THE CONSTITUTION OF THE NONPROFIT ENTERPRISE: IDEALS, CONFORMISM AND RECIPROCITY[#]

Gianluca Grimalda[*], *Lorenzo Sacconi*[**]

1. Introduction

Studies dealing with the economic and social function of the nonprofit enterprise can be traced back into two major strands of literature. The first emphasises the peculiar failures – mainly median voter and asymmetry of information - of both the political and the market system in the provision of public or welfare goods (respectively Weisbrod, 1988; Hansmann, 1980), thus arguing for the necessity of new types of organisational forms of productive activity in those sectors. However, these models do not actually explain what in the peculiar institutional nature of the nonprofit should help to solve this kind of inefficiency.

The second approach does offer a "positive" explanation for the nonprofit firm, which draws on the idea that the agents involved in the nonprofit sector have *other-regarding* motivations such as altruism, will to conform to the established system of norms, a disposition to reciprocate the perceived fairness of others' action (for a review, see Rose-Ackermann, 1987). However, in our view this approach does not provide a sound theoretical foundation for these attitudes, which risks making the whole explanation void. Moreover, such a theory is at odds with evidence on the conflicts of interests that affect the agents involved in the nonprofit activity, as highlighted by the frequent practise of self-imposing norms involving fiduciary duties and codes of conduct even in the nonprofit sector. In fact, the reality of the nonprofit sector appears much more variegated than what would result from this approach.

The model we develop in this paper seeks to address both shortcomings that we perceive in the received theory. First, it takes on the question of individual motivations to choice, providing a general model of choice in which a variety of possibly conflicting motives to action is weighed up by an agent. In this setting, a seemingly altruistic behaviour is not a mere attitude of the individual, but is one of the

[#] Support received by the MIUR and the CARIPLO Foundation under the national research project "Economic comparative analysis of institutions and institutional complexity of governance forms, in the perspective of incomplete contracts" is gratefully acknowledged.

[*] Department of Economics, University of Trento and University of Southampton

[**] Department of Economics, University of Trento and CELE, Centre for Ethics, Law and Economics, University Cattaneo-Liuc, Castellanza

possible outcomes emerging from a process of rational evaluation of different motives to action. In the application of this model to the case of the nonprofit enterprise, we shall assume that agents' preferences are represented by a *comprehensive* utility function, in which two basic motives to action are considered: the first is a (standard) *self-interested* motivation, whereas the second is a *conditional willingness to conform to an ideology*, or a *moral principle*, which for brevity we shall call a *conformist,* or *ideal*, motive to action. Ideology, or morality, is shaped as a normative criterion of evaluation for collective modes of behaviours, which provides the agents with a ranking of the states of affairs based on their greater or lesser correspondence with the fulfilment of this normative principle. The ideology is seen as the result of a (possibly hypothetical) contract between the agents involved in the interaction in an *ex-ante* phase. This sets on a normative principle that offers an assessment in the *ex-post* phase of the social outcomes, broadly described, in terms of the fulfilment of the principle itself; that is to say, the normative principle boils down to a social welfare function that measures the correspondence of the outcome with the normative prescriptions provided by the ideology. Agents, therefore, use such a shared principle in order to measure their own and any other's degree of conformity with it, and we assume that one's own motivation to act in conformity with the principle increases with others' (expected) conformity. In other words, individual compliance with the ideology is *conditional* on others' compliance with it, as perceived by the agent. This peculiar feature of *reciprocity* over others' behaviour calls for an extension of the usual equipment of decision theory, which is provided by the theory of Psychological Games (Geanakoplos *et al.,* 1989).

Second, we propose a possible way in which the model is capable of accommodating for the piece of evidence mentioned above, namely the possibility of conflict of interests within the nonprofit. In fact, the nonprofit is one of the possible outcomes in a "game of production" where some relevant agents setting up the productive activity, ideally an entrepreneur and a worker, determine the *nature* of the organisation through their decisions. Since the structure of the interaction turns out to be that of a coordination game, then codes of ethics and self-imposed rules of conduct can be justified as devices extending the structure of the game in order to select the outcome corresponding with the nonprofit organisation, or, as we suggest for future extensions of the work, investments made to "reveal" the "true" type of the firm to external agents, e.g. donors, in a context of asymmetric information.

Overall, two are the features that are needed to turn the nature of the firm from *profit* to *nonprofit oriented* in the production game. First, agents must attach a sufficiently high weight to the conformist motive to action in comparison with the material loss that this may bring about. Second, the ideology that agents incorporate into their system of ends is shaped as the result of a (possibly hypothetical) 'social' contract between the relevant figures participating in the venture. In particular, ideology is *inclusive* in that not only does it take into account the interests of the agents active in the productive enterprise, but also the interests of other beneficiaries and stakeholders of the good produced. This additional category is represented in the model by the presence of a third agent, the consumer, who does not have an active role in the after-constitutional phase; i.e. she is a *dummy* player in the stage game of production. By conforming to the ideology, therefore, the promoters are aware that they are giving "voice" to some categories otherwise *excluded* from social consideration. Moreover, the ideology is assumed to apply a

fair and *efficient* distribution of the surplus in accordance with a *contractarian* criterion, since the interests of each participant – consumer included - are symmetrically accounted for (Brock 1979, Sacconi 1991, 2000). Given such an impartial perspective that characterises the ex-ante stage of agreement on the set of distributive principles, the resulting choice can also be said to embody a peculiar *moral* character. On an operative ground, the Nash bargaining solution is taken as the function representing this ideology. Overall, the nonprofit organisational form is seen as the result of a - possibly hypothetical – *internal* contract agreed upon by the relevant figures setting up the enterprise, which brings in the interests of the stakeholders *external* to the firm in an equitable manner. Therefore, ideology stands out as a crucial asset for the nonprofit organisation.

The first part of the paper is devoted to the development of the individual model of choice. Section 2 introduces the distinction between *consequentialist* and *conformist* individual preferences. The *material* and *ideal* game are then presented as representations of the same interaction though assessed from different standpoints, which adopt the self-interested consequentialist and the conformist attitude respectively. Finally, a general version of the comprehensive utility function is presented. Section 3 offers a specification of the conformist motive to action, introducing a peculiar notion of reciprocity in compliance with the ideology, which is based on an extension of Rabin's model of fairness (1993).

The second part of the paper aims to apply such a model of behaviour to the account of the nonprofit enterprise as a peculiar organisational form. Section 4 illustrates the setting of the "production game", where both the active players have one action improving the quality of the good and another one that leaves it unaltered with respect to a free market standard. The surplus of the consumer is directly linked with the number of agents performing the quality-improving action. It is then shown how this stage game leads to different solutions depending on whether it is evaluated from the self-interested standpoint (material game) or from the ideal one (ideal game).

Section 5 explores the final solution of the production game when the two conflicting attitudes are blended into the comprehensive utility function. We show how an equilibrium is possible that leads both active agents to perform the quality-improving solution, provided that the weight attributed to the ideological motivation is sufficiently high. However, we observe that under the same conditions there exists another equilibrium in which agents perform the non-quality-improving action, besides a third equilibrium in mixed strategies. Noticing that the structure of such a psychological game resembles that of a coordination game, we suggest that the issuing of codes of ethics by the firm may act as a cognitive device able to generate determinate expectations for the agents over the quality-improving equilibrium. We finally interpret this result as a main underpinning for the nonprofit firm.

2. The System of Choice of the Agents

2.1. Self-Regarding and Other-Regarding Motives to Action: an Overview

The idea that individuals take into account a large number of reasons to action when making decisions, which extend well beyond the stereotypical self-interested motive, is now largely accepted

among students of rational choice. As Binmore puts it (1994: 19), "not even in Chicago are the views [that homo-economicus strictly abides by her own self-interest] given credence any more". This set of supplementary motivations may include altruism, the willingness to act in accordance with the received sense of morality, or the want to conform to the behaviour or the expectations of the other members of the community. In principle, every type of motivations, even those dictated by a person's whims, or by self-destructive and anti-conformist desires, can be included in one's system of ends.

Therefore, according to this view, the range of the agent's possible motives to action is left as ample as possible. In other words, there is no constraint on the set of ends that the agent may like to pursue, but the correspondent choices need to satisfy requirements of internal consistency in order to be called rational. In particular, when a sequence of choices made under different circumstances – that is, under different values of the "parameters" that frame the context of choice – fulfils the basic axioms of transitivity, completeness, reflexivity, and possibly some others, then the internal consistency and thus the rationality of the action can be said to hold. The utility function does not have any intrinsic meaning if not for acting as a formal device to represent such a coherent system of choice[1]. In particular, individual rationality is not assessed on the grounds of the agent's effectiveness in pursuing some notion of self-interest, but rather on the logical internal coherence of her choices with respect to her ends: even the behaviour of a saint can be assessed in terms of rationality in much the same way as that of a *homo-economicus*[2].

Ben Ner and Puttermann (1998) provide a theoretical underpinning for such a model of individual choice, by distinguishing between *self-regarding*, *other-regarding* and *process-regarding* motivations. The difference between them depends on whether the agent is concerned with the consequences of her action on herself, on others, or on the way outcomes are brought out, respectively. We shall expand on this point in the next section. Another way of representing these ideas has been put forward by Copp (1997), who associates different reasons to action to different *standpoints* that can be adopted in assessing a particular social outcome. In particular, a self-regarding motivation stems from the adoption of a standpoint that is *internal* to the individual, where the *standard* of assessment is some form of her well-being. In the case of the other-regarding motivations, the agent uses a perspective *external* to that of her own self. In this case, she may adopt the standpoint of a single agent different from her, which may lead to altruism, or that of the "team" she is part of (Sugden, 2000), or the point of view of an impartial observer sympathetic to each member of the group of agents (Harsanyi, 1977).

Only recently have some contributions been put forward that build on this background theoretical framework to provide working models of choice. In particular, Bernheim (1994) and Sugden (1998a, 1998b) add to the self-interested motivation a second one given by the desire to obtain the commendation and avoid the disapproval of others with respect to one's own actions. In these models the other-regarding motive is thus associated with the desire to live up to others' expectations, which is the reason why these approaches are generally referred to as *normative expectations* models (Sugden, 1999).

Another strand of contributions is connected with the flourishing body of literature in Experimental Economics, where the evidence gathered in laboratory experiments on individual behaviour, somewhat unaccountable by relying only on self-interested motivations, have spurred the elaboration of new

hypotheses in choice models. Fehr and Schmidt (2001) distinguish theories where agents are endowed with 'social preferences' - that is, their utility function also depends in some way on the payoffs distribution amongst them – and theories where agents are motivated by 'intentions-based' reciprocity; that is, the individual is spurred to replicate the 'intention' perceived in others' actions.

In particular, the social motives taken into account in the first approach include aversion to inequality in surplus distribution, or some form of altruism, or concern for the individual position within the payoffs ranking. The second approach builds on Rabin's seminal model of fairness (1993). The main idea is that the agent may assign a different value to others' actions depending on how she perceives their *intentions* in bringing them out. For instance, an action may be deemed as kind when it brings about an extra utility with respect to what expected in relation with some standard of behaviour, or it may be perceived as nasty when it leads to an unexpected loss. As a matter of fact, according to investigations in Psychology, a key trait in human behaviour is to reciprocate the intention perceived in others behaviour with an action of the same sign. On this view, Rabin's model is a formal device to incorporate these observations into individual choice theory.

The theory of Psychological Games provides with some tools to embody these considerations into a formal analysis. In fact, it introduces *beliefs,* of every possible order, on each other behaviour into the utility function (Geneakoplos *et al* (1989)). In this fashion, it is possible to model the idea that an agent can be more or less satisfied depending on how others' actual action correspond to her initial expectations. In particular, for simplicity restricting the analysis to the case of two-person interactions, Rabin considers a pair of 'kindness functions', which measure the extent to which the agent's and her counterpart's actions increase or diminish one another's expected payoff. This estimate is used by each agent to appraise the kindness of the other party to herself, on the grounds of her second order expectation, and the kindness of the subject herself toward the other agent, as perceived on the basis of her first order expectation. The way in which these functions are constructed is to consider the best and the worst payoff that each agent can cause to the other on the basis of the reciprocal expectations, and then to consider how the payoff actually brought about lies between those two extremes[3].

Other models have been developed in which agents' social preferences and intention-based reciprocity attitudes are both present in individual motivations. For instance, in Charness and Rabin (1999) the 'weight' that each individual attaches to each other individual in her own social preference depends on the disesteem with which the agent herself thinks of the others, which is appraised in terms of the 'distance' of others' behaviour from a purely disinterested one. Likewise, in Falk and Fischbacher (1999) each agent computes a 'benevolence term' for any other agent, which depends on the degree to which any other agent's action has increased or diminished the inequality in the overall distribution. This term is then multiplied by a 'reciprocity term' that is positive or negative in relation to the other agent's action being perceived as kind or hostile. Finally, another parameter measures the relative weight attached to material utility with respect to that of reciprocity on the social distribution.

The model that we intend to build is similar to those now illustrated in that the aspects of reciprocity are related to some forms of normative evaluation of the social states. However, as we shall argue in section 3 and 5.1, it differs from them in the content of the normative function.

2.2. Conformist Preferences

2.2.1. Consequentialist Preferences Vs. Deontological Reasons to Prefer

In the present paper we shall embrace the view outlined in the previous section that the number of motivations that agents consider extends beyond the standard self-interested reasons to action. However, we believe that prior to the distinction between self-regarding and other-regarding reasons to action there exists an even deeper distinction between *consequentialist* and *conformist* types of preferences of the individual, on which our model will be grounded. Given the importance of the matter, we devote the present section to put forward in detail the theoretical underpinnings of the individual system of preference.

Simply stated, preferences can be said to be consequentialist when they are defined over the consequences of the agents' actions. Consider a situation of strategic interaction involving many agents. This generates states of affairs that can be differently described according to their different characteristics. If these are meant as consequences, they are understood as *what happens* to the *decision maker* in a state – i.e. as the consequences *to the decision maker* herself - or what happens to *any* subset of individuals or to *every* individual – that is the consequence *to anyone* in the same state. In the first case the characteristics under consideration would be an attribute of the single agent herself – such as her wealth, leisure, effort, etc. - and they are the result of a *one to one* mapping between the state set and the consequence set held by *one particular* individual (the decision maker). In the second case the characteristics under consideration would be attributes of some set of individuals (possibly all of them), and they could be defined by a *one to many* correspondence between the state set and the consequences sets held by all the concerned individuals.

The distinction between self-regarding and other-regarding consequentialist preferences thus depends on whether the list of characteristics takes into account only self-referred consequences, or also consequences to other agents. In the former case we have *self-regarding* consequentialist preferences. When instead the agent takes account of the consequences of social interaction on other individuals, other-regarding consequentialist preferences obtain. Notice that this definition does not necessarily imply a benevolent disposition of the self towards other people, but only that individual preferences are affected by the outcomes occurring to other people as well as theirs. For instance, even a sentiment of hate towards another person would be regarded as implying an other-regarding preference for the very fact that the set of consequences that the agent considers includes that of other agents. To be sure, however, other-regarding preferences are the natural source for individual *moral* preference of a consequentialist type - namely, preferences over every individual's consequences impartially weighted. In particular, we have utilitarian moral preferences if, besides accounting for everyone's consequences, we further require that each agent's consequences are assessed from the point of view of each agent's preferences, thus implicitly calling for interpersonal comparisons of utility, in addition to the requirement of summing up each agent's utility. Altruistic preferences are another special case, in which the agent attaches a high, weight to the consequences for other agents, as assessed from the point of view of *their* own self-referred preferences, rather than from her own.

Let us now come to the second basic type of preferences of the self, which we call *personal conformist preferences* as opposed to personal consequentialist preferences. As well as the first type of preferences, conformist preferences are defined over states of affairs, but nevertheless they are not described in terms of consequences occurring to any individual, but as patterns of individual, interdependent or collective behaviours, and as beliefs about such modes of behaviour. We put a *deontological* element at their basis, since these preferences are grounded on some intrinsic characteristics of the agents' actions rather than on their consequences. In other words, agents are motivated to act by the awareness that their actions satisfy some formal properties rather than from the mere outcomes of their actions[4]. For instance, the agents may attach utility to the knowledge that the decision procedure they follow is "fair" according to some definition. Again, it is possible to draw a secondary distinction between self-regarding and other-regarding *conformist* preference, where the former refers to the case in which the agent only cares about the intrinsic characteristic of *her* own action, whereas the latter points to the characteristics of both her own and the others' participants actions.

In order to better understand the distinctions between the two basic concepts of preferences, the following elements are to be considered in sequence: i) the relevant description of states of affairs (sec. 2.2.2), ii) the preference ordering over states of affairs as it depends on the relevant description of the states (sec. 2.2.3), iii) the induced preferences ordering over the actions set of each individual player, iv) the numerical representation of such preferences by an utility index that we call *ideal utility*(sec. 2.2.4)[5].

2.2.2. The Relevant Description of States of Affairs

States of affairs are now primarily *described* as sets of interdependent actions - to whom each player' beliefs over the other's actions are appended. These are considered with respect to their conformity (or lack of conformity) to a given abstract principle of justice. Under this description states are modes of deontological individual or collective behaviours performed by the players. We may fix a pattern of behaviours (a vector of strategies) that is meant as perfectly *deontological* because it fully conforms to an abstract principle of fairness or to a fair criterion of benefits distribution amongst the concerned parties. Call such a state the *ideal*. Then we may look for the degree of conformity to the ideal displayed by each state of affairs resulting from the individual choices actually performed by all the players (or by each player's choice given other players' choices.) In other words, we allow for the possibility that agents experience different levels of 'utility' - that is, different degrees of motivational strength - in relation to the degree to which the normative principle can be said to be fulfilled. In particular, in order to define the character of *mutuality* of the preference (see next section), it will be important to single out the individual contribution to the accomplishment of the ideal state and, conversely, the individual responsibility in the deviation from it.

Another point deserves some comments. The principle of justice to which agents desire to conform their actions, may well be a principle of *distributive* justice, and this will indeed be the case in our model. Therefore, the outcomes for the agents *have* to be taken into account in order to check the degree of fairness of the surplus distribution. This does not reduce the second type of preferences to the first. First type utilities are no more than the *rough materials* of the second type. We must know about outcomes where *utilities for consequences* are allocated amongst the players in order to describe whether they

correspond to the ideal distribution defined according to an abstract principle. A principle of fairness (some given fair bargaining solution to a social contract model) accounts for each state according to a distribution criterion. This enables us to say whether the occurring vector of strategies in each state determines a payoffs distribution consistent with the abstract principle of fairness. But what matters for the relevant description of the states of affairs are not consequences or material payoffs as such, but the description of a *distributive* property of the payoffs. Under this description there is no individual to whom the relevant state of affairs happens as a *consequence*. We simply have a distribution stating the *ratio* according to which a pie, which provides an amount of overall surplus as high as possible, is partitioned amongst different players. Consequently, we may say that the concern for outcomes is in this case only indirect, as the interest of the agents lies in the compliance with the ideal principle of justice rather than on the consequences that this brings about. The content and the features of such a principle will be specified in more detail in section 2.2.4 and 5.1.

2.2.3. Mutual Conformity

Preferences are then defined not directly over consequences, but over *acts* n the grounds of their conformity to an abstract norm, i.e. a distributive principle. It is apparent that the preference ordering over states depends on an objective measure of conformity of any vector of actions to the abstract principle of fairness as it is built into the description of each state of affairs (as seen through the beliefs of the players over the others' actions). The more a state of affairs is expected to conform to the ideal, the more it is preferred by a player, i.e. the degree of expected reciprocal conformity is used as the basis for defining each player's preference ordering over states. Therefore, conformity is the characteristic that we assume is considered by players in order to say how desirable a state is. In particular, besides allowing for the 'measurability' of the extent to which the set of interdependent actions fulfils the abstract normative principle, we also assume that the expectation of greater conformity by other agents spurs a greater incentive in the agent to conform as well. In this sense, at the basis of conformist preferences lies a measure of how much deontology is built into the expected pattern of behaviour displayed by all the players in each state. This type of preference may be deemed as *conformist*, in that it consists of the desire to have the rules *ex-ante* accepted by an agent to be obeyed by everyone else. The type of conformism we are describing is nonetheless *moral*, in that the principle whose general observance triggers utility is, in our model, the result of an *ex-ante* unanimous and impartial rational choice.

In this aspect there lies the major difference between approaches like Sugden's and ours: in Sugden (2000) there does not exist an independent normative condition shaping the rule that agents are required to conform to: in fact, agents pay a disutility (a penalty) for not living up to anyone else's expectation. This implies that virtually any outcome of the game can emerge as the "moral" rule to be followed, since every convention can find support by means of the motivational force engendered by the expectations of the community members (Grimalda, 2001). Hence, the heuristic power of such an approach can appear questionable, since it seems that every norm can command conformity for the very fact of having come into existence. In our model, we take a different route in modelling conformism in that the rule must reflect an abstract principle of justice, whose only requirements are to be rationally acceptable and fair in an ex-ante perspective. In other words, not all of the patterns of mutual conformity, but only those

satisfying ex-ante properties of rational acceptability, are those embraced by the agents. We shall take the Nash solution to bargaining games, and the corresponding social welfare function, to represent such a principle (see section 5.1).

One important feature of our approach is that, despite the deontological element put at the basis of conformism, we must not give up the ultimately subjective nature of preferences over states of affairs, meant as some sort of subjective affection of the players (Gauthier 1986). In fact there is no reason to conclude that the preference criterion should be based on some objective *value* having an ontological reality "out there", completely independent on the affections or the judgement of those who are asked to express their preference. Notice that while conformist preferences depends on degrees of conformity - which are an objective measure of the levels of deontology built in the description of states - nonetheless deontology is meant as conformity of actions to a fair distribution principle that we have simply rationally agreed upon. As we shall illustrate in section 5.1, rationally agreed principles of fair distribution are in our model meant as what players would accept in an hypothetical bargaining situation amongst symmetrically rational bargainers, who are all equally driven by rationality postulates derived from the same principle of utility maximisation under strategic interaction, but as well equally incapable to identify their own particular name and role in the game.[6]

To clear up the matter, let us state the hierarchy within which the different pieces of the argument should be understood so far. First of all, for each player it is taken for granted the existence of some first order utility function defined on states, which are initially described in terms of the consequence that each player gets from feasible surplus allocations. Second, players accept some terms of agreement concerning surplus distributions. This agreement is worked out according to a fundamentally subjective notion of unanimous rational choice under ideally symmetrical bargaining conditions. Moreover, it defines a norm for distributing benefits in any game situation of the kind under consideration. Third, this principle is adopted as the ideal term of reference in order to measure "conformity" of states of affairs - described as vectors of interdependent actions - to a principle of fairness, and this introduces a deontological assessment of states of affairs.

The result is a preference ordering defined over states of affairs, which we hold not merely because of our primitive psychological desires for material payoffs or preferred consequences, but *because* it conforms to a rationally agreed abstract principle. That conformist preferences are based on a principle derived in turn from a rational bargaining model (over payoffs distributions), does not make less deontological the reason of preference at this second level of the argument. Nonetheless the deontological nature of these second order preferences does not make them dependent on values (ontologically) objective in nature or completely independent of the decision maker's affectivity or judgment. Duties are simply those we have rationally agreed upon in a hypothetical bargaining situation.

2.2.4. Preference Orderings and Ideal Utility

In the end, what really matters are each player's preferences over her own actions. As consequentialist preferences induce personal preferences over the actions' sets of every player, this must also be true for conformist preferences. Simply, these are induced by the conformist preferences over states described so far. If a player thinks that a strategy combination conforming to the principle of fairness is currently the

most probable state of affairs, then she will prefer her action that conforms to the duty – call it the deontological action – exactly *because* it contributes to bring about a state of affairs conforming to the duty.

To state it a bit formally, agent A *conformistically* prefers action X_1 to action X_2 if A observes an action Y by the other player B that would bring about a state of affairs S (a strategy vector) that conforms to the principle P if chosen together with action X_1 more than together with action X_2.

This definition however hides how important beliefs are to the definition of personal conformist preferences. We must account for the fact that a player, while he does not *observe* vectors of actions as such, on the contrary holds beliefs over other players' actions and over other players' beliefs over his own action. Thus he holds preferences over actions according to whether these actions, along with what she believes other players will do and what she believes other players will believe about what she does, contributes to bring about states of affairs that conforms to a rationally agreed principle of fairness.

To give again a formal definition, agent A *conformistically* prefers action X_1 to action X_2 if she believes that agent B will adopt the action Y, given that he (B) believes that A chooses action X_1, so that by choosing action X_1 (together with act Y) agent A believes to bring about a state of affairs S that conforms to principle P more then by choosing action X_2. This definition makes it natural explaining personal conformist preferences of agent A as resting on a hierarchy of mutual beliefs, within which any layer of beliefs of each party is justified by a higher order layer of beliefs.[7]

Since conformist preferences are also two-place relationships, by assuming that they satisfy the usual conditions of completeness and transitivity, we can derive a standard preference ordering over the strategy set of an agent [8]. Thus, even if conformist preferences are defined over characteristics of joint actions, rather than on their consequences, this does not impede to represent them by means of a utility function, which would satisfy in addition the usual axioms of expected utility. We call it individual *ideal utility* of actions as it is based on the agent's conformist preference ordering on actions.

In what follows, we will provide an example of a utility function that additively compounds the self-interested consequentialist motive to action and the deontological-conformist one. The two will be associated with what we call a *material* and a *conformist*, or *ideal*, (source of) utility, which, under a reasonable assumption of separability, make up the individual *comprehensive utility function*. The existence of this pair of different attitudes calls for two different types of analysis, coming down to two different concepts of solution of the same basic game situation under scrutiny. We call the first type of analysis the *material* game, in which the self-interested attitude is dominating and agents are only concerned with their material utility: this will be given a formal illustration in sec. 2.3. The second is the *ideal* game, where instead the deontological source of preference is the relevant one and agents are concerned with their ideal utility, as shown in section 2.4. The final choice of the agent will be based on how these two prompts to action are combined in the comprehensive utility function, and in particular on the weight that the agent assigns to one rather than to the other prompt to action.

2.3. The Material Game

It is given a game *G*, made up as usual by a triplet of elements: a set *I* of players, a set of strategies S_i and a utility function U_i for each agent. Formally, $G = \{I, S, U\}$, where $S = \underset{i \in I}{\times} S_i$ defines the set of feasible strategies profiles, and likewise *U* is the set of vectors of utilities. Allowing for the use of mixed strategies by the agents, we can further introduce the operator Δ(X) to express the randomisations over a set of elements X. We can thus define the set of possible randomisations over the strategy sets of the agents: $\Sigma_i := \Delta(S_i)$; finally, we can consider the vector including a randomisation for each agent: $\Sigma := \underset{i \in I}{\times} \Sigma_i$, where the generic element is indicated with $\sigma \in \Sigma$.

In the game G, the utility functions represent a measure of the self-interest of the agents, thus reflecting the first type of motivations illustrated in section 2.2. They are defined, as customary, firstly over the outcomes of the games - that is over the consequences to any player attached to a given way of playing the game, such that they are functions of the profiles of pure strategies: $\overline{U}_i(S)$.[9] Furthermore, taking on standard assumptions regarding expected utility, we introduce Von Neumann-Morgestern utility functions defined over mixed strategies profiles, $U_i(\Sigma)$, where $U_i(\sigma) := \sum_{s \in S} P_\sigma(s) \overline{U}_i(s)$. $P_\sigma(s)$ represents the probability that the pure strategy profile *s* is played according to the mixed strategy profile σ. Provided that the nature of this game does not differ from the standard, the relevant concept of solution would be the Nash's one.

2.4. The Ideal Game

The ideal game differs from the previous one in that agents evaluate the social situation from a different standpoint than the self-interested consequentialist one, possibly including the evaluation of the material payoffs of other agents who are *affected* by their actions but *cannot affect* the final outcome. Hence, we introduce an *ideal* game *G** as an extension of the material game *G*, in which the set of players is possibly larger than in the material game thus modifying the corresponding set of utilities. Formally, this game is defined by the triplet: $G^* = \{I^*, S, U^*\}$., with $I \subseteq I^*$ and $U^* = \underset{i \in I^*}{\times} U_i$. Notice that the set of actions *S* is left unaltered with respect to the material game: by definition the players now included in the game are *dummy* players in the original one.

Resting upon this construction, we can now introduce the notion of a *normative principle* used to appraise social state of affairs resulting from strategic interaction. This generates a ranking of the strategy combinations made on the grounds of the ideology, or the moral principle, which is ex-ante accepted by the agents. Notice that this ranking is established according to the level to which the vectors of material utilities (the standard payoffs vectors) satisfy a given formal *distributive* property, that is whether, attached to any outcome, a distribution of the material utilities does materialise that satisfies a normative property T. Consequently, we are assuming that it is possible to measure on some scale the

correspondence of the social states of affairs to an ideal norm of assessment, which is represented by a function of the social outcomes. This is analogous to an *individualistic* social welfare function in that it is dependent on the material utilities of the agents involved in the interaction and establishes a certain formal property of the material utilities' distribution amongst the agents themselves:

$$\overline{T} := \underset{i \in I^*}{\times} \overline{U}_i(S) \rightarrow R$$

Therefore, such a normative principle permits the creation of an ordering over the possible states of affairs (strategy vectors like s∈ S), which represents the assessment that an impartial spectator would give to the different social situations on the basis of the relevant normative criterion of distribution. A higher value of the function T, defined over outcomes, implies that the associated social state of affairs satisfies to a higher degree the normative criterion.

Of course, taking the structure of the game as granted, it is possible to make the function directly dependent on the pure strategy profile set S, and, also, on the mixed strategies of the game:

$$T(\sigma) := \sum_{s \in S} P_\sigma(s) \overline{T}\left[\overline{U}(s)\right].$$

In analogy with individual expected utility, the expected normative function is simply a weighed sum of the indexes of welfare distribution under all possible pure strategies profiles, with weights given by the probabilities that each outcome is actually played.

2.5. The Comprehensive Utility Function

As already pointed out, we allow for an agent having various, possibly conflicting, motives to action in her own system of deliberation. The first is given by the usual self-interested motivation, whereas the second hinges upon the ordering of the social outcomes that is carried out by means of the normative principle *T* introduced in the previous section. It consists of the utility derived from the knowledge that the action performed by the agent, given her expectation on others players' action, satisfies, to some extent, the normative principles *T* with respect to the assessment of the social states of affairs based on the ranking of the corresponding outcomes.

We now introduce what we call a comprehensive utility function, whose components are given by the material and ideal utility. In what follows we shall assume that the agents are able to fully compare this pair of reasons to action and to take a decision, thus leaving aside the issue of commensurability of different sources of value[10].

The comprehensive utility function will then have the following form:

$$V_i(\sigma) = U_i(\sigma) + \lambda_i f\left[T(\sigma)\right] \quad i \in I^*$$

The first term U_i represents the material utility and is shaped in accordance with the agent's self-interested consequentialist preferences. The second term is the ideal utility and reflects the agent's concern with other types of reasons to action, meant in general as the degree of conformity of the social state of affairs - the agent's and the others participants' behaviours - to the normative principle of welfare distribution *T*. This is expressed as a function *f*, shared by all agents, of the social normative criterion *T*.

For simplicity, the two components enter the function additively, and the parameters λ_i, possibly different for the agents, measure the weight attributed to their ideal rather than material source of utility. The function f may be specified in different ways in order to account for various possible forms of the morality-grounded motive to action. In the following section we shall provide a particular specification based on an idea of expected mutuality in conforming to the normative prescriptions.

3. Mutual Conformism

3.1. A Reciprocity-Based Account of the Ideological Motive

The model that we wish to develop emphasises the aspects of reciprocity in acting in accordance to a shared normative principle embodying an ideology, as represented by the welfare distribution function *T*. In particular, the idea we want to capture by means of our model is germane to the common approach in the literature on moral philosophy that sees agents as available to sustain a 'just' action, but possibly detrimental in terms of self-interest, only insofar they expect other agents to do the same. Indeed, this is a restatement of the usual notion of reciprocity, where this is now intended in a general sense and with respect to a normative principle, rather than in a two-side relationship where agents are concerned with each other's payoffs.

We model this account of reciprocity by building on Rabin's model of fairness (see sec. 2.1). In particular, Rabin's kindness functions are substituted by functions of expected conformity with the normative principle, so that each agent's incentive to perform an action satisfying the moral principle, and possibly contrasting the self-interested reason to act, is positively linked with the extent to which the opponent is expected to perform an action consistent with the same normative principle. In this way, we model the idea that agents derive utility from their expected reciprocal conformity to a shared normative principle, rather than from an expectation about how kind they are one toward the other in terms of the satisfaction of their own consequentialist preferences.

3.1.1. Expected Conformity to the Ideology

To model these ideas, we need a further extension of the analytical structure of individual preferences, derived from the approach of Psychological Games (Geneakoplos *et al*, 1989). In principle, the formal apparatus requires the construction of hierarchies of beliefs of infinite order, but this aspect is much simpler here since, for our purposes, beliefs of the first two orders are all of what is needed in order to give an account of reciprocity.

A first order belief for player i is a probability measure over the other players' mixed strategy set, namely $B_i^1 := \Delta(\Sigma_{-i})$; thus the generic element $b_i^1 \in B_i^1$ indicates the probability with which i believes that the other players are going to implement the profile of strategies σ_{-i}. In the same fashion we can define $B_{-i}^1 := \times_{j \neq i}(B_j)$. Obviously, when there are just two active players, we have $B_i^1 := \Delta(\Sigma_j)$ and $B_{-i}^1 := B_j$. A second order belief for player i is a conjecture over the belief of j over i's strategies. Therefore, it consists of a probability measure over the Cartesian of other players' beliefs of first order:

$B_i^2 := \Delta\left(B_{-i}^1\right)$. Thus the generic element of this set, $b_i^2 \in B_i^2$, represents *i*'s probability that the belief of *j* over *i*'s strategies is $b_j^{1\,11}$. We shall indicate with $b_i = \left(b_i^1, b_i^2, ...\right)$ the infinite-dimension vector collecting the beliefs of each order for player *i*.

We now restrict our attention to a two-person game, even though a generalisation to the case of n players would be straightforward. In analogy with Rabin's pair of *kindness* functions, measuring the mutual impact of one's actions on the other's individual utility, we can now introduce functions computing the degree of conformity to the ideal - i.e. a moral principle (we call it thereafter the ideology). We first define *i*'s conformity to the ideology in the following way:

$$f_i\left(\sigma_i, b_i^1\right) = \frac{T\left(\sigma_i, b_i^1\right) - T^{MAX}\left(b_i^1\right)}{T^{MAX}\left(b_i^1\right) - T^{MIN}\left(b_i^1\right)}$$

where $T^{MAX}\left(b_i^1\right) = \arg\max_{\Sigma_i} T\left(\sigma_i, b_i^1\right)$ and $T^{MIN}\left(b_i^1\right) = \arg\min_{\Sigma_i} T\left(\sigma_i, b_i^1\right)$. In other words, $T^{MAX}\left(b_i^1\right)$ and $T^{MIN}\left(b_i^1\right)$ are respectively the maximum and minimum value that the welfare distribution function, representing the normative principle or ideology, can assume, depending on *i*'s action, given *i*'s first order belief b_i^1 over the action that *j* is going to perform[12]. Therefore, if $T^{MAX}\left(b_i^1\right)$ $\left(T^{MIN}\left(b_i^1\right)\right)$ is obtained, then agent *i* is maximising (minimising) the welfare function given his first order belief. $T\left(\sigma_i, b_i^1\right)$ is instead the value of the welfare function corresponding to *i*'s actual choice σ_i, given what he expects from player *j*.

Hence, $f_i\left(\sigma_i, b_i^1\right)$ is an index varying between -1 and 0 expressing the extent to which *i*'s action satisfies the normative criterion associated with the function *T*. When $f_i\left(\sigma_i, b_i^1\right)$ is equal to 0 (-1) it means that *i* is exactly performing the strategy maximising (minimising) the welfare function, given *i*'s first order belief, and this proves that his action is consistent with the normative prescriptions at the maximum (minimum) degree. In other words, conformity to the agreed upon normative principle is measured by the extent to which one's action reduces the distance between the actual state of affairs and the ideal one, that is the state where the value of the welfare distribution function is maximised over the agent's strategy set, given the expected choice by the counterpart.

To model the concept of reciprocity in the individual motivational system, we need to introduce a function symmetric to that set out above. This is the esteem that player *i* forms about *j*'s compliance with the ideology:

$$\tilde{f}_j\left(b_i^1, b_i^2\right) = \frac{T\left(b_i^1, b_i^2\right) - T^{MAX}\left(b_i^2\right)}{T^{MAX}\left(b_i^2\right) - T^{MIN}\left(b_i^2\right)}$$

where $T^{MAX}\left(b_i^2\right) = \arg\max_{\Sigma_j} T\left(b_i^2, \sigma_j\right)$ and $T^{MIN}\left(b_i^2\right) = \arg\min_{\Sigma_j} T\left(b_i^2, \sigma_j\right)$. Therefore, $T^{MAX}\left(b_i^2\right)$ and $T^{MIN}\left(b_i^2\right)$ are the value that the welfare function takes when player *j* respectively

maximises or minimises it, given the second order belief of player *i*. In other words, those functions indicate the maximum and minimum values that player *j* can attribute to the welfare function, given the belief he has about *i*'s action as perceived by *i* himself. In fact, recall that such a function measures the *esteem* of *j*'s compliance to the ideology as measured from *i*'s standpoint. Thus, if player *i* has formed a belief b_i^2 about the player *j*'s belief over *i*'s action, she will judge *j*'s actions from this point of view. She will then consider the best and the worst value that *j* can do with respect to the welfare function, and then compare these values with $T\left(b_i^1, b_i^2\right)$, which is the actual value that *i* expects the welfare function to take according to his beliefs. Alike the twin function $f_i\left(\sigma_i, b_i^1\right)$, a value of $\widetilde{f}_j\left(b_i^1, b_i^2\right)$ equal to 0 (-1) indicates the maximum (minimum) degree of conformity by player *j* to the ideology as embodied in the welfare function *T*.

3.1.2. The Comprehensive Utility Function

We can now introduce the final version of the utility functions. Notice that, as in every psychological game, the utility of an agent depends on her beliefs over the different possible outcomes (strategy vectors). We assume the following representation, which blends the two functions of compliance to the ideology:

$$V_i\left(\sigma_i, b_i^1, b_i^2\right) = U_i\left(\sigma_i, b_i^1\right) + \lambda_i\left[1 + \widetilde{f}_j\left(b_i^2, b_i^1\right)\right]\left[1 + f_i\left(\sigma_i, b_i^1\right)\right]$$

The fact that b_i^1 now substitutes σ_j depends on the fact that only in equilibrium the two are assumed to coincide. The ideal utility, again weighted by the coefficient λ_i, consists of the product of the two conformity functions augmented by 1.

The idea we wish to capture through this specification is twofold. On the one hand, the agent's utility depends positively on the realisation of the "best" social state of affair, in terms of the satisfaction of the normative criterion; indeed, the ideal utility is increased when an agent performs an action increasing the value of T, whoever she is. The second aspect is the character of "reciprocity" in the compliance with the normative criterion: in fact, the (esteemed) conformity of the other player, as expressed by $\widetilde{f}_j\left(b_i^1, b_i^2\right)$, may be seen as the "marginal incentive" that the subject has in pursuing her ideal motivations, as represented by $f_i\left(\sigma_i, b_i^1\right)$. Therefore, the ideal utility increases as the counterpart's action is perceived as more consistent with the ideology, thus eliciting a similar behaviour in the agent herself. In the extreme case in which $\widetilde{f}_j\left(b_i^1, b_i^2\right)$ is equal to –1, which denotes the worst action that agent *j* can perform in terms of the normative principle, the coefficient of the ideological motive gets equal to zero, thus leaving the self-interest as the only relevant motive to action[13]. Conversely, when $1+\widetilde{f}_j\left(b_i^1, b_i^2\right)$ is positive and sufficiently "large", then agent *i* may accept to pursue an action that is contrary to her self-interest but conform to the normative principle[14]. In general, the evaluation of the opponent's conformity to the normative principle magnifies or shrinks the individual motivation to act in accordance with the normative principle as well.

3.1.3. The Psychological Nash Equilibrium

The peculiar innovation introduced in the comprehensive utility function, that is the inclusion of beliefs in the arguments of the function, calls for an extension of the standard concept of solution of games, namely the Nash equilibrium. We shall adopt the original notion of Nash psychological equilibrium put forward by Geanakoplos *et al.* in their seminal contribution, although some refinements of this notion have been suggested (Van Kolpin, 1992) and others appear possible.

The underlying idea of this concept is that, if we are in equilibrium, then the beliefs of rational players must be coherent with the strategies that are there being played. As an example, if in equilibrium I observe my opponent playing the (possibly mixed) strategy $\sigma_j \in \Sigma_j$, then my first order belief must assign probability one to that particular strategy and 0 to all of the others. This is tantamount to saying that once an equilibrium has been reached, all of the first order beliefs must be single-point distributions assigning probability one to the equilibrium strategy. The higher order beliefs are then generated upon a condition of *coherence* with this initial condition (Geanakoplos et al, 1989: 64). We shall call $\beta_i(\sigma)$ the distribution of beliefs associated with the distribution that is coherent with assigning probability 1 to the strategy σ, and with $\beta(\sigma) = (\beta_1(\sigma), \ldots \beta_n(\sigma)) \in B$ the profile of such beliefs for the *n* players.

Recalling the definition of b_i as the vector collecting the beliefs of each order for player *i*, and consequently of $b = (b_1, \ldots b_n)$ as the profile of beliefs for each of the *n* players, we are now able to provide the definition of Psychological Nash equilibrium (Geanakoplos et al, 1989: 65):

A psychological Nash equilibrium for a n-person normal form psychological game G is a pair $(\hat{b}, \hat{\sigma}) \in B \times \Sigma$ such that:

i) $\hat{b} = \beta(\hat{\sigma})$

ii) for each $i \in I$ and $\sigma_i \in \Sigma_i$, $V_i(\hat{b}_i, (\sigma_i, \hat{\sigma}_{-i})) \leq V_i(\hat{b}_i, \hat{\sigma})$

Condition (ii) is a simple restatement of the standard Nash equilibrium condition, affirming that for each player the equilibrium strategy must confer a payoff not smaller than what attained by any other feasible strategy, *given* the opponents' strategies *and* the beliefs[15]. Condition (i) restrains the beliefs to be coherent with the equilibrium strategy. Notice that if beliefs are not part of the utility function then condition (i) becomes redundant and the definition boils down to the standard Nash equilibrium definition.

4. The Game of Production

After the philosophical and analytical underpinnings of the system of choice of the agents have been set out, we can now apply this model of choice to the analysis of the nonprofit enterprise (NPE hereafter). First, we depict a situation of interaction in the production of a good (section 0), whose outcomes correspond to a variety of different behaviour of a firm corresponding in turn to different organisational form. This game is analysed in accordance with the two attitudes that make up the utility functions of the

players (section 0). In section 5, the Nash social welfare function is adopted as the normative principle used by the agents, and we analyse the conditions under which a nonprofit organisational form can be an equilibrium of the game.

4.1. The Setting of the Game

We suppose that three players are involved in the game of production: a worker (W), an entrepreneur (E) and a consumer (C)[16]. The latter is actually a dummy player, her actions not having any impact on the others' payoffs, though her payoff is affected by the others' actions. The worker and the entrepreneur work together in a firm, and are to decide the degree of their commitment to the company, which is supposed to be measurable along some scale. Their different degree of involvement brings about different organisational forms for the firm. More specifically, each of the active agents has two available strategies; one prescribes performing an action that would be standard in a free-market, profit-oriented context. The other action permits the improvement of the quality of the supplied good with respect to such a forprofit, free market, standard, but triggers an extra-cost, with respect to the alternative strategy, that is to be sustained by the agent herself. For instance, the entrepreneur may decide to adopt a productive practise, or a technology, which permits to increase the quality of the good, where, though, this technology is more costly with respect to that adopted in a purely competitive context. Analogously, the entrepreneur may renounce to a part –or all- of his profits in order to reinvest them in the productive process either by improving the quality or increasing the quantity of the good supplied at the same price. We shall indicate with h_E and l_E the adoption of the good's quality-improving action and that leaving the quality of the good unaltered with respect to market standards respectively, where the letters *h* and *l* refer to the high or low quality-enhancing purpose of the action, and the subscript *E* stands for the entrepreneur.

Likewise, the worker may decide to work at a lower wage than that fixed in a free market context, thus partially – or totally - supplying his labour contribution in a *voluntary* form. Similarly, he may increase his effort in the provision of the good at the same wage. In both cases, either the quality of the good is improved, or this is offered in a larger amount at the same price. We shall indicate this pair of action with h_W and l_W. The consumer does not have actions affecting the utility of the other two agents, but the surplus derived from the consumption of the good depends on its quality, thus on the level of effort put in by the producers.

Following the formalisation introduced in section 0 and 0, we distinguish between the set $I=\{W,E\}$ of the active players and the set $I^*=\{W,E,C\}$ that includes the dummy player C. A strategy set for the two agents can be easily introduced by considering that both have an action that improves the quality of the good and another that leaves it unaltered with respect to a competitive context. We indicate this with $S_i = \{hi\ ,\ li\ \}$, $i \in I$. Also recall that $S = \underset{i \in I}{\times} S_i$, where the generic element $s \in S$ indicates a vector of pure strategies for the two players, and that $\Sigma := \underset{i \in I}{\times} \Sigma_i$ is the set of mixed strategies profile, with generic element $\sigma \in \Sigma$.

The game representing the interaction depicted so far is then as follows:

	h_E	l_E
h_W	$\underline{w}, R-\underline{w}-c, s$	$\underline{w}, R-\underline{w}, \frac{s}{2}$
l_W	$\overline{w}, R-\overline{w}-c, \frac{s}{2}$	$\overline{w}, R-\overline{w}, 0$

Figure 1

The first, second and third terms in each box represent the material payoff for the worker, the entrepreneur and the consumer respectively. *c* stands for the extra cost that must be paid for by the entrepreneur if she wants to engage in the quality enhancing action of the good, namely h_E. *R* indicates the revenues of the selling of the good, which is assumed to be constant in all of the four possible outcomes, and *w* is the wage, which enters as a cost for the entrepreneur and as the only source of material utility for the worker[17]. There are two possible levels of the wage: $\overline{w}$ is a comparatively high level that obtains when the worker supplies a level of labour in accordance with a market standard (strategy l_W), whereas $\underline{w}$ is a lower level that the worker is available to earn when engaged in the good's quality enhancing action (strategy h_W). Therefore, the difference between $\overline{w}$ and $\underline{w}$ is the cut in the real wage that the worker is available to accept in order to improve the quality of the good.

The consumer's utility is given by the surplus gained in the four possible outcomes. This depends on the effort put in by the other agents in improving the quality of the good. In particular, we normalise to 0 her level of surplus in the outcome where neither the worker nor the entrepreneur engage in the quality improving action, that is (l_W, l_E). We then assume that when both agents agree to enhance the quality of the good, the surplus gained by the consumer is comparatively higher, equal with the level *s*, whereas when only one of the two agents contributing to production provides such an activity the surplus reaches an intermediate level, for simplicity equal with *s/2*.

We identify the outcome in which both agents perform the quality improving actions as that leading to the constitution of a nonprofit venture. The intuition is quite simple: provided that by construction the outcome (l_W, l_E) is associated with the level of effort supplied in a free market context, (h_W, h_E) takes on all the relevant characteristic of a nonprofit-oriented firm; that is, the entrepreneur gives up her profits to invest in a quality-enhancing technology, or simply to increase the quality or the quantity of the good, while the worker supplies a larger amount of effort or some voluntary work. The surplus of the consumer is then as high as possible. The other pair of outcomes represent different situations: (h_W, l_E) gives the best payoff for the entrepreneur as she can count on the worker giving the maximum of his effort while not performing any quality-increasing action; conversely (l_W, h_E) provides the worst payoff for the entrepreneur as the extra-costs that she sustains cannot be compensated by the provision of some extra-work by the worker.

If the game were played by the two active players without any concern for the dummy player, then the game of fig.1 would degenerate to the following standard game, where only the payoffs of the agents representing their self-interest are depicted, as they are the only relevant to the solution of the game:

	h_E	l_E
h_W	$\underline{w}, R-\underline{w}-c$	$\underline{w}, R-\underline{w}$
l_W	$\overline{w}, R-\overline{w}-c$	$\overline{w}, R-\overline{w}$

Figure 2

It is apparent how a unique Nash equilibrium in dominant strategies exists, in which both agents perform the low-quality action. In fact, neither agent has any incentive to perform the quality-enhancing action, being the utility of the consumer neglected in this game. One could say that a non-profit form of enterprise could emerge only if some other-regarding attitude toward the beneficiary is sufficiently developed amongst the active agents. However, in what follows this attitude is not directly modelled as altruistic toward the dummy player, but as a conformist preference for mutual compliance with an accepted principle of fairness or toward the non profit ideology. How this ideology can be selected is the argument of the next section.

5. The Psychological Equilibria of the Game

5.1. Contractarianism and the Ideology of the Nonprofit Enterprise

As already pointed out, the set of normative criteria moulding the conformist motive to action (see section 2.5 and 3.1.3) has not been attributed a specific shape yet. In fact, to the purpose of building up a model of choice, our main point was to emphasise the existence of a prompt to action different from the self-interested one, which emphasises the conditional willingness of the agents to abide by some general moral or ideological principle. But the question of the exact shape of such a general principle had been somehow left on the backstage of the argument. To be sure, this is nothing but a secondary question, which conveys other relevant matters like the convergence of every agent to embrace the same general principle as a reference point in the evaluation of their actions. Needless to say, seeking a general answer to those questions lies beyond the scope of this paper.

However we suggest here a conjecture that we take as reasonably suitable for an account of the non profit enterprise, both from the positive and normative standpoints, which is based on the consideration that both the entrepreneur and the worker of the non-profit enterprise are "ideologues" (Ackerman, 1996). We make this point by introducing two assumptions in sequence. These are meant to capture two distinct roles of morality in the NPE: the first is the "rational justification giving" role that we capture in terms of contractarian ethics. The second is the motivational role, which we model by a particular interpretation of the ideal utility of the NPE members. It is a basic tenet of this paper that these two roles must be considered as both indispensable but irreducible to one another, so that both should be squarely faced by any endeavour to explain how morality can play a role in economic organisations[18].

Hp.1:*The NPE internal players' ideology states that the NPE is based on an hypothetical 'social contract' amongst all the players - the consumer included - affirming a principle of fairness.*

The situation has to be understood as if, before playing the actual game, a hypothetical cooperative bargaining game amongst all the players would be played. This game captures the *ex ante* perspective according to which the players could agree to join the organisation in the different roles of entrepreneur, worker and consumer. In doing this they look for a *justification* of their joining the organisation. Thus, they take an impartial or moral point of view, which means that the decision of joining must be rationally acceptable from whichever point of view. In other words, the terms of agreement must be rationally acceptable under the permutation of the personal or role-relative point of views, so that the agreement must result invariant when it is considered under both two apparently distinct perspectives: the perspective of each particular player, choosing according to his best payoff, and the perspective of 'anyone' - that is the perspective of whichever player who would consider the problem of finding an acceptable agreement without any knowledge of his name and personal role in the game (Sacconi, 1991).

In fact the impartial perspective is adopted in order to settle the mission and the conjoint strategy of the organisation, which is intended as the one that would be agreed upon amongst all the internal members and the external stakeholders of the NPE as well. In particular, this perspective is taken in order to identify the reasonable and acceptable balancing amongst the claims of all the interested participants, from which the internal players derive the fiduciary duties that the NPE must discharge toward the beneficiaries (consumer). Thus the "social contract" works as a "Constitutional" ideology legitimating the enterprise as an institution from the *ex-ante* perspective.

At the very core of the contractarian approach lies the idea that a fair distribution can be worked out through a rational agreement for mutual advantage of all the interested parties. The inclusion also of the consumer within the set of bargaining players is due to the impartial perspective taken in this justificatory exercise. As it is an example of the justificatory role of ethics, it disregards the effective influence of the dummy players in the actual game. On the contrary it considers the ex ante perspective in which also the consumer would have a voice about the terms of agreement on the cooperative venture in which the beneficiary essentially contributes, as he accepts to consume the organisation's output. A rational agreement in this hypothetical game thus requires an efficient production of the surplus and its fair distribution amongst the internal and external players as well.

Formally this can be modelled as the requirement that the NPE distributes the surplus according to the Nash Bargaining Solution for cooperative bargaining games, i.e. we pick up the distribution maximizing the product of the three players' payoffs net of the status quo (Nash 1950). Note that Nash Bargaining Solution always selects an outcome reflecting the degree of symmetry of the payoff space, which means that if the payoff space is symmetric the solution is perfectly symmetric amongst the players (i.e. it splits the pie in equal parts). Consequently the solution is covariant with any asymmetry in the utility representation of the outcome space. This solution excludes any discrimination against whichever player (of course the utilities' product becomes zero if any factor in the multiplication is zero) and always selects equality in so far as equality is represented in the shape of the payoff space. In sum, we adopt to the Nash bargaining solution as a normative criterion for defining a moral preference over the outcomes of the original game, which orders outcomes according to 'fairness'[19].

With respect to the non-cooperative game of the foregoing section, the constitutional ideology is what can be called the result of a "pre-play communication" phase, an agreement that players endorse before the beginning of the actual non-cooperative game on surplus allocation. However the actual game of the foregoing section is non-cooperative. This means that commitments on the ideological principle are not binding *per se,* and there is nothing in the rules of the game that make sure that the precepts of the ideology will be enforced or put in practice by the players. Moreover, due to the payoffs structure of the actual game and its Nash equilibrium, we know that the players *do not have* the appropriate incentives to put in practice the precepts of the constitutional ideology. Why then the active players, the entrepreneur and the worker, do comply with their constitutional ideology? Here comes in our second hypothesis.

Hp 2.The internal players of the NPE take the expectations of reciprocity in conformity to the constitutional ideology as a source of utility per se.

In other words, there is an intrinsic source of utility in acting according to the ideology in the event that you believe that, whilst you act according to the ideology, other players are also conforming to the same ideology, and you also believe that they in fact expect you are acting according to the ideology whilst they act according to it. This is where ideal utility based on conformist preferences enter the production game, but now the resulting comprehensive utility function of the players is specified by the contractarian form of the NPE members' ideology**.**

5.2. The Nonprofit Enterprise as a Psychological Equilibrium

Recall that the expression of the Nash welfare function is as follows:

$$N(U_1,...U_N)=\prod_{i=1}^{N}(U_i-d_i)$$

where d_i represents the reservation utility that agents can get when the process of bargaining breaks down, that is when they renounce to act in mutual cooperation. In the present context, we think appropriate to set all of these reservation utilities to the level of zero[20].

Applying this function to our model, and expressing it with respect to the pair of the relevant agents' actions, we obtain the following values:

$$N_{hh} \equiv N(h_W,h_E)=\underline{w}(R-\underline{w}-c)s$$

$$N_{hl} \equiv N(h_W,l_E)=\underline{w}(R-\underline{w})\frac{s}{2}$$

$$N_{lh} \equiv N(l_W,h_E)=\overline{w}(R-\overline{w}-c)\frac{s}{2}$$

$$N_{ll} \equiv N(l_W,l_E)=0$$

For a significant set of the parameters, we can assume that the Nash function is maximised in (h_W,h_E)[21]. Recalling what set out in the previous section, this would be the allocation obtained in the process of bargaining between the three agents.

It is now straightforward to show how the agents can view this outcome as optimal when the conformist utility is sufficiently high with respect to the material. Specifically, we want to prove that (h_W,h_E) can be sustained as a Nash psychological equilibrium, as defined in section 0. Let us first consider the position of the worker and compute his level of utility associated with such an outcome. His material utility is clearly the lower wage; what about his conformist utility? Recalling the expressions of the two functions measuring conformity to ideology, we can notice that, provided that N_{hh} is the maximum for the function, both compliance functions will be equal to zero, thus attributing the maximum value to the ideological source of utility: $V_W\left(h_W, b_W^1 = h_E, b_W^2 = h_W\right) = \underline{w} + \lambda$. Notice that in the computation of this value we have used the definition of the Nash psychological game equilibrium, which implies that the beliefs of the agents must be confirmed by the agents' actual choice. Accordingly, the beliefs assign probability one to the equilibrium strategies (see section 0).

Let us now test whether the worker finds this allocation optimal or he has any incentive to deviate. In Psychological Games, a deviation from a certain allocation consists of a change in the agent's strategy, *given* the set of beliefs held in that allocation. In other words, when deviating, the agent must take into account what the expectations of the other agents on his behaviour are, and then compute the possible change in his own comprehensive utility deriving from not conforming to such expectations. In our case, we shall generically indicate with σ_W<1 the probability with which the worker plays h_W in the mixed strategy adopted in the deviation. The estimation of the entrepreneur's compliance to the ideology is unaffected by this deviation, since by construction the worker knows that she still believes that he is going to perform h_W.

However, the worker's very conformity to the normative principle must change. Given that the entrepreneur is still going to perform with probability one h_E, the resulting value for the Nash function is: $N\left(\sigma_W, h_E\right) = \sigma_W N_{hh} + \left(1 - \sigma_W\right) N_{lh}$. Given the worker's belief, his action that maximises (minimises) the Nash function is to play h_W (l_W). Formally: $N^{MAX}\left(b_W^1 = h_E\right) = N_{hh}$, and $N^{MIN}\left(b_W^1 = h_E\right) = N_{lh}$. Substituting these values into the function measuring the conformity of the worker with the normative principle, we obtain:

$$f_W\left(\sigma_W, b_W^1 = h_E\right) = \frac{\left(1 - \sigma_W\right)\left(N_{lh} - N_{hh}\right)}{N_{hh} - N_{lh}} = -\left(1 - \sigma_W\right)$$

Hence, the comprehensive utility of the deviation is:

$$V_W\left(\sigma_W, b_W^1 = h_E, b_W^2 = h_W\right) = \sigma_W \underline{w} + \left(1 - \sigma_W\right)\overline{w} + \lambda\sigma_W$$

The ideal source of utility is now smaller: the worker is paying the fact that he is not reciprocating the action of the counterpart. Knowing that the entrepreneur is dong her best to act in accordance with the normative principle, the fact that he is partly failing in doing the same causes a lesser satisfaction deriving from the conformist motive. A different but related interpretation is that the worker feels guilty for not having conformed to the counterpart's expectations. On the other hand, the expected value from the

material utility is certainly higher. To ensure the optimality of the choice of the quality improving action for the worker, we therefore need a further condition:

$$V_W\left(h_W, b_W^1 = h_E, b_W^2 = h_W\right) > V_W\left(\sigma_W, b_W^1 = h_E, b_W^2 = h_W\right) \Leftrightarrow \lambda_W > \overline{w} - \underline{w}$$

This condition states that the weight attributed to the ideological source of utility must be sufficiently large so to compensate the loss in material utility caused by not performing the best action in terms of self-interest.

An analogous condition ensuring the pursuing of the quality improving action holds for the entrepreneur:

$$V_E\left(h_E, b_E^1 = h_W, b_E^2 = h_E\right) > V_E\left(\sigma_E, b_E^1 = h_W, b_E^2 = h_E\right) \Leftrightarrow \lambda_E > c$$

We therefore have a simple intuition of how the presence of a conformist motivation in the individual system of preferences helps the emergence of an equilibrium associated with what we can identify as the NPE's behaviour . When the importance attributed to this is sufficiently high in comparison with the material gain that must be given up when acting in conformity with the normative principle, then the outcome in which both agents perform their best action in terms of the interests of the third party involved in the interaction, going against what the pursue of their mere self interest would prescribe, does emerge as an equilibrium of the game. Hence, the presence of two agents motivated to act in accordance with the normative principle, which we identify with the NPE constitutional ideology, emerges as a necessary condition for the emergence of an equilibrium state where we observe the typical behaviour of the NPE.

Up to now this result seems fairly natural: whenever two agents are sufficiently concerned with the conformity to the normative criterion, and when they entertain reciprocal expectations that both will abide by such a criterion, then an conformist equilibrium emerges as a solution of the game. However, there are some questions still unanswered: is the presence of ideology-"motivated" agents a sufficient condition in order to ensure the emergence of this outcome? As we shall argue in the next section, the answer is negative: even when the agents have conformist preferences, the type of interaction resembles a coordination problem, where the outcome corresponding to the for profit behaviour of the firm can emerge as an equilibrium too.

5.3. Multiple Equilibria and Codes of Ethics as Devices for Selection

We now want to investigate if other types of solutions are feasible in the game. First, let us examine whether the "opposite" outcome to that until now considered, in which both agents perform the best action in terms of self-interest (l_W, l_E) can be sustained as a psychological equilibrium. The answer is in fact positive. Consider the worker's situation. Since each agent is performing the *worst* action in terms of the maximisation of the normative function *given* the belief on the other's action, the worker derives utility only from the material component: $V_W\left(l_W, l_E\right) = \overline{w}$. However, the worker cannot gain any benefit from the deviation from this outcome: in fact, the esteem accorded to his counterpart is at the minimum level, namely $\widetilde{f}_E\left(b_W^1 = l_E, b_W^2 = l_W\right) = -1$. Therefore, he does not have any incentive to

perform an action going against his self-interest and somehow respecting the moral principle. Every other strategy cannot do but worse than the current outcome.

Obviously, similar considerations hold for the entrepreneur, thus making (l_W, l_E) a psychological Nash equilibrium for the game. This is indeed a relevant fact: even when agents are inclined to act in accordance with the moral principles reigning in a society, that is their λs are sufficiently high, and this is known to them, there exists an equilibrium where the agents do not care about such morality-grounded motivations and just perform the action respecting their self-interest. This may be indeed be seen as a sort of "nonprofit failure": even when the necessary conditions to build a nonprofit enterprise are present, the self-interested outcome can nonetheless emerge.

The situation is therefore similar to a coordination problem, where the existence of a multiplicity of roughly similar equilibria leaves open the problem of the selection of one of these. That this is indeed the case can be shown more generally: the problem of the choice of each agent's best reply to the opponent's is represented in the graph of fig.3.

It is noticeable how there exists a threshold level in the best reply functions such that each agent performs the "good" action only if the action of the counterpart is sufficiently "good" and vice versa. This gives rise to a third equilibrium, this time in mixed strategies, for the game.

Therefore, the presence of a significant attitude by the agents to perform the actions prescribed by the fairness principle to a full extent is a necessary condition in order that the NPE be derived as an equilibrium of the game. However, this condition is not sufficient: even when agents assign a large "weight" to their conformist motive to action, a failure in signalling their attitude to their counterpart may lead to the selection of the for profit organisational form as the equilibrium.

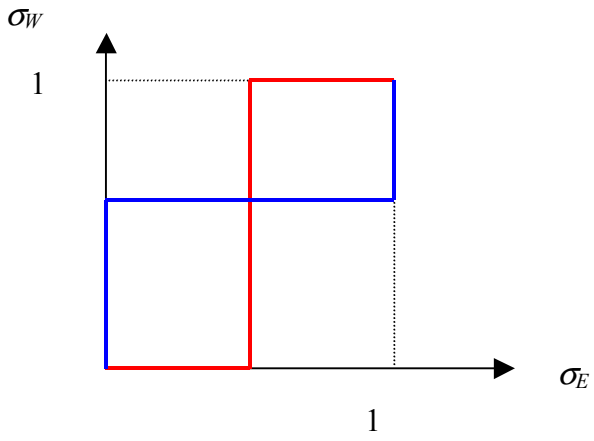

Figure 3

This observation points to the importance of some characteristics of the possible coordination equilibria that may be "external" to the inner structure of the game, and that can act as focal points to make one of the equilibria "salient" with respect to the others. As suggested extensively in the literature (Schelling, 1960; Lewis, 1969; Sugden, 1986), the ability to recognise the salience of one outcome within a set of available results rests on the sharing by the agents of some common cultural traits that makes the convergence to that outcome common knowledge for all the agents. Codes of ethics can be thought of as

an effective devices in order to signal the agents' disposition to coordinate on the socially more efficient outcome in a context of a coordination game or, more generally, in a situation of contract incompleteness. In the present context, codes of ethics would take on the role of signalling the agents' mutual disposition to comply with the moral principles in a pre-play phase; this would make common knowledge their assigning large importance to the ideology-grounded motivation within their individual system of choice. In other words, agents coupled to play the productive game should make it manifest to the counterpart their major attitude to follow their conformist attitude. In our view, this may take the shape of an announcement by both the entrepreneur and the worker directed to the other party concerning the main goals that the partnership in the productive activity should attain. Indeed, this type of announcement is exactly what is embedded in a code of ethics. In such a fashion, codes of ethics can act as focal points generators in solving the coordination problem and in attracting agents with "good" dispositions to the equilibrium associated with the setting up of a nonprofit firm.

Whereas in the context of a purely co-ordination game, such as the psychological game under scrutiny, the present argument does not seem too problematic, the consideration of the most realistic situation of incomplete information on each other's types rises some concerns as to whether the announcement of the constitutional ideology through a code of ethics is sufficient to create the appropriate reciprocal expectation system leading to the NPE. This would be a situation in which the disposition of the agents as to their conformist attitude is a private hidden characteristic, namely a type; in other words, each weight λi would be unknown to the counterpart. It is clear that in such a situation, viewing a code of ethics as a cost-free announcement would not help solving the co-ordination problem between players with high disposition to conform to the ideology. In fact, the possibility that such an announcement is used strategically by a profit-oriented entrepreneur in order to attract the collaboration of non profit-oriented workers, thus bringing about extra-profits for her, would make this device ineffective. In other words, agents with low λ have an incentive to 'cheat' in the pre-play phase, thus leading to the well-known result of a pooling equilibrium in a signalling game.

However, a code of ethics can be seen as a substitute for the commitments within a game of reputation under unforeseen contingencies, where standard commitments on specific and concrete strategies of the game (the standard "types" of the reputation game literature) are made void because of the impossibility to specify *ex ante* their requests contingently upon the unforeseen states of the world that have revealed *ex post* (Sacconi 2000, 2001). In a related work one of us (Sacconi 2002) suggests that a code of ethics may therefore work as the basis for introducing reputation affects in a *repeated trust game* between the NPE as a whole and its external consumers and stakeholder in general, which has been modelled as a game under unforeseen contingencies and incompleteness of contracts. This is in fact the typical context within which it can be expected that an institutional form of firm like the NPE is constituted, such that the firm is endowed with some authority toward the beneficiaries under the condition that it discharges some fiduciary duties toward the beneficiaries themselves. In this case the existence of strong reciprocal effects can be proved between the game of production internal to the firm (the interaction between the entrepreneur and the worker, with the consumer as a dummy player) and the game involving the NPE as a whole and its external stakeholders. On the one hand the existence of ideology and conformist preference

provide the underpinning for assuming that the "type" of the enterprise which discharges its duties according to the commitment (the "type" coinciding to the code of ethics) has positive prior probability. On the other hand, the beliefs dynamics of the reputation model, which proves the existence of an equilibrium of reputation such that the firm complies with its code, makes also salient the outcome of the internal game where the active players give up some of their material utility to the advantage of the consumer. Then the expectations system is formed that supports the emergence of the psychological equilibrium of the internal game in which the ideal utility of the agents plays the main role in guiding their strategy choices (l_W, l_E).

6. Conclusions

The goal of the paper was to offer a characterisation of the nonprofit enterprise to some extent different from the others put forward in the literature so far. Our main point has been to emphasise the importance of the sharing of a common ideology by the participants to the productive venture, whose main feature is the inclusion of all the relevant stakeholders in the decision over the organisational form of the enterprise, the nature of the productive activity, and the distribution of the surplus. In order to attain this goal, we have introduced an individual model of choice encompassing a self-interested and a mutually conformist prompt to action. We have then developed a specification of the latter to bring in the simple intuition that the disposition to comply with moral principles is greater when the other participants to the social interactions are doing so.

Through this model of reciprocity in individuals' system of choice, we have been able to account for the constitution of the nonprofit enterprise as one equilibrium in a Psychological game where the weight assigned to the morality-grounded motivation is sufficiently high compared with the self-interested one. The role of codes of ethics has then been emphasised as a helpful device in order to solve the coordination problem that arises in this type of interactions.

Of course, the analysis is not complete in that some other important aspects of the nonprofit enterprise have been overlooked. First, the question of the efficiency of the nonprofit firm has been somehow neglected, although it is apparent how its constitution can help reducing transactions costs in the "market" of the demand and supply of welfare goods. This aspect has been elaborated in a different work (Sacconi, 2002). Moreover, the extension of the model to the case of incomplete information, which has been sketched in section 5.3, opens the analysis to the relevant issue of the 'external' relation of the NPE with other stakeholders than the consumers, such as donors, where reputation effects become relevant. Said that, we believe that focussing on the "internal" framework of the constitution of the nonprofit venture was a helpful starting point in order to develop a comprehensive theory on this subject, something that we aim to develop in the further stages of our work.

References:

BEN NER, A. and L. PUTTERNAM (eds.) (1998), *Economics, Values, and Organization*, Cambridge: Cambridge University Press, pp. 3-69.

BERNHEIM, B. (1994): "A Theory of Conformity", *Journal of Political Economy*, Vol. 102, N. 5, pp.841-877.

BINMORE, K. (1994), *Game Theory and the Social Contract Volume 1: Playing Fair*, Cambridge MA: The MIT Press,

BINMORE, K. (1997), *Just Playing; Game theory and the social contract, vol.2,* Cambridge MA: The MIT Press

BROCK, H. (1979), "A Game theoretical Account of social Justice", *Theory and Decision*, V. 11, pp.239-265.

BROOME, J. (1999), *Ethics out of economics*, Cambridge: Cambridge University Press.

COPP, D. (1997), "The Ring of Gyges: Overridingness and the Unity of Reason", *Social Philosophy and Policy*, Cambridge University Press, Vol. 14 N.1.

GAUTHIER D. (1986)*, Morals by Agreement,* Clarendon Press, Oxford.

GEANAKOPLOS, J., PEARCE, D., and STACCHETTI, E. (1989), "Psychological Games and Sequential Rationality", *Games and Economic Behavior*, Vol. 1, pp. 60-79

GRIMALDA, G. (2001), " A Survey on the Nature, Reasons for Compliance and Emergence of Social Norms", *LIUC Papers n. 92, Suppl. Oct*.

HANSMANN, H. (1980), "The Role of Nonprofit Enterprise", *Yale Law Journal*, Vol. 89, pp. 835-901.

HANSMANN H. B., (1987), "Economic Theory of Nonprofit Organisation", in Walter W. Powell, (ed.) *The Nonprofit Sector*, Yale UP, New Haven, pp.27-42

HANSMANN H. B., (1988), "Ownership of the firm", *Journal of Law, Economics and Organisation*

HARGREAVES-HEAP, S. (1989), *Rationality in Economics*, New York: Blackwell.

HARSANYI, J. (1977), *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge University Press.

HOGARTH, R., and M. REDER (eds.), (1986), "The Behavioural Foundations of Economic Theory", Journal of Business (supplement).

KOLPIN, V. (1992), "Equilibrium Refinements in Psychological Games", *Games and Economic Behavior*, Vol. 4 N. 2, p. 218-228

KREPS D.M (1990), "Corporate Culture and Economic Theory", in J. Alt and K. Shepsle, (eds.), *Perspectives inn Positive Political Economy*, Cambridge U.P.

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

MERETNS J.F., ZAMIR S. (1985), "Formulation of Bayesian analysis for games with incomplete information", *International Journal of Game Theory*, Vol. 14, No. 1, pp. 1-29

NELSON, R. and WINTER, S. (1982), *An Evolutionary Theory of Economic Change*, Cambridge: Harvard University Press.

NORTH, D.C. (1990) *Institutions, Institutional Change and Economic Performance*, Cambridge: Cambridge University Press.

RABIN, M. (1993): "Incorporating Fairness into Game Theory", *American Economic Review*, Vol. 83, N. 5, pp. 1281-1302.

ROSE ACKERMAN, S. (1987), "Ideals Versus Dollars: Donors, Charity Managers, and Government Grants"*, Journal of Political Economy*, Vol. 95 N. 4, pp. 810-823.

ROSE-ACKERMAN, S. (1996), "Altruism, Nonprofits, and Economic Theory", *Journal of Economic Literature*, Vol. 34, pp. 701-728.

SACCONI, L. (1991), *Etica degli affari, individui, imprese e mercati nella prospettiva dell'etica razionale,* Milano: Il Saggiatore.

SACCONI, L. (1997) : *Etica, economia ed organizzazione*, Bari: La Terza.

SACCONI, L. (2000), *The Social Contract of the Firm*, Berlin: Springer,.

SACCONI, L. (2001), *Incomplete contracts and corporate ethics: a game theoretical model under fuzzy information*, Liuc Papers n.91 October.

SACCONI L. (2002); *The efficiency of the non profit enterprise: constitutional ideology, conformist preferences and reputation*, Liuc Papers n. 111, July; forthcoming in B. Hogdson (ed.) *The Invisible hand and the common good*, proceedings of the SEEP Conference, Berlin: Springer Verlag.

SCANLON, T.M. (2001), "Symposium on Amartya Sen's Philosophy: 3 Sen and Consequentialism"*, Economics and Philosophy*, Vol. 17, pp. 39-50.

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

SEN, A. (1985), 'Well-being, Agency and Freedom', *Journal of Philosophy*, 82: 169-221

SEN, A. (2000), 'Consequential Evaluation and Practical Reason'. *Journal of Philosophy*, 97: 477-502

SEN, A (2001), "Symposium on Amartya Sen's Philosophy: 4 Reply"*, Economics and Philosophy*, Vol. 17, pp. 51-66.

SUGDEN, R. (1998a), *The motivating power of expectations*, mimeo

SUGDEN, R. (1998b), "Normative expectations: the simultaneous evolution of institutions and norms", in Ben-Ner, A. and Putterman, L. (eds): *Economics, Values, and Organization*, Cambridge: Cambridge University Press, pp. 73-100.

SUGDEN, R. (2000) "Team Preferences", *Economics and Philosophy*, Vol. 16, pp. 175-204.

VERBEEK, B. (2001), "Consequentialism, Rationality and the Relevant Description of Outcomes", *Economics and Philosophy*, V. 17, pp. 181 – 205.

WEISBROD, B. (1988) *The Non Profit Economy*, Cambridge, MA: Harvard University Press.

Notes

[1] For the evolution of the concept of "utility" in economic theory, see Broome, 1999: Chapter 2.

[2] Despite this change in the perspective and scope of rational choice theory, this approach cannot be said to be immune from various types of criticism, both of empirical and theoretical nature. On the one hand, critics stress the bulky informational assumptions that are needed in order that such a logically coherent set of choices be made. On the other hand, experimental economics single out the existence of systematic violation of the axioms underlying the standard theory of rational choice, especially under conditions of uncertainty, in the individuals' actual choices. See for instance Hogarth and Reder: 1985, and Hargreaves-Heap (1989). For a review: North, 1990: ch. 3. Also, Nelson and Winter (1982) are among the first authors who have extensively argued on this subject.

[3] In particular, the threshold level that Rabin thinks of as appropriate in order to classify an action as kind or hostile is what he calls the 'equitable payoff', which consists of the middle point between the best and the worst payoff the agent can obtain, provided that both of the associated outcomes are Pareto efficient.

[4] Our argument may be subject to the following type of criticism: an outcome can always be defined so that it comprises every characteristic to which the agent assigns value, thus also possibly including deontological properties of the patterns of actions. In fact, these elements can be included in the description of the state of affairs. This is essentially the theory advocated by Sen (1985, 2000, 2001), with particular reference to the notion of *freedom* as the significant deontological property. As Scanlon (2001) points out, this approach calls for a subjective theory of value, whereas only on an objective account can the distinction between 'consequences' and 'actions' leading to such consequences still be said to be significant and neat. In particular, the latter case corresponds to what Scanlon calls *Foundational* Consequentialism, which is consistent with classical Utilitarianism, as opposed to *Representational* Consequentialism, where value is subjectively determined by the agent *and* some notion of fairness pre-exists to that, so that the traditional means-ends relationship is no more than a formal construct of rational choice. Verbeek (2002) even holds a more radical position in arguing that the inclusion of the agent's concern for the fairness of the process is incompatible with any notion of Consequentialism. Notice that Sen's theory would in fact lie in the middle between these two categories, in that he endorses a subjective account of value *but* moral values are not pre-determined: in this there would lie the properly consequentialist trait of his theory. According to this view, the distinction between 'consequences' and 'set of actions' that leads to such consequences may appear somehow redundant. This would also undermine our distinction between consequentialist and deontological preferences. However, we believe that this separation at any rate significant in that it helps to clarify the different sources of value that the agents deem as relevant. See also note 9 on this point.

[5] More about this in Sacconi (2002)

[6] Harsanyi (1977) states the set of symmetrical rationality postulates from which the bargaining solution is derived, Binmore (1997) shows a symmetrical bargaining game suitable for ethical theory. See also Sacconi (1991) for a different account.

[7] Hierarchies of beliefs are typical game theoretical constructions built on David Lewis' seminal account of common knowledge (Lewis 1968); see Mertens and Zamir (1985) and Tan and Werlang (1988). They are also basic for the theory of psychological games (Geanakoplos et al. 1989).

[8] As long as conformist preferences are assumed to satisfy the formal conditions for being represented by a utility function, I suggest that this is an example of the *betterness relationship* proposed by John Broome (Broome 1999), which is a binary relations expressing whichever reason for saying that in one state of affairs or action there is "more good" (it is better) than in another. Therefore, these preferences can be represented by a utility function, even if it does not corresponds in any sense to the typical "desire" or "revealed" interpretation of preference.

[9] Outcomes are here intended as what happen to each single player in consequence of the result of a certain way of playing the game by the participants (see Harsanyi 1977, p. 90); see also Binmore (1992, p.27): "Each terminal node [of a game tree] must be labelled with the consequences for each player if the game ends in the outcome corresponding to that terminal node". In this sense outcomes are the relevant description of the state of affairs resulting from strategic interaction required by

consequentialist preferences. However, the same state of affair can be described also directly in terms of characteristics of the action *per se*. We have suggested in the foregoing section that the relevant characteristic is the fairness of the strategy combination, but we also argued that this can be detected by a property of the utility distribution attached to the outcomes as compared to an abstract principle of fair distribution.

[10] To be sure, that agents are able to fully compare different values is far from an trouble-free question: for the question of incommensurability of values, see Broome (1999, Ch. 8 and 9). For a sceptical view doubtful as to the possibility of comparing various reasons to action in an individual's system of practical choice, see Copp (1997).

[11] Although beliefs are probability distributions iteratively defined over probability distributions, the associated probabilities over pure strategies can be easily obtained by means of the following formulas: $P_{b_i^1}(s_j) = \int_{\Sigma_j} P_\sigma(s_j) P_{b_i^1}(\sigma_j) d\sigma_j$; $P_{b_i^2}(s_i) = \int_{B_j^1} P_{b_j^1}(s_i) P_{b_i^2}(b_j^1) db_j^1$.

Thus the first formula indicates the overall probability that player *j* is going to play s_j, according to the belief b_i^1 held by player *i*, and the second the overall probability that player *j* holds about *i*'s performing s_i, according to the second order belief b_i^2 .

[12] Notice the dependence of $T^{MAX}(b_i^1)$ and $T^{MIN}(b_i^1)$ on the belief b_i^1 . Indeed the belief is necessary in order to determine the probabilities for the expected value of the welfare function, which is: $T(\sigma_i, b_i^1) = \sum_{S_i}\sum_{S_j} \overline{T}(s_i, s_j) P_{\sigma_i}(s_i) P_{b_i^1}(s_j)$, where the probability $P_{\sigma_i}(s_i)$ is what prescribed by the mixed strategy σ_i, and $P_{b_i}(s_j)$ is the probability computed in accordance to the formula of the previous note.

[13] To be sure, if agent *i* performs her worst action in terms of conformity to the normative principle, the fact that agent *j* acts contrarily or in favour of the same normative principle does not affect *i*'s overall utility function. Therefore, we can interpret the situation where one or both the agent perform the action leading to the worst outcome according to the welfare distribution function as one in which the social contract between the agents breaks down.

[14] Of course this is only one of the possible models of the ideological motive to action. Another one, which, *mutatis mutandis,* coincides with Rabin's specification is the following:

$$V_i(\sigma_i, b_i^1, b_i^2) = U_i(\sigma_i, b_i^1) + \lambda_i \left[\frac{1}{2} + \tilde{f}_j(b_i^2, b_i^1)\right]\left[\frac{3}{2} + f_i(\sigma_i, b_i^1)\right]$$

An "equitable" payoff in the normative function is here identified with half of the difference between T^{MAX} and T^{MIN} . Hence, agent *i* will experience a positive incentive to perform an action increasing the social welfare only when the opponent performs an action above this level. However, if agent *j* executes an action below this equitable level, then agent *i* would be subjected to an incentive to act *contrarily* to the normative criterion. This specification seems to emphasise the aspect of reciprocity *per se* partly neglecting the other aspect of the will to contribute to the normative principle satisfaction. We think, however, that this emphasis would be somehow inappropriate in the present context, thus opting for a specification in which the incentive provided by the opponent in acting according to the normative principle is always non-negative, and nil only in the extreme case of him inflicting the least value to the social welfare function. A specification in which the agent is not concerned with the action of the counterpart would be the following:

$$V_i(\sigma) = U_i(\sigma) + \lambda_i[T(\sigma)]$$

This account captures the idea that agents are interested in the fulfilling of the normative principle of distribution through the materialisation of appropriate social outcomes, without any concern for the other agents' commitment to the same principles. This specification can be taken as a useful reference point with respect to the more elaborated version of the next section, other than an interesting account of the ideal motive to action *per se*.

[15] The refinements of such a notion of equilibrium deal with the possibility that the beliefs of the player that is "deviating" from the equilibrium can vary as well, reflecting the "direction" of this deviation.

[16] As customary, we attribute different sexes to the players: E and C are both females, whereas W is a male.

[17] For simplicity the material utility of both worker and entrepreneur is assumed to be linear in the monetary revenue.

[18] For a similar point see Gauthier (1986), where he makes the basic distinction between *internal rationality* of the social contract, what can be solved in terms of rational bargaining theory, and *external* rationality of the social contact, i.e. the compliance problem, a point that however we face in a completely different way by introducing conformist preferences.

[19] The idea to base the Social Contract on Nash's bargaining solution was first given by Horace Brock (Brock 1979) see also (Sacconi 1986, 1991, 2000). It is also adopted in a somewhat different way by Ken Binmore (Binmore1997). I admit that using here the words "Social Welfare Function" can be misleading, because they induce to think that there exists a sort of super-individual decision maker whose objective function is defined according to the SWF. That is not the case however. By this SWF I only mean an ethical criterion of fairness useful to judge the outcomes of the game. It is not a consequence that a decision maker would bring about for herself. This is clear given the underlying contractarian account of the Nash Bargaining Solution.

[20] This choice calls for some justification. Many authors would argue that the proper choice for the "exit option" would be the Nash solution of the material game played non-cooperatively. However, this choice is not immune for criticism as a possible situation of prevarication of one party over the other in the *status quo* would carry over to the final "moral" solution. This is the reason why other authors have proposed the notion of a "moralised" status quo, in which some minimal form of reciprocal respect are already in place. Therefore, one may consider our choice equivalent with a, perhaps naive, notion of moralisation of the status quo from which the "bargaining" starts.

[21] In particular, $N_{hh} > N_{hl} \Leftrightarrow R - \underline{w} > 2c$ and $N_{hh} > N_{lh} \Leftrightarrow 2\frac{R - \underline{w} - c}{R - \overline{w} - c} > \frac{\overline{w}}{\underline{w}}$. The first condition implies that the extra cost required for the quality improving technology is not too large in comparison with the profits of the firm when the worker accepts the lower wage. The second condition ensures that the increase in the consumer and entrepreneur's utility when the worker partly acts voluntarily compensates the loss in the earnings of the worker himself.