold. However this is an unconvincing argument because, at least in one-shot free-rider interactions, values of λ_i significantly greater than zero cannot be explained unless agents are motivated by something beyond an urge to increase their direct utility.[18] Below we examine well-known suggestions as to what that 'something' might be.

(b) *Hume's natural sympathy*, *Smith's moral sentiments and utilitarian altruism*: Moved by sympathy, the "chief cause" of moral practice according to David Hume, the agent may think of others' interests as her own (though in inverse proportion to the psychological distance between her and 'them').[19] Similarly with generosity occasioned by Adam Smith's moral sentiments.[20] Given sufficient sympathy or sentiments for members of group N, the value of λ chosen by a Humean/Smithian can be quite substantial. On the other hand, the fact that neither sympathy nor sentiments extend to all people and all groups is what creates the need for, and the possibility of, justice. To be just is to be generous to those for whom one harbours no 'natural sympathy' or 'moral sentiments'. Though not necessarily an end in itself, pleasure derives from acting justly toward others; something that can only imply that a sacrifice was made on their behalf at odds with one's narrow self (or family, or class) interest. "With regard to all ... benevolent and social affections", wrote Smith (1759) "it is agreeable to see the sense of duty employed rather to restrain than to enliven them, rather to hinder us from doing too much, than to prompt us to do what we ought. It gives us pleasure to see a father obliged to check his own fondness, a friend obliged to set bounds to his natural generosity, a person who has received a benefit, obliged to restrain the too sanguine gratitude of his own temper".

Utilitarians have a simple explanation of positive sacrifices $s_i > 0$ on the behalf of target groups. Having reduced all of the agent's passions (including her natural sympathy to others) to a single one (i.e., the maximisation of utility function u_i),[21] positive s_i values stem from an inner cost-benefit analysis. To be precise, an altruistic act a_i°, involving sacrifice level $s_i > 0$, is performed when $a_i^\circ = \arg\max_{a_i}\{u_i[s_i(a_i), w(a_i)]\}$; i.e., because this sacrifice leaves the agent at a higher point of her scale of ordinal preference. In this case, both the co-ordination and the free rider problems (examined above) recede in proportion to the valuation of others' welfare (i.e., to $\partial u_i/\partial w$).[22] However, we note that such sacrifices do not qualify automatically as cases of λ-*generosity* – recall Definition 1 and its insistence that generosity must involve a loss of net utility. However, utilitarians may get around this requirement by distinguishing between

direct and indirect utility; namely, between utility that does not take into account the psychological benefits from having acted selflessly and utility that does.

(c) *Kantian, rule-utilitarian and Rawlsian generosity*: A Kantian propensity to be generous is independent of any pleasure she might derive from it. Generosity, of this ilk, is a matter of doing one's 'duty'; and, in Kant's (1788) infamous words, "the majesty of duty has nothing to do with the enjoyment of life". In the same way that the Kantian is duty-bound not to break a promise (since she cannot will that everyone should break theirs), our Kantian refuses to set her λ_i's equal to zero [even when her net utility suffers as a result]. Thus, Kantians are, by construction, maximally solidaristic. In both games thus far examined, Kantians set $a_i = 10$ even though they are fully aware that smaller choices (i.e., contributions to the social group) are individually more lucrative. For Kant has defined rationality as a capacity to overcome the temptations of hypothetical reasoning and to stick to its categorical variant which enables, indeed forces, the rational person to recognise her duty to do what is right as opposed to what is expedient.[23]

Rule-utilitarians follow a similar, but quite distinct, logic. They ask: "What degree of generosity would maximise my utility were it to be chosen by all, including myself?" Again the unique answer in both relevant games is to select, as part of a rule or a disposition, the maximal sacrifice. Interestingly, both Kantians and rule-utilitarians end up with higher pay-offs (e.g., as a result of successful co-ordination and/or co-operation). But rather than being the *reason* for their generosity, this welfare improvement is merely a satisfying by-product. To recap, a Kantian's λ-*generosity* makes itself felt in the form of sacrifices performed in the *line of duty*; that is, independently of any cost-benefit calculation and unmoved by the expectations of others. It is in this sense that a Kantian's minimum[24] level of λ_i is always independent of the other arguments (ξ_i, Ξ_{Nj}, $\Xi_{M\sim i}$, Λ_{Ni}, $\Lambda_{M\sim i}$) in her λ-*profile*. Rule-utilitarians are less high-minded than Kantians (as utility is their ultimate guiding force) and more generous than straightforward utilitarians (since, unlike the latter, they are capable of generosity *as a rule*).

An analytically equivalent interpretation of *maximal generosity* can be attained by invoking Rawls' (1971) veil of ignorance. It is akin to a willingness, by an agent belonging to group M, to select an action after imagining that, *ex post*, one will end up either as still a member of group M or of another, less fortunate, group N (without knowing *ex ante* which of those M or N people one will turn into). If that 'blind' choice were to be made under the influence of infinite risk aversion, the resulting λ-

generosity would equal $\lambda_i \equiv \xi_i = \max_{\lambda_i \equiv \xi_i} \lfloor \min_{k \in M \cup N} u_k \rfloor$, irrespective of i's expectations.[25]

(d) *Conformity with others' predictive beliefs*: Olson (1965) makes the obvious point 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 λ_i in proportion to Λ_{Nj} and/or Λ_{Mi} because when, say, Λ_{Mi} is high she loses utility if seen to act selfishly (i.e., if seen to choose $\lambda = 0$). Akerlof (1980) produced a dynamic version of this story by modelling the relative weight of Λ_{Nj} in one's utility as an increasing function of Λ_{Mi}. In other words, as long as a minimum level of sacrifice (or λ-*generosity*) is anticipated, then a bandwagon effect begins to unfold and 'selfless' acts spread inexorably.[26] More recently, Brennan and Pettit (2000) extend these ideas in their study of the urge to cultivate esteem.

Geanakoplos, Staccheti and Pearce (1989) and Sugden (2000) delve deeper in suggesting a direct link between beliefs and preferences. They model an agent's preferences as a direct function of her second order beliefs; that is, an agent might prefer to act in solidarity with group N, *even if no one is to know*, as long as she thinks that this is what is expected of her. To see how this idea differs fundamentally from Olson (1965) and Becker (1974), consider two examples. First, in the models by Olson and Becker, if my actions are unobservable by others then there is nothing that would motivate me to be generous. Invisibility would remove the lure of prestige acquisition or the threat of losing face. However, in Geanakoplos et al (1989) and Sugden (2000) the mere fact that some people *expect* me to make a sacrifice makes me *want* to make that sacrifice (irrespectively of whether I am being monitored or not). Secondly, Geanakoplos et al. (1989) allow for the possibility that agents who act on these reasons might, nonetheless, regret the fact that others entertain 'great' expectations of them; a case of what we might term *reluctant* generosity.[27]

(e) *Conformity with others' normative beliefs*: This is a variant of (d) above with others' normative beliefs replacing their calculative ones in i's λ-*profile*. Once more, others gain a hold on one's utility, either directly or indirectly, as their moral beliefs influence the agent's preferences. Of course, the moment predictive beliefs are 'allowed' to contaminate preferences (Geanakoplos et al., 1989; Sugden, 2000), the distinction between positive and normative beliefs becomes really fine. If one's behaviour is influenced by an urge not to frustrate others' beliefs, and this is common knowledge, beliefs appear simultaneously as predictive and normative. Nevertheless, we think that the appearance of a fully collapsed distinction is deceptive. It is one thing to help a needy person because others *predict*

you will do so (and know that their predictions matter to you), and it is quite another to help because, otherwise, that they would think of you as morally defective.

(f) '*Biblical*' *generosity*: Imagine that person i plans to make sacrifices for group N because she thinks that, had *they* been in *her* shoes, they would be prepared to make similar sacrifices. Note a crucial difference between this and straightforward utilitarian reciprocity (which we have referred to previously as *enlightened selfishness*). In the latter case you help others because the expected benefits are significant (e.g., tit-for-tat cooperation in a repeated free-rider game). The same applies, though at the level of the unconscious, to or socio-biological reciprocity. Here, however, we are referring to a different motivation altogether: An agent i is prepared to act selflessly, and at a cost, *independently* of any *actual* benefits to be had from such action. The mere thought that group N members are well-disposed to her, that they would have helped her if they had swapped places, is sufficient reason to want to help them even if she thinks it impossible that such a reversal of fortune will occur. In this sense, i's λ-*generosity* will be positive regardless of whether she expects to benefit materially from it. It is intentions that count alone and, therefore, such beliefs can potentially lead to positive λ-*generosity* even in one-shot free-rider interactions.

However, this type of generosity has a nasty underbelly and it is for this reason that we use the term *biblical* to describe it. The ugly flipside transpires when we consider the possibility that M-group members fear that their N-group counterparts would be willing, if they could, to make positive sacrifices (s_i) in order to harm them. As a result, they are motivated also to make positive sacrifices to hurt them back. Indeed when both groups feel the same way about one another, we may end up in equilibrium with positive s_i values, negative welfare effects w_i, and no λ-*generosity* (since the latter is zero under these circumstances even if product $s_i\ w_i$ is non-zero).[28] A generalisation of this idea allows for the possibility that cohesion and mutual generosity *within* one group (M) might well be dependent either on mutual generosity or hostility with another (N).[29]

4. SOLIDARITY

The last section examined six other-regarding categories of generosity. In this section we argue that *solidarity* should be added to these as a separate analytical category of other-regarding motives and acts. To demonstrate why we think this, we re-visit Sugden's (1993) example of the *British Lifeboat Service*; an institution financed entirely through public donations. "Why do people contribute money to it?" asks Sugden. He points out that

the answer cannot lie in utilitarian altruism. For if donors are motivated by an interest in ensuring that the Service has sufficient funds to perform its lifesaving duties, they ought to think of each contributed pound as a perfect substitute for each pound contributed by someone else. Yet the econometric evidence contradicts this hypothesis.[30]

Selten and Ockenfels (1998) make a similar point. They report that, in an experimental setting, winners of a simple lottery proved quite willing to donate a portion of their winnings to the losers but, surprisingly, their donations turned out to be largely independent of how much the latter collected from other donors, or even of how the donations were to be divided amongst a number of recipients.[31] This result, just like the econometric evidence reported in Sugden (1993), amounts to a violation of utilitarian altruism's requirement that donors' valuations of recipients' utility from contributions be symmetrical vis-à-vis the contributors.[32]

In both examples, donors are channelling their empathy to a particular target group (e.g., the 'shipwrecked', the 'lottery losers'). The question is: On what basis is this group selected? The usual explanations turn on (a) personal characteristics and (b) universalisable principles. We are generous to persons from whom we expect something back (even if it is only their gratitude); who belong to the same team/group as we; for whom we care individually; or toward whom we have a sense of universalisable duty. With this paper we seek to highlight a different motivation: We may be generous to a class of persons (even when none of the above apply) simply *because we identify with their condition.* Our definition of *solidarity* draws on this capacity.

Before proceeding further with the definition of solidarity, it is important to note that solidarity may, of course, co-exist with reciprocity,[33] person-specific sympathy, team-reasoning and Kantian duty. The point, however, is that solidarity motivates generosity *independently* (that is, even in the absence) of these other-regarding motivations. The source of its power comes from nothing more than the fact that these are people unwittingly connected by some shared condition (e.g., ship-wrecked, HIV-infected) which fuels our solidarity toward *whoever* might be afflicted by it. Therefore, we envisage solidarity as a *condition-specific* disposition.

Given that solidarity (as defined here) does not rely on the expectation of reciprocal generosity, and in view of its impersonal (and condition-specific) nature, it is obvious that solidarity cannot be a species of 'enlightened selfishness' or utilitarian altruism.[34] The same applies to team-reasoning and Kantian duty, neither of which explain this aspect of human motivation. Our reasons for thinking this follow:

Team reasoning requires team spirit and, by definition, excludes all acts of solidarity by non-members. Though we lack the hard evidence on this, it seems likely that a large part of the funds received by the Life Boat Service come from non-sailors. Why would, for instance, a poor land-bound single mother give money to support a sea-rescue service? It seems far-fetched to suggest that her motivation is tantamount to natural sympathy or altruism toward rich round-the-world yachtsmen with more money than sense. Nor is it plausible that she fancies herself as part of their jet-setting 'team'. However, she may well contribute if she feels that the shipwrecked are *entitled* to *her* help in virtue of being shipwrecked and independently of who they are or how much others help them. Similarly with the winners in Selten and Ockenfels (1998). Given the experimental design, it is hard to imagine that subjects managed to develop in the laboratory the bonds which occasion team-reasoning. It is more credible to suggest that the winners donated money to losers, not because of some concern about how much money fellow players leave the laboratory with, nor because winners feel they belong to the same group as *losers*, but due to a feeling of solidarity with the losers as losers; a feeling which breeds an obligation to share with them part of one's winnings.

Why is this obligation not some form of Kantianism? Kantians are λ-*generous* because they *ought* to, even if they feel no empathy with the person afflicted by the condition that gives rise to their duty. They are capable of donating to the Life Boat Service (independently of their feelings toward sailors) because of a (universalisable) maxim about the (Kantian) rationality of helping the ship-wrecked. So far, this seems similar to our notion of solidarity-with-the-shipwrecked. However, a Kantian's universalisable logic means that she cannot pick and choose *between* maxims consistent with this logic. To give an example, if visiting cancer patients in hospital is a Kantian maxim, and so is donating to the Life Boat Service, the Kantian is duty-bound to do both. Thus, one characteristic of solidarity (as perceived here) that sets it apart from Kantian duty is the former's contingency; the possibility that one can be disposed to visiting cancer-patients but not to donating to the Life Boat Service, even if both are demanded by similarly universalisable maxims. This difference flows onto a second one.

When a Kantian visits a cancer patient, it is conceivable that she does so without love, pity, pleasure in helping a sick person, or from being in her company.[35] She visits because she must, in precisely the same manner that she is honest because of a maxim that prohibits lies. However, here lies a paradox. The patient is less likely to be helped by the Kantian's visit if she feels that it is performed coldly, out of duty, and without empathy. In Smith's (1759) words, a "... benefactor thinks himself but ill requited,

if the person upon whom he has bestowed his good offices, repays them merely from a cold sense of duty, and without any affection to his person". The Kantian knows this but is structurally unable to pretend to care personally (when she does not) because her visit is motivated by exactly the same 'force' that causes her to be honest, to respect red traffic lights and, of course, to visit cancer patients.

It might be argued that the same paradox emerges when someone visits our patient out of solidarity; motivated by empathy not toward her individually but due to her 'condition'. Not quite. Although solidarity is also impersonal in this sense, it differs crucially from Kantianism because the 'condition' responsible for it is not pre-determined by some steely, universalisable logic. The patient sees that her visitor is perfectly capable of disregarding all sorts of high-minded maxims (e.g., she lies when it suits, jumps red lights when impatient, ignores pleas for donations from the Life Boat Service). And yet, her visitor is moved by the plight of cancer sufferers like herself. This inconsistency that solidarity allows for (and Kantianism bans) makes for a more fruitful hospital visit.

To recap, team-reasoning confines generosity to team members; natural sympathy limits it to those for whom we feel *as persons*; and Kantian generosity recognises no special entitlements to one's generosity. A Humean's $\lambda_i > 0$ can only be attributed to i thinking of the sufferers' ends as a *means* to i's own; a Kantian's $\lambda_i > 0$ reflects i's eagerness to treat *all* others as ends-in-themselves. And while the former will only be generous to *persons* whose interests she can adopt as her own, the Kantian ends up performing her 'duty' to all but lacks in real compassion. By contrast, the notion of solidarity steers a middle course. It identifies a *condition* which makes those who 'suffer' it worthy of one's generosity *independently of who they are and what interests they have*. Some misfortune beyond their control defines a group of N persons as those entitled to one's λ-*generosity*; thereafter, the agent feels an emotionally charged urge to help them *out of solidarity with their condition*. And because the selection of this *condition* does not derive from some rationally determinate formula, solidarity packs the emotional element that Kantian duty is missing.

DEFINITION 5. A person's σ-*solidarity* toward some group N is given as

$$\sigma_i = \begin{cases} \lambda_i & \text{iff conditions (I) to (IV) apply} \\ 0 & \text{otherwise} \end{cases}$$

(I) *Personality-invariance*: i selects target group N *independently* of any personal characteristics of its members.

(II) *Condition-specificity*: Target group N is identified on the sole basis of an *adverse condition* which is *shared* by N's members. This *condition* is selected by an unspecified, non-universalisable method.

(III) *Belief-Irrelevance*: λ_i is independent of beliefs (Λ_{Nj}, $\Lambda_{M\sim i}$, $\Xi_{M\sim i}$, Ξ_{Nj}).

(IV) *Non-instrumentality*: Agent i's choice of the set of persons N is irreducible to the maximisation of expected net gains from the future behaviour of others (N-members and non-N-members).

Condition (I) differentiates solidarity from utilitarian altruism, personal sympathy etc. by ruling out personal motives and interests as a possible source. Condition (II) identifies solidarity exclusively with generosity directed at victims of misfortune, rather than of serendipity.[36] It also allows for a narrow and highly subjective focus of one's solidarity (by the virtue of the non-universalisability of the selection criteria) which is consistent with the often puzzling observation that a sighted person, who has no blind friends or relatives, may be prepared to go to incredible lengths to help with the education of blind children while remaining distant from similar efforts with deaf children. Condition (III) reflects the thought that solidarity cannot be motivated by an urge to impress others, or conform to their expectations (calculative or normative). Indeed it requires an autonomous moral judgment that some group N is somehow entitled to one's generosity, even if no well-recognised principle of justice so prescribes.[37] Condition (IV) is technically redundant (since a positive λ always comes at a personal cost – see prelude to Definition 1) but is included here in order firmly to remind us that we exclude from the realm of solidaristic acts those which, in the final analysis, are no more than shrewd self-interested investments.

So far we have established that, courtesy of our four conditions above, solidarity has been decisively distinguished from the previous section's other-regarding categories (b), (d), (e) and (f). The same conditions disqualify explanations (a) and (c) (team-reasoning and Kantian duty).[38] Conditions (I) and (II) ensure that σ*-solidarity* remains irreducible to team-reasoning since the λ*-generosity* underpinning it is not due to i belonging to target group N. Condition (II) keeps σ*-solidarity* analytically separate from some variant of Kantianism by introducing contingency into the selection of the 'condition' that motivates it. Taken at once, these conditions forge a notion of *solidarity* which can be juxtaposed usefully against the related ideas regarding fairness and justice. Such a juxtaposition, however, falls outside the scope of this paper.[39]

The urgent question that needs to be addressed next derives from Condition (II). If not on a basis of a universalisable principle, how does one select the condition that motivates her solidarity? While different 'conditions' tussle for our 'targeted empathy' (e.g., 'shipwrecked', 'loser in a lottery', 'redundant worker', 'refugee', 'victim of torture' etc.), only a small number, if any, succeed in eliciting σ-*solidarity*. This eclecticism lends emotional and moral weight to the ensuing acts of λ-*generosity* (e.g., makes hospital-visiting worthwhile) but also calls for an explanation. Why are some moved by the plight of the deaf, others by the plight of the blind, while many more remain unmoved by either? This paper offers no definitive answer. [Perhaps there can be no such answer if Condition (II) is to be met (i.e., the selection process is not unique and thus non-universalisable).] What it does claim, however, is (a) that σ-*solidarity* is probably as rare a phenomenon as it is socially important, and (b) that the reasons for selecting *the* condition(s) on which our solidarity trades may be either internal or external to our preferences.

Beginning with (a), there is little doubt that Conditions (I) to (IV) will remain dissatisfied more often than not. Most acts of generosity violate *personality-invariance* (in that they are directed to kin or friend); are *belief-contingent* (i.e., people are motivated to perform them because they are expected to); and verge on the *instrumental* (e.g., sacrifices are seldom independent of the hope that it will be reciprocated). However, just as dishonest acts trade on the fact that not everyone is dishonest, generosity that is not motivated by solidarity finds fertile ground on which to grow only in social settings where σ-*solidarity* has not been eradicated completely.[40] 'Other-regarding' deeds, which deep down are self-serving, must always remain parasitic on something resembling either our σ-*solidarity* or Kantian duty. Indeed, if perfectly egotistical acts can masquerade as other-regarding, selfless, solidaristic etc., this is so only because σ-*solidarity* not only makes sense but is also possible (and perhaps easier to relate to than Kantian high-mindedness).

Turning to (b), i's choice of some 'misfortune' or 'adverse condition' with which to empathise can be motivated by two types of explanation. An internalist explanation is fundamentally Humean in that it places the burden of explanation on the evolving passions and the feedback effects between the latter and the corresponding social conventions (or 'equilibria') that they spawn. Of course, there are a variety of explanations consistent with this. For instance, a neo-Humean might argue that a rich tapestry of solidarity is woven gradually over time (e.g., some people develop solidaristic feelings toward the homeless, others toward the refugees etc.); its genesis resembling the spontaneous emergence of conventions in

indeterminate social interactions while its survival depends on how successfully it regulates social life. In effect, neo-Humean solidarity (just like all other conventionally evolved patterns) adds to the evolutionary fitness of the community within which it sprung and, in a never-ending circle, is strengthened by it.[41]

Of course internalist accounts are not all neo-Humean. For instance, consider the following two-stage, rule-utilitarian account of i's σ-*solidarity* toward members of group N: In the first stage i selects the condition which determines set N (e.g., those who are 'shipwrecked', 'HIV carriers' etc.) on the basis of some principle external to both her preferences and to any social expectations. In the second stage, i chooses $\lambda_i = \arg\max_{\lambda_i}(U'[u_i(\lambda_i), W_N(\lambda_i)])$. 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 should be allowed to pass.

By contrast, those arguing in favour of fully external reasons for action (Hollis, 1987,1998) might insist that genuine solidarity requires a moral psychology which enables i to distance herself completely from her own preferences and passions; to show her solidarity to N-members for reasons pertaining to them, rather than reasons appealing to some desire or urge in her own bosom. Most economists would dismiss this idea and would associate non-optimising choices with bounded rationality. This is due to their insistence that reasonableness reduces to instrumental rationality or (in the term coined by Hollis (1998)) to philosophical egoism. However, there is no reason why this identification should be taken for granted. Unlike *homo economicus*, reasonable people can pass judgement on their own passions or desires and one way in which they rebel against the tyranny of preference is to do what is 'right' by some group of persons who are 'entitled' to their generosity. To the extent that this 'rebellion' is expressively (as opposed to instrumentally) rational,[42] and indeed finds expression in solidarity with sufferers of some misfortune, human motivation is underexplained unless solidarity is acknowledged as an important and distinct aspect of the human experience.

5. RADICAL SOLIDARITY: EMPATHISING WITH THE VICTIMS OF SOCIAL POWER

In the previous sections solidarity was defined as empathy with persons afflicted by some shared misfortune (e.g., cancer victims or shipwrecked sailors). When the latter is a social artefact, as opposed to an accident of nature, solidarity turns radical. The 19th century anti-slavery movement, for instance, was an expression of radical political solidarity with the victims of humanity's darkest artifice. It is a general tendency of human societies in all places and at all times to generate social power structures which place whole groups of people, quite arbitrarily, into 'unfortunate' roles and situations. Spontaneously, and through no fault of their own, they become victims of an evolved social force which expels them to the periphery of social life. A disposition toward making sacrifices on their behalf will be defined below as *radical solidarity*. First, however, we need to define arbitrary, evolved, social power and the hierarchies which it fashions.

DEFINITION 6. Suppose that the distribution of resources and social roles within a community $\mathcal{C}$ is determined by a series of interactions between its members. Suppose further that $\mathcal{C}$ is subdivided in at least two groups *arbitrarily*; that is, according to criteria irreducible to differences in their personal talents, application, or 'worth'. Members of group $K \subset \mathcal{C}$ are said to exercise two types of power over members of group $N \subset \mathcal{C}$:[43]

(a) *structural social power* if the structure of the interactions is consistently biased in their favour (and, therefore, so are their outcomes),[44]

(b) *conventional social power* if the outcomes of interactions between agents $i \in K$ and $j \in N$ conform to some discriminatory *evolutionary equilibrium* even though the interactions are structurally symmetrical.[45]

DEFINITION 7. *Radical* or *ρ-solidarity* is defined as σ-solidarity (see Definition 5) directed *consciously* to those who live under the structural or conventional social power of others. More precisely, i's $[\forall i \in M \subset \mathcal{C}]$ radical solidarity ρ_i equals σ_i *if and only if* it is directed to some group N for the reason that the latter's members are subjected to group K's social power. Otherwise, $\rho_i = 0$ (even if $\rho_i > 0$).

Strategies	Left	Middle	Right
Up	1,3	0,0	0,2
Middle	0,0	2,2	0,0
Down	3,0	3,0	**3,1**

Game 1: 1-2-3
Pure strategy equilibria in **bold**

	h	d	c
h	-2,-2	**2,0**	4,-1
d	**0,2**	1,1	0,0
c	-1,4	0,0	3,3

Game 2: Hawk-Dove-Cooperate
Pure strategy equilibria in **bold**

As an example, consider two games being played repeatedly among different identical opponents drawn from a large community $\mathcal{C}$. Game *1-2-3* has a unique equilibrium (Nash and evolutionary) which awards payoffs 3 and 1 to the players selecting among the rows and among the columns respectively. A case of *structural social power* emerges if some social process systematically selects K-players to choose among the rows in meetings with N-players. By contrast, *hawk-dove-cooperate* (HDC hereafter) is symmetrical and features two equilibria in pure strategies [(2,0) and (0,2)] and one in mixed strategies (play h with probability 1/3 and c with zero probability). Because of its symmetry, this game leaves room only for the *conventional* type of *social power*.

Although symmetrical in terms of its payoff structure, HDC spawns asymmetrical and highly discriminatory evolutionary equilibria *even if players are identical* in every respect other than their group membership.[46] As long as group membership is observable, it can be used as a behaviour-conditioning device whenever player $I \in K$ interacts with $j \in N$. To see why, consider the first few rounds during which differences in behaviour between the two groups can only be due to randomness. Once one of the two groups is *observed* to have selected h with higher probability (for reasons similar to why three tosses of a fair coin may yield three tails), a bandwagon effect begins to roll intensifying the originally random inter-group differences in aggression.[47] When the evolutionary equilibrium is reached, one of the two groups (say K) dominates the other (say N) in that its members play h consistently against members of the other group who acquiesce (i.e., respond with d). Thus the latter are, by Definition 6, subject to the *conventional social power* of the former.

In *1-2-3* the process manufacturing the subservience of N-members works through the assignment of row-column social roles; an assignment which has not been explained here. One possible explanation of its origins can be based on a fully endogenous analysis in the context of a conflictual interaction such as *hawk-dove*.[48] In such games, as discussed above, some groups gain the upper hand for reasons that have nothing to do with their personal qualities (see notes 46, 51 and 52). Indeterminacy conspires with asymmetry in order to spawn some non-rational social hierarchy. Once *conventional social power* has been established, the dis-

criminatory conventions which it produces spread from one interaction to another, perhaps by the force of analogy, and determine the allocation of social roles in a manner which favours the already dominant groups. For example, the group that dominates in *hawk-dove* ends up with the row-role in subsequent plays of game *1-2-3*! Thus *conventional social* power may oversee and spontaneously lead to the creation of *structural social power*. In the process, whole groups of people are arbitrarily assigned the lesser roles and, through no 'fault' of their own, are subjected to the misfortunes reserved for them by unconscious, supra-intentional social design.

Juxtaposed against such evolutionary accounts, a number of interesting issues flow from our definition of *ρ-solidarity* as solidarity with the victims of discriminatory social design. For example, instinctively, the notion of solidarity-with-an-oppressor seems strained. Interestingly, and encouragingly, this 'strain' shows up in our taxonomy of solidarity above. Consider a case in which a group of socially powerful agents is threatened with loss of power and thus privilege; e.g., the dissolution of a Mafia-type organisation or white rule in South Africa. To the extent that their loss of privilege, wealth and status can be thought of as a 'misfortune' afflicting them as a group, there is nothing in our original definition of a *solidarity profile* (see Definition 2) to rule out solidarity as targeted empathy toward (or within) such groups. Indeed, it is even possible that such sentiments qualify as *σ-solidarity*, provided the conditions of *personality-invariance*, *condition-specificity*, *belief-irrelevance* and *non-instrumentality* hold (see Definition 5).[49] The anomaly is however revealed when we submit these cases to the test of radical solidarity; a test which they cannot but fail since radical solidarity is directed solely toward groups who fall on the short side of evolved, arbitrary social power.

So it seems that our last refinement of the solidarity definition (*ρ-solidarity*) drives a wedge between the sentiments underpinning the collusion between holders of arbitrary social power and those shoring up acts of sacrifice (on behalf) of its victims. Things get messier however in the presence of interpenetrating patterns of discrimination, where the same group may be, at once, the victims in one type of interaction and the perpetrators in another.[50] And if discriminatory patterns have a tendency to survive by dividing and multiplying,[51] then evidence of *ρ-solidarity* and coercive collusion, whose purpose is to maintain some form of discrimination, may be found within most groups.

A related issue concerns the connection between philanthropy and solidarity. Whether, and to what extent, the philanthropist's motives can be deemed solidaristic depends both on her reasons and cognition of the beneficiary's situation. In our account, the identification of a group as

worthy of her concern and sacrifice is the first prerequisite. To qualify for *σ-solidarity*, her motives must be untainted by a concern for what others expect of her, or what there is 'in it' for her (a 'condition' also imposed by Christian and other religions). And to meet the criteria of *ρ-solidarity* she must be conscious of the specific social design which manufactures and arbitrarily assigns misfortune to undeserved victims. By these criteria, few Victorian philanthropists' acts and motives would qualify as *solidarity*[52] and even fewer as *radical solidarity*.[53]

Perhaps the natural limit of *radical solidarity* is a capacity to focus one's endeavours on undoing the root-causes of others' systematic disadvantage and misfortune, even if this means undoing also the sources of one's own privileges. Such radical solidarity transcends mere palliative efforts; it threatens to dismantle whole networks of privilege and destitution but carries enormous risks for both 'donor' and 'recipient' as it combines opportunities for progress with the risk of gigantic folly characteristic of all radical change.

6. CONCLUSION

Hurley (1989) castigates *homo economicus* for lacking the *nous* effectively to engage in the bewildering enterprise of acting in a manner *organically* consistent with the objectives of the team to which she belongs. This paper takes Hurley's theme further by focussing on organic connections of the self with groups of 'others' to which one does *not* belong. A rational person may expect nothing *of* them, may care not one iota *for* them individually, may feel she has no duty *to* them in particular, that she neither belongs to their 'team' nor wants to. *Homo economicus*, who only acts when there is something 'in it' for her, would not lift a finger on their behalf under the circumstances. However typical of men and women this model might be, there are exceptions whose importance relates inversely with their frequency. *Some* intelligent people, *some* of the time, are capable of selfless sacrifices, moved neither by expected gain nor altruism nor duty, but by a fierce repugnance for the suffering caused by some accident of nature or of social evolution.[54]

Of course empirical observation cannot help us distinguish genuine solidarity from impostors, just like it cannot settle disputes between, say, Humeans and Kantians. Yet this does not lessen the importance of exploring philosophically the notion of authentic solidarity. For its very possibility, however faint it might be, provides the toehold necessary for shallower forms of solidarity to proliferate. Tiny as these ripples of genuine solidarity may be, they often turn into torrents of targeted empathy through

imitation, social influence, even sheer hypocrisy. When they do, the social scenery is transformed and the cement of society is inserted between the bricks of individualist endeavours.

Rational choice theory is a powerful tool for explaining behaviour in response to preferences inhabiting the well-defined space within the walls separating one self from an 'other'. Solidarity, on the other hand, refers to a phenomenon made possible because these walls are more porous than rational choice theory would permit; it alludes to a series of human interactions unfolding in the space *between these walls*, in a kind of no man's land where the plight of others inspires us to experiment with violations of our current 'preferences', rationally toy with alternatives to the prevailing constraints of 'rationality', throw away the masks of self-sufficiency, reach out for one another, re-discover something 'real' and authentic about our nature and, at rare moments, believe that there is more to us than some weighted sum of desires. Those of a romantic disposition may even conclude that solidarity-with-others is a prerequisite for throwing out a bridge over to our 'better' self.

ACKNOWLEDGMENT

We wish to thank Nicholas Theocarakis and seminar participants at the Chaire Hoover, Université Catholique de Louvain and the Economics Department, University of Sydney for helpful comments. In particular, we are indebted to an anonymous referee for her/his many valuable queries, suggestions as well as criticisms.

NOTES

[1] There is hardly a European politician who, in the aftermath of monetary union, has not called for the blending of stringent monetary policies with a new commitment to solidarity with weaker members of society. Such calls have been reinforced from an array of institutions ranging from the churches and social activist networks to the Confederation of European Industries. For a recent example see Rouille D'Orfeuil (2002).

[2] Evolutionary biologists tell us that altruism is not a puzzle, in the sense that there is plenty of evidence from the animal world supporting the idea that altruistic behaviour does indeed improve a species' fitness (Dawkins, 1976; Midgley, 1994). Economists favour models of enlightened selfishness in which bargain-hunting agents, though incapable of resisting the lure of a marginally higher payoff, are nevertheless led to the conclusion that it pays to be 'good'. Whilst this is the rational choice theorist's favourite explanation of humanity's mysterious, other-regarding side, it is by no means the only one. Some (Sugden, 1986) still rely on Hume's (1739, 1962) distinction between selfish and self-interested actions, and the notion of conventionally reinforced natural sympathy that is founded on this distinction. Others turn to bounded rationality and evolved social reciprocity, as opposed

to instrumental or economic reciprocity; that is, to norms of cooperative or seemingly altruistic behaviour which jump from game to game through analogy and habit (Hoffman et al., 1996). Non-utilitarian thinkers, meanwhile, have been focusing on explanations turning on kin selection, rationally deduced obligations to others (or duties, e.g., Kant, 1788, 1949) and ideas about justice and fairness (Rawls, 1971).

[3] "To breed an animal capable of promising – isn't that just the paradoxical task which Nature has set herself with mankind, the peculiar problem of mankind?" Nietzsche (1887, 1957).

[4] Under the assumption of cardinal utilities, a particular case would be a Benthamite aggregation such that set N comprises the *complete* human population (and W_N is the average cardinal utility). Another particular case would be for set N to contain a single person: the one with the lowest utility (a type of welfarist-Rawlsian solidarity).

[5] Liberals should beware the assumption that an act is 'generous' when the actor deems that she has benefited others through her own sacrifice. Sen (1970) issues an early warning. In our context it takes the form of a query: What if i feels that group N members need to be 'saved' from themselves by, for example, being burnt at the stake? Is burning them an act of kindness? A simple retort is that, naturally, it is anything but an act of kindness. But, on the other hand, if i genuinely thinks that she is benefiting them, we should accept that she is performing an act which she perceives, misguidedly of course, as kind.

[6] Act a_i is generous ($\lambda > 0$) when both $s(a_i) > 0$ and $w(a_i) > 0$. When $s(a_i) < 0$ and $w(a_i) < 0$, we have an act that causes hurt at no expense to the agent and, therefore, $\lambda = 0$ even though $s(a_i) \times w(a_i) > 0$. Spiteful acts set $s(a_i) \times w(a_i) < 0$ as they imply $s(a_i) > 0$ and $w(a_i) < 0$. Product $s(a_i) \times w(a_i)$ is also negative in cases of reciprocal kindness, i.e., when agent i benefits others $[w(a_i) > 0]$ but does so expecting something back in return [i.e., $s(a_i) < 0$]. In both these cases (spite and reciprocity) Definition 1 sets λ-*generosity* equal to zero. Finally, note that the intersection of groups N and M may well be non-empty.

[7] For example, Rabin (1993) argues convincingly that the same action can be deemed fair or unfair depending on the agent's first and second order beliefs. Chapman (1998) takes this idea further by examining how rational behaviour might be affected if agents had to give well argued reasons for their actions; as they must in a court of law.

[8] Geanakoplos et al. (1989), Rabin (1993) and Sugden (2000) model instrumentally rational actions which transcend the Humean divide which keeps beliefs separate from motives (e.g., utility). The common thread running through these three articles is that a person's valuation of a certain outcome depends, among other things, on her second order beliefs (that is, on what she thinks her opponents/friends expect her to do).

[9] Calculative or positive beliefs are mere predictions. We use these epithets in order to distinguish them from normative beliefs which pertain to beliefs regarding what *ought* to happen; as opposed to what *might* happen.

[10] Note that this second order belief is not a truly normative one. A truly normative second order belief would correspond to what i thinks that j ought to think that i will do.

[11] This game is identical in structure to Rousseau's stag-hunt game. Rousseau's original narrative had a group of hunters choosing between combining their efforts to catch a stag (the grand prize capable of feeding the group for days) or, alternatively, hunting skinny hares individually. The stag would escape if even a single hunter broke the 'chain' and sought to capture hares (i.e., everyone's payoffs is determined by the effort expended by the least committed members). Rousseau's point was that were the hunters to trust one another to pursue the stag diligently, they would all do so. However, pessimism about

the group's solidarity would force them all to the suboptimal pursuit of hares. In recent times, experimental work has shown co-ordination to converge on inefficient outcomes in this type of game. It seems that Pareto-dominated Nash equilibria are selected because risk-dominance overpowers Pareto-dominance. See van Huyck et al. (1990).

[12] Note that, unlike the prisoner's dilemma or the free riding game, there are no built in incentives in this game to cheat/defect. If one expects everyone else to contribute maximally one would follow suit.

[13] Suppose the expected minimum choice equals m, but player i is prepared to choose $a_i = m+x$. The sacrifice involved equals x since sacrifice level $s_i = (A-1)m-[(A-1)m-x]$. When commonly anticipated, this sacrifice will lead all to make it. In this sense, i's sacrifice x has increased the welfare of the rest of the group to the tune of $w = (N-1)(A-1)x$. Thus, i's λ-*generosity* equals $\lambda_i = s_i \times w = x^2(N-1)(A-1)$. Under minimal generosity, the sacrifice is minimal, i.e., equals ϵ, and therefore $\lambda_i = s_i \times w = \epsilon^2(N-1)(A-1)$; a value of lower order viz. the degree of sacrifice involved. On the other hand, for λ-*generosity* to be of ϵ-order, $\lambda_i = \epsilon = s_i \times w = x^2(N-1)(A-1)$, in which case the relevant sacrifice level is $x = \sqrt{\epsilon/(N-1)(A-1)} + o(x)$.

[14] This being a one-shot game, the 'algorithm' described here unfolds in logical, rather than in historical, time. It simply captures the train of thinking that leads players to the unique equilibrium (in a manner analytically identical to the process of iterative dominance or, as it is sometimes known, the successive elimination of dominated strategies).

[15] Note that the difference between this variant of the game and the original is that here the average choice of number in the group has replaced the minimum choice in each player's utility function. Obviously this changes the character of the game from that of a coordination/stag hunt type to a N-person free-rider problem since, by choosing a number smaller than the average choice, your payoff rises as long as $N > A$. To see this, note that the derivative of player i's pay-off function u_i s.t. a_i is negative as long as $N > A$. And since there can be no fewer than 1 player, $N > A > 1$ is the condition under which each of the N players has a dominant strategy: "Set $a_i = 1$!" In short, it pays to undercut the 'contribution' of the average player in the group.

[16] Note however that the amount of generosity required to sustain the co-operative outcome varies. For if they all expect maximal generosity of each other, then the actual sacrifice of each $i \in M(s_i)$, and the welfare benefit of others (w) following this sacrifice, is smaller than it would have been if co-operation was not envisaged.

[17] It is easy to see that co-operative behaviour requires $a = 10$, a value that maximises λ. Taking the limit as N tends to infinity, we note that, in games involving many players, a co-operative outcome requires mutual λ-generosity equal to $81A$.

[18] For a summary of why instrumentally rational agents cannot be reasonably expected to choose a cooperative disposition in free rider (or prisoner's dilemma) interactions, see Hargreaves-Heap and Varoufakis (1995), Chapter 5.

[19] For a modern version, complete with empirical evidence, see Andreoni (1990).

[20] "If he was to lose his little finger tomorrow, he would not sleep tonight. But provided he never saw them, he will snore with the most profound security over the ruin of a hundred million of his brethren, and the destruction of that immense multitude seems plainly an object less interesting to him, than this paltry misfortune of his own" Smith (1759).

[21] Note that the passage from Humean to *homo economicus* is not as straightforward as some seem to think. Indeed 'sanitising' the passions so as to turn them into preferences (cardinal or ordinal) is philosophically problematic. See, for instance, Sugden and Hollis (1993). For a different perspective on the same issue, see Margolis (1981).

[22] Nevertheless, the paradox of 'rational saints' remains. If each player is motivated by a selfless urge to satisfy the preferences of others, then in the context of a prisoner's dilemma agents may still get caught up in a mutual-minimum since each will be failing to make a sufficiently satisfying sacrifice on others' behalf.

[23] An anonymous referee made the point that "... Kant meant us to ask ourselves whether our action is possible as such if all selected that action. Hence the categorical interdiction of lying and cheating, as one literally cannot cheat if nobody honours agreements ..." This is not the place to enter into hermeneutical debates around what Kant really meant. However, it is fascinating to note that, if we were to accept the referee's interpretation, Proposition 1 would be threatened. The latter shows that *minimal generosity* leads to an equilibrium in which generosity is rendered impossible. In a sense, Kant would be censoring not only lies but also contributions to the Public Good.

[24] We say 'minimum' because there is nothing stopping a Kantian from boosting her generosity beyond the level determined by her 'duty' in cases in which she does feel sympathy for the target group or person.

[25] Rawls' (1971) argument is that rational agents will exercise infinite risk aversion behind the veil and will thus choose the best outcome from the perspective of the person who will end up being worst off. Thus if agents are forced to go behind the veil, and choose while there, their choices (which amount to a maximal λ) are deemed, by Rawls, to be merely rational. However, in view of the fact that no one is ever forced to go behind the veil, a willingness to decide what to do on the basis of what one would have done *had one found oneself behind the veil*, is a willingness tantamount to a *generous* predisposition.

[26] Akerlof (1980) utilises this idea in order to model the decision of unemployed workers not to undercut the wages of their employed colleagues and Varoufakis (1989, 1990) tells a story about wage and employment determination when a trades union's power stems from worker solidarity during (actual or threatened) strikes.

[27] Geanakoplos et al. (1989) examine a situation in which person A must choose between acting courageously or cowardly (nb. this is not really a game in the sense that there is only one player: A). Her utility from these two outcomes hinges crucially on what others' expect of her. So, if A believes that others expect her to act courageously, she will *want* to do so. If not, she will prefer to act like a coward. There is nothing to suggest that in the former case A's utility will not be lower than in the latter.

[28] Rabin (1993) labels a similar situation an *un-fairness equilibrium*.

[29] For example, suppose that for $i \in N$ the utility function is given by: $U^i = u_i(\pi_i, \lambda_i) + \gamma_i[\lambda_i \times \Lambda_{Mj}]$ where π_i, is i's material payoff and $\gamma_i > 0$ is some constant which reflects i's relative valuation of the means by which certain payoffs are produced. Similarly, let $U^j = u_j(\pi_i, \lambda_i) + \gamma_j[\lambda_j \times \Lambda_{Ni}]$ be the utility payoffs to $j \in M$. Such a maximand instructs i and j (as long as the γ's are large enough) to set λ_i, $\lambda_j > 0$ if they anticipate $\Lambda_{Mj} > 0$ and $\Lambda_{Ni} > 0$ respectively. However it also urges them to set their $s > 0$ in order to cause $w < 0$ (i.e., to make positive sacrifices in a bid to hurt the other group) if they expect a similar disposition from members of the other groups.

[30] See also Sugden (1982).

[31] Three subjects A, B and C participated in a lottery which would award each DM10 with probability 2/3. Subjects where asked *ex ante* to state how much of their winnings they were prepared to share with the other subjects in their team of three who won nothing. Subject A was invited to declare the sum she would donate to B (or C) **if** A were to win DM10 **and** B (or C) was the only loser in the trio. Let us call this sum X. Then A was asked to select her donation to both B and C if neither B nor C were to win any money. Let this

sum equal Y and assume that 'losers' B and C split Y between them. 52subjects chose X $\cong$ Y (up to a rounding error), a finding which the authors label *fixed total sacrifice* (FTS) and show to be inconsistent with standard utilitarian altruism.

[32] For instance, in the Selten and Ockenfels experiment, symmetry means that, in A's eyes, *ceteris paribus* the loss of one expected currency unit (e.g., DM1) by a 'losing' subject B yields the same disutility for subject A as the loss of DM1 by a winning C who nevertheless donates DM1 to some other 'loser'.

[33] The willingness to make a sacrifice on behalf of others based on the expectation that, if roles are reversed, members of this target group will/should come to one's aid.

[34] By 'enlightened selfishness' we mean generosity motivated by the (selfish) hope that the beneficiary will re-pay the donor in the future. Furthermore, utilitarian altruism requires a specific person's utility to be introduced as a variable in the donor's utility function. But our definition of solidarity rules out person-specific motivation in two ways: First, by identifying solidarity as a subset of λ-*generosity* (which in itself rules out self-serving sacrifices as potentially λ-*generous* acts). Secondly, by tying solidarity up with other peoples' condition, rather than with their disutility from it.

[35] There is of course no doubt that a Kantian motivation may coincide with feelings of love, sympathy etc. However, Kant's point is that even when the latter are absent, the visit ought to take place. Our interest lies in the effects and nature of such purely Kantian acts of generosity.

[36] For example, i might be λ-*generous* to a group of pop-stars that she worships. However, given condition (1) this does not qualify as a case of σ-*solidarity*.

[37] Thus norm or custom-following [à la Akerlof (1980) and Varoufakis (1989)] do not qualify as examples of σ-*solidarity*. In this sense nor do the concerns for one's image within a group mentioned by Olson (1965) or Becker (1974) since, according to our definition, σ-*solidarity* is irreducible to social norms or public expectations.

[38] Effectively, we argue that, whenever $\lambda_i > 0$ but $\sigma_i = 0$, the explanation of i's λ-*generosity* must be sought in some of the other-regarding categories in Section 3.

[39] We believe, nevertheless, that σ-*solidarity* has important implications for justice: According to one perspective on justice, the latter flourishes when altruism reaches its limits. It comprises a set of constraints regarding our behaviour toward persons for whom we harbour no natural sympathy (for if we did, we would not need moral constraints in our dealings with them). In this paper we argue that something else is also born, in addition to justice, at the limits of altruism: Solidarity! It pertains to instances of sacrifice and generosity motivated by 'worthy causes', rather than by an altruistic urge to contribute to specific individuals. The single mother of our Boat Service example may feel no ethical obligation to yachtsmen on the grounds of any principles of 'justice'; and yet, she may contribute in response to an antipathy toward the abstract idea of a lone figure helplessly fighting a losing struggle against menacing seas. Similarly with the subjects in the Selten and Ockenfels (1998) experiment: Solidarity with the losers is a feeling quite distinct from a commitment to fairness. The interaction between solidarity and justice is an obvious area of further study.

[40] Sugden (1993) describes instrumental accounts of moral behaviour as: "parasitic on moral theories that enjoin us to behave in ways that are not instrumentally rational". Thus the presence of even a small percentage of persons capable of σ-*solidarity* may be the necessary initial condition for some bandwagon to start rolling (Akerlof, 1980; Varoufakis, 1989).

[41] Though not narrated in terms of solidarity, Sugden's (1986) main thesis is consistent with this account.

[42] For a discussion of expressive, versus instrumental, rationality see Hargreaves-Heap (1989).

[43] Hereafter the analysis will proceed on the assumption that the two groups do not overlap. However, the analysis generalises naturally when there are more than two groups and a person can belong to more than one at the same time.

[44] For example, a game with a unique equilibrium which awards higher payoffs to *K*-players than to *N*-players.

[45] For example, a symmetrical game with twin equilibria one of which favours the *K*-players, the other the *N*-players. If a convention evolves selecting the former equilibrium, *K*-players will, according to Definition 6, enjoy conventional social power over *N*-players. And vice versa.

[46] For the theoretical proof see Weibull (1989). Hargreaves-Heap and Varoufakis (2002) report on an experiment which confirms this theoretical intuition. In it, players were divided in two groups ('red' and 'blue') and only their colour was made known to their opponent. And yet, in repeated play of the hawk-dove game, one of the two groups (in some sessions the 'red', in others the 'blue') emerged as dominant. When later they played the HDC game above, the same pattern continued with one important difference: when dominant colour players were matched with one another, they never cooperated whereas when disadvantaged colour players met, they cooperated most of the time (a case of solidarity among the discriminated?).

[47] Selecting *h* can be interpreted as aggressive behaviour, *d* as acquiescent and *c* as cooperative.

[48] For example, in the context of conflict over property rights.

[49] There is, for instance, plenty of documented evidence of selfless, reciprocal sacrifice among the ranks of otherwise abhorrent groups and organisations (e.g., SS officers).

[50] Much ink has been expended in an attempt to come to terms with situations in which, for instance, the male victims of racial discrimination struggle to retain their exercise of arbitrary social power over their wives, mothers and sisters. In the sense of this paper, they pose simultaneously as the potential recipients of *ρ-solidarity* (in interactions with the white community, labour market etc.) and as parties to a collusion which fails the conditions of *ρ-solidarity* outright.

[51] See Hargreaves-Heap and Varoufakis (1995), Chapter 7 for an evolutionary model of how discriminatory conventions gain evolutionary fitness through division and multiplication.

[52] Since most philanthropical activity was part of the facade of Victorian socialising.

[53] Since the last thing on most Victorians' mind was the social process manufacturing systematic, large scale deprivation. Instead, they tended to focus on the personal responsibility of the wretched and the poor for the condition they found themselves in.

[54] Of course an economist might argue that the amelioration of the repugnant suffering, and the indirect utility so procured, is the solidaristic agent's reward. This is neither here nor there. Whether the reason for acting in solidarity with an 'other' is internal (e.g., indirect utility) or external to one's preferences is too rarified a question to delve into here.

REFERENCES

Akerlof, G.: 1980, 'A Theory of Social Custom of Which Unemployment May be One Consequence', *Quarterly Journal of Economics* **95**, 749–775.

Andreoni, J.: 1990, 'Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving', *The Economic Journal* **100**, 464–477.

Arnsperger, C. and Y. Varoufakis: 1999, 'Solidarity and Rational Contemplation', mimeo, University of Sydney, Department of Economics, May 1999.

Bacharach, M.: 1999, 'Interactive Team Reasoning: A Contribution to the Theory of Cooperation', mimeo.

Becker, G.: 1974, 'A Theory of Social Interactions', *Journal of Political Economy* **82**.

Brennan, G. and P. Pettit: 2000, 'The Hidden Economy of Esteem', *Economics and Philosophy* **16**, 77–98.

Camerer, C. and H. Thaler: 1995, 'Anomalies: Ultimatum, Dictators and Manners', *Journal of Economic Perspectives* **9**, 209–219.

Chapman, B.: 1998, 'More Easily Done Than Said: Rules, Reasons, and Rational Social Choice', *Oxford Journal of Legal Studies* **18**, 293.

Frankfurt, H.: 1971, 'Freedom of the Will and the Concept of Reason', *Journal of Philosophy* **68**, 5–20.

Gauthier, David: 1986, *Morals by Agreement*, Oxford University Press, Oxford.

Geanakoplos, J., D. Pearce, and E. Stacchetti: 1989, 'Psychological Games and Sequential Rationality', *Games and Economic Behavior* **1**, 60–79.

Hargreaves-Heap, S.: 1989, *Rationality in Economics*, Blackwell, Oxford.

Hargreaves-Heap, S. and Y. Varoufakis: 1995, *Game Theory: A Critical Introduction*, Routledge, London and New York.

Hargeaves-Heap, S. and Y. Varoufakis: 2002, 'Some Experimental Evidence on the Evolution of Discrimination, Co-operation and Perceptions of Fairness', *The Economic Journal* **112**, 678–702.

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, Cambridge.

Hollis, M.: 1998, *Trust Within Reason*, Cambridge University Press, Cambridge.

Hollis, M. and R. Sugden: 1993, 'Rationality in Action', *Mind* **102**, 1–35.

Hurley, S.: 1989, *Natural Reasons*, Oxford University Press, Oxford.

Hume, D.: 1739, *A Treatise of Human Nature*, edited by D. G. C. Macnabb, Fontana/Collins, London, 1962.

Kant, I.: 1788, *Critique of Practical Reason*, translated and edited by L. W. Beck in *Critique of Practical Reason and Other Writings*, Cambridge University Press, 1949.

Margolis, H.: 1981, 'A New Model of Rational Choice', *Ethics* **91**, 265–279.

McPherson, C. B.: 1973, *Democratic Theory: Essays in Retrieval*, Oxford University Press, Oxford.

Midgley, M.: 1994, *The Ethical Primate: Humans, Freedom and Morality*, Routledge, London.

Nietzsche, F.: 1887, *Genealogy of Morals*, Doubleday, New York, 1956.

Nowak, A. S. and T. Radzik: 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*, Harvard University Press, Cambridge, MA.

Rabin, M.: 1993, 'Incorporating Fairness into Economics and Game Theory', *American Economic Review* **83**, 1281–2302.
Rawls, J.: 1971, *A Theory of Justice*, Harvard University Press, Cambridge, MA.
Rouille D'Orfeuil, H.: 2002, *Finances et Solidarité*, La Découverte, Paris.
Schelling, T.: 1960, *The Strategy of Conflict*, Harvard University Press, Cambridge, MA.
Selten, R. and A. Ockenfels: 1998, 'An Experimental Solidarity Game', *Journal of Economic Behavior and Organization* **34**, 517–539.
Sen, A.: 1970, 'The Impossibility of a Paretian Liberal', *Journal of Political Economy* **78**, 152–157.
Sen, A.: 1974, 'Choice, Orderings and Morality', in S. Korner (ed.), *Practical Reason*, Basil Blackwell, Oxford.
Sen, A.: 1977, 'Rational Fools: A Critique of the Behavioral Foundations of Economic Theory', *Philosophy and Public Affairs* **6**, 317–344.
Sen, A.: 1999, *Development as Freedom*, Oxford University Press, Oxford.
Sprumont, Y.: 1996, 'Axiomatizing Ordinal Welfare Egalitarianism When Preferences May Vary', *Journal of Economic Theory* **68**, 77–100.
Smith, A.: 1759, *The Theory of Moral Sentiments*, edited by D. Raphael and A. Macfie, Clarendon Press, Oxford, 1976.
Sugden, R.: 1982, 'On the Economics of Philanthropy', *The Economic Journal* **92**, 341–350.
Sugden, R.: 1986, *The Economics of Rights, Co-operation and Welfare*, Blackwell, Oxford.
Sugden, R.: 1993, 'Thinking As a Team: Towards an Explanation of Non-Selfish Behavior', *Social Philosophy and Policy* **10**, 69–89.
Sugden, R.: 2000, 'The Motivating Power of Expectations', in J. Nida-Rümelin and W. Spohn (eds), *Rationality*, *Rules and Structure*, Kluwer, pp. 103–129.
Thomson, W.: 1995, 'Population Monotonic Allocation Rules', in W. Barnett et al. (eds), *Social Choice*, *Welfare and Ethics: Proceedings of the Eighth International Symposium in Economic Theory and Econometrics*, Cambridge University Press, Cambridge.
Van Huyck, J., R. Battalio, and R. Beil: 1990, 'Tacit Coordination in Games, Strategic Uncertainty and Coordination Failures', *American Economic Review* **80**, 238–248.
Varoufakis, Y.: 1989, 'Worker Solidarity and Strikes', *Australian Economic Papers*, June, 76–92.
Varoufakis, Y.: 1990, 'Solidarity in Conflict', in Y. Varoufakis and D. Young (eds.), *Conflict in Economics*, Harvester Wheatsheaf and St Martin's Press, Hemel Hempstead and New York.
Varoufakis, Y.: 1991, *Rational Conflict*, Blackwell, Oxford.
Weibull, J.: 1995, *Evolutionary Game Theory*, MIT Press, Cambridge, MA.

Christian Arnsperger
Chaire Hoover and FNRS
Université Catholique de Louvain
Belgium
and
Yanis Varoufakis
Department of Economics
University of Athens, Greece
and
University of Sydney, Australia

Manuscript submitted 20 November 2001
Final version received 21 February 2003